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

Student Name: Kevin Peng

Student Number: A16253731

University of California: San Diego

Final Project

ECON 178

Economic & Business Forecasting

Professor: Ying Zhu

August 2023

Word Count: 3318

Introduction

How well do various indicators predict a household's total wealth? In this paper, a predictive model for total wealth (tw) in US dollars is constructed using data from the 1991 Survey of Income and Program Participation (SIPP). Such a model would be immensely useful for businesses and policymakers in many scenarios, such as pinpointing families that are in need of welfare programs. The dataset consists of 7933 observations, focusing on households with reference persons aged 25-64 years old, in which at least one person is employed and none are self-employed.

The dataset offers a diverse array of independent variables that can potentially influence total wealth. These variables can be broadly categorized into the following groups:

Variable to predict (outcome variable):

- tw: Total wealth (in US $), which is defined as "net financial assets, including Individual Retirement Account (IRA) and 401(k) assets, plus housing equity plus the value of business, property, and motor vehicles."

Retirement-related variables:

- ira: individual retirement account (IRA) balance (in US $).
- e401: Binary variable, where 1 indicates eligibility for a 401(k)-retirement plan, and 0 indicates otherwise.

Financial variables:

- nifa: Non-401k financial assets (in US $).
- inc: Income (in US $).

Home ownership-related variables:

- hmort: Home mortgage (in US $).
- hval: Home value (in US $).
- hequity: Home value minus home mortgage.

Other variables:

- educ: Education (in years).
- male: Binary variable, where 1 indicates male and 0 indicates otherwise.

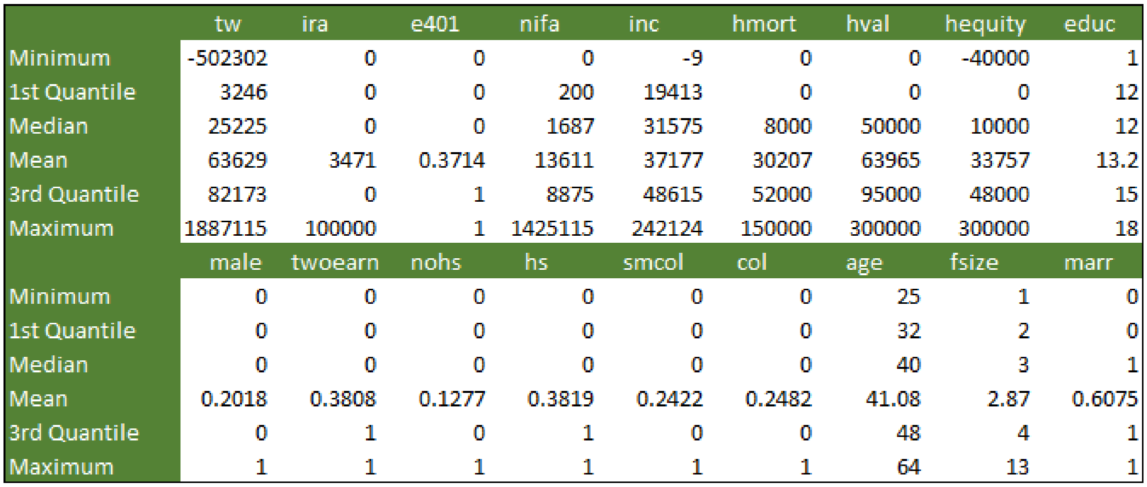
- twoearn: Binary variable, where 1 indicates two earners in the household, and 0 indicates otherwise.
- nohs, hs, smcol, col: Dummy variables for education levels - no high school, high school, some college, college.
- age: Age.
- fsize: Family size.
- marr: Binary variable, where 1 indicates married and 0 indicates otherwise.

By understanding the aforementioned features, which is another word for covariates, and their relationships with the outcome variable (total wealth), we can explore how different factors are associated with wealth accumulation within a household.

Statistical Analysis

Inspecting the Variables

In the analysis conducted, a thorough examination was performed on the variables in the dataset to gain insights into their distributions and potential susceptibility to outliers. Notably, the variables ira, nohs, smcol, col, and male exhibited a value of 0 at the 3rd quantile. This implies that these variables are heavily positively skewed, with a significant number of data points taking on the value of 0. Consequently, these variables' mean values are susceptible to the influence of outliers, as the presence of extreme values could significantly impact the mean. Therefore, it is crucial to exercise caution when performing subsequent analyses involving these variables.



	tw	ira	e401	nifa	inc	hmort	hval	hequity	educ
Minimum	-502302	0	0	0	-9	0	0	-40000	1
1st Quantile	3246	0	0	200	19413	0	0	0	12
Median	25225	0	0	1687	31575	8000	50000	10000	12
Mean	63629	3471	0.3714	13611	37177	30207	63965	33757	13.2
3rd Quantile	82173	0	1	8875	48615	52000	95000	48000	15
Maximum	1887115	100000	1	1425115	242124	150000	300000	300000	18
	male	twoearn	nohs	hs	smcol	col	age	fsize	marr
Minimum	0	0	0	0	0	0	25	1	0
1st Quantile	0	0	0	0	0	0	32	2	0
Median	0	0	0	0	0	0	40	3	1
Mean	0.2018	0.3808	0.1277	0.3819	0.2422	0.2482	41.08	2.87	0.6075
3rd Quantile	0	1	0	1	0	0	48	4	1
Maximum	1	1	1	1	1	1	64	13	1

Avoiding Perfect Multicollinearity

Two variables (covariates) are said to be multicollinear when they are highly correlated. This poses an enormous threat to linear regression, notably ordinary least squares (OLS) regression. Since OLS regression estimates the coefficient of its covariates in the form of the following expression, it involves inverting a matrix consisting its covariates. A matrix that contains perfectly collinear covariates, is not invertible, therefore running the OLS regression inoperative, in which case the programming language R would drop one dummy anyway.

$$\hat{\beta} = (X'X)^{-1}X'y$$

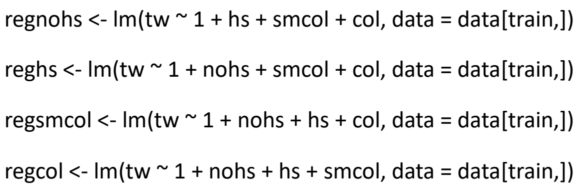
An argument can be made that ridge regression, which estimates the coefficient of its covariates in the following expression, making the regression operable even in the case of having highly collinear covariates. However, including such a perturbation term comes at a cost of increasing the bias of the estimator, which in certain cases decreases its predictive power.

$$\hat{\beta}^{RR} = (X^T X + \lambda I)^{-1}(X^T Y)$$

In addition, removing one of the highly multicollinear terms offer better interpretability of the coefficients within the model, whether under the setting of ridge regression or any other regressions. Hence, the rational decision to make is to drop one of the highly multicollinear terms irrespective of the specific model selection.

The presence of perfect multicollinearity among the three variables, hval, hmort, and hequity, is evident due to the relationship described by the expression hequity = hval – hmort. To alleviate the effects of multicollinearity, it is the convention to drop one of the variables. In this case, the rational decision is to drop the variable hequity, given that it encapsulates both hval and hmort. By discarding hequity, two distinct features, hval and hmort, are retained.

A similar phenomenon can also be observed amongst the four education dummies, namely, nohs, hs, smcol, and col. For each observation, the sum of these four dummies is always equal to 1, indicative of perfect multicollinearity. To evaluate which dummy to drop, a test-train split, in which 1/10 of the data is randomly attributed to the testing set and the other 9/10 is attributed to the training set, is created to evaluate the performance of the following four simple linear regressions, each of which omits one of the four dummies. The rationale behind partitioning the dataset into 10 subsets for validation purposes is to strike a balance between achieving reasonably accurate results, leveraging the substantial size of the dataset comprising 7933 observations, and avoiding undue computational burden during the validation process. Under the seed 123, an arbitrary number created to ensure that running the code would yield the same exact result, the mean square predicted error (MSPE) for all four models is the same, having a value of 11322526334.8804. Hence, a conclusion can be made that dropping any variable from the four dummy variables would yield the same predictive performance.



```
regnohs <- lm(tw ~ 1 + hs + smcol + col, data = data[train,])

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

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

regcol <- lm(tw ~ 1 + nohs + hs + smcol, data = data[train,])
```

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$y_i = \textit{actual value}$$
$$\hat{y}_i = \textit{predicted value}$$
$$n = \textit{\# of observations}$$

Choosing which education dummy to drop, however, does affect the interpretability of the coefficients. Whichever variable that is dropped serves as a reference point for the model. The coefficient of col, for example, in the case that hs is dropped, is interpreted as how much more wealth households with a reference person with a college degree would be expected to have when compared to households with a reference person with a high school degree. Since the prompt has no clear preference for an education level reference point, I choose to drop the hs dummy because it is the only variable that doesn’t have a value of 0 on the 3rd quantile, an indication that high school degree is a relatively accurate representation of the population’s average education attainment.

Creating a Baseline Model for Future Benchmark

After a preliminary inspection of the variables and the removal of any perfect multicollinearity, a baseline model can be created for benchmark. The baseline model aims to construct a linear model that minimizes the MSPE within its designated category (linear models).

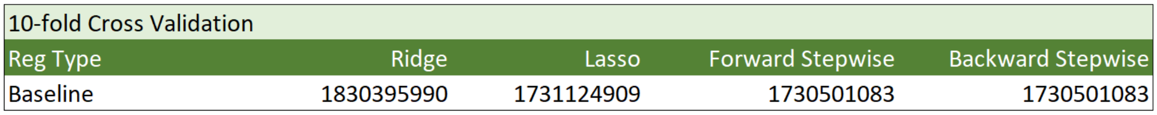
This model serves as a foundational reference for the subsequent application of nonlinear transformations, which will be elaborated upon in later sections of this paper.

Assuming for linearity in the relationships between tw and individual covariates, there are several options: OLS regression, forward/backward stepwise regression, ridge regression, and lasso regression. Given that forward/backward stepwise regression entails a sequential selection or removal of covariates based on OLS regression, conducting a comparison between OLS regression and stepwise regression is redundant.

A comparison is made between the three types of regressions using K-fold cross validation. The dataset is divided into K subsets. The proposed model (regression) is trained on K-1 subsets and tested on the remaining one, repeating K times. A value of 10 is chosen for K for the aforementioned reason. Given that the investigation extends beyond the mere assessment of the effects of dropping different dummy variables, the adoption of K-fold cross-validation, rather than a simple test-train split, safeguards the MSPE against the inadvertent influence of outlier-rich data partitions.

An alternative to this approach is to use Leave-One-Out cross validation. Leave-One-Out cross validation is very similar to K-fold in its approach, except that the model is trained on n-1 observations and tested on the remaining one observation, repeating n times (n is defined to be the number of observations within the dataset). Leave-One-Out validation is optimal in the case of a small dataset because it is accurate but computationally heavy. The considerable size of our dataset, comprising 7933 observations, serves as a protective measure against any excessive variation arising from the choice of K-fold over Leave-One-Out cross validation. Hence, this paper will exclusively use K-fold validation as its validation method from now on.

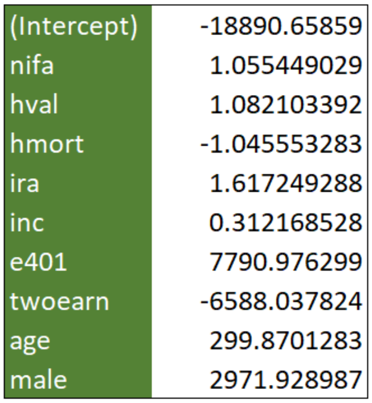
The 10-fold cross validation is performed, and the results are as follows. It can be observed that forward and backward stepwise regressions are surprisingly equally potent in their predictive performance, while ridge and lasso regressions provide lackluster results in comparison.



10-fold Cross Validation

Reg Type	Ridge	Lasso	Forward Stepwise	Backward Stepwise
Baseline	1830395990	1731124909	1730501083	1730501083

Since it is unclear whether every single covariate is highly predicative of tw, choosing forward stepwise regression is likely to provide a more accurate and concise model than choosing backward stepwise regression. Running a forward stepwise regression on the entire dataset obtains the finalized baseline model, whose coefficients are shown below.



(Intercept)	-18890.65859
nifa	1.055449029
hval	1.082103392
hmort	-1.045553283
ira	1.617249288
inc	0.312168528
e401	7790.976299
twoearn	-6588.037824
age	299.8701283
male	2971.928987

Assessing the Relationships between Individual Features with Total Wealth (tw)

Recall that the baseline model was constructed under the assumption that the relationships between the features and the outcome variable, total wealth, are strictly linear. However, this may not be true for many of the features. To create a more accurate model, the model shall fit different functional forms tailored to distinct segments (covariates) of the dataset.

To achieve this, transformations will be made to individual covariates of the dataset, and they will be replacing the original covariates to create a flexible linear model. This method is called generalized additive model (GAM).

$$\begin{aligned} y_i &= \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \epsilon_i \\ &= \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \epsilon_i. \end{aligned}$$

This paper will utilize splines rather than polynomials for the following reasons:

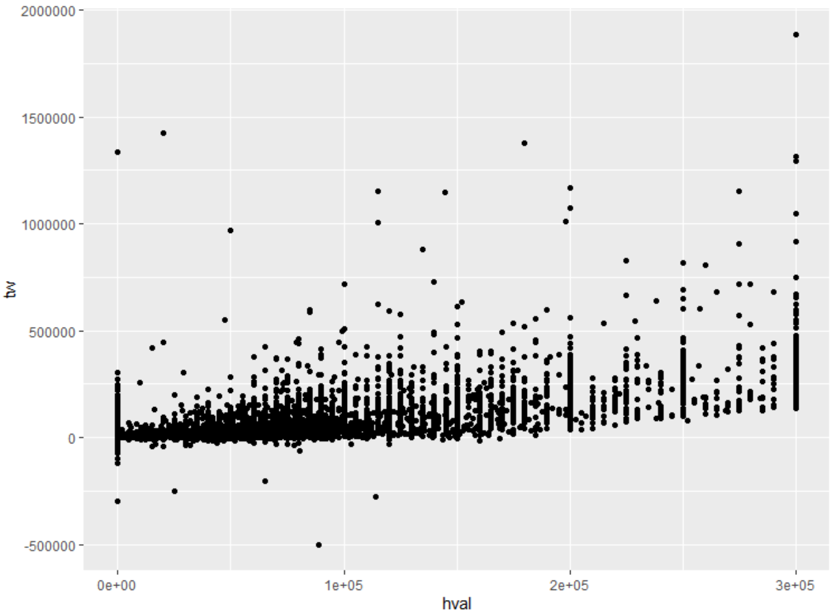
a) Polynomials must use a high degree for flexible fits, but splines are able to do so with the degree fixed. This is likely to produce more stable estimates.[1]

b) Polynomials lack the ability to incorporate thresholds like splines, leading to undesirably global outcomes. In other words, observations within one range of the predictor strongly influence the model's behavior across different ranges.[2]

[1] Springer Texts, 2013, *An Introduction to Statistical Learning*

[2] Magee, 1998, *The American Statistician*

c) A polynomial fit is likely to produce undesirable results at the boundaries.[3]



tw ~ hval

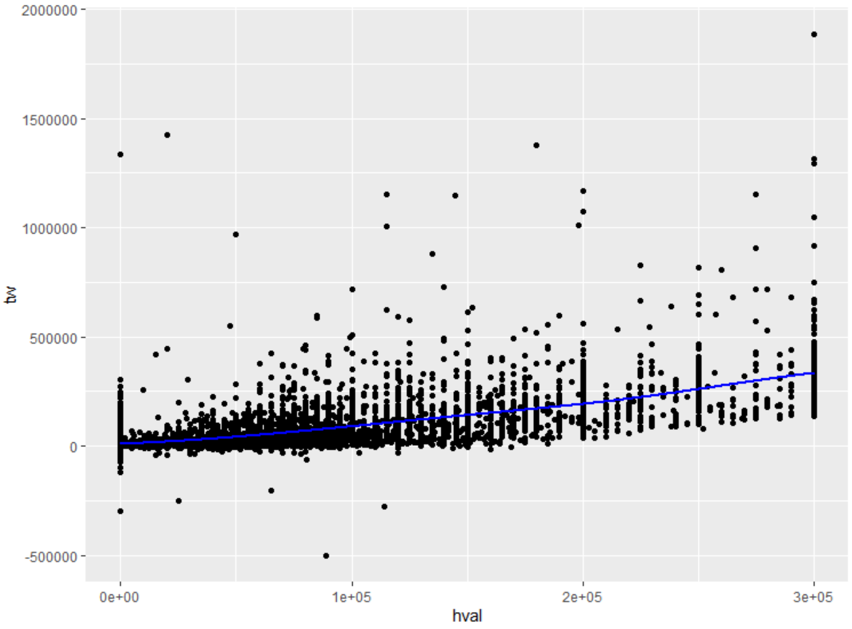
From the graph above, it is evident that the relationship between hval and tw is nonlinear, as tw soars higher at a faster rate at high ranges of hval. From a conceptual standpoint, individuals tend to allocate more of their financial resources towards expenditures beyond their residential properties once their housing conditions have reached a level of comfort and adequacy. Consequently, households with higher property values are likely to possess even greater levels of overall wealth.

It should be noted that there are large values of tw across all ranges of hval, and these observations are not likely to represent the average household in the population. Hence, all observations with tw>1000000 are removed for the selection of the optimal number of knots.

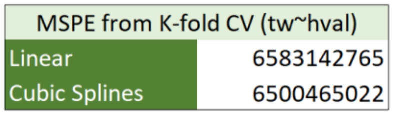
A 10-fold cross validation is utilized to ascertain the optimal number of knots for the cubic splines. A double loop is constructed for this purpose. The inner loop iterates through various quantities of knots, while the outer loop traverses the 10 partitions of the dataset. To ensure that knots are not placed at the minimum and the maximum value of hval, two additional knots are added and specified to be boundary knots. The cross-validation process ascertains that including 3 knots, excluding the two boundary knots, in the cubic splines minimizes the MSPE.

[3] Springer Texts, 2013, *An Introduction to Statistical Learning*

Subsequently, a splines model is trained on the complete dataset, encompassing the previously excluded outliers. This approach is adopted in consideration of the subsequent integration of splines into the Generalized Additive Model (GAM), where the inclusion of outliers remains uncertain in this stage. The result is shown below.

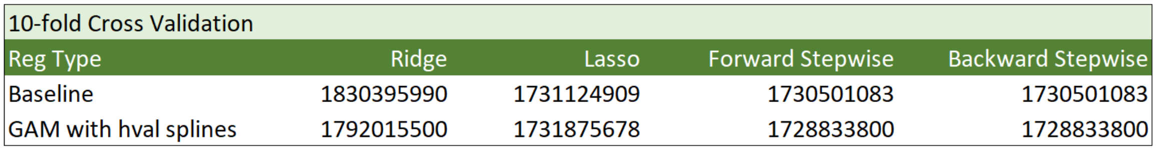


To establish the superiority of the model over its linear counterpart, a 10-fold cross validation is performed, revealing a significant reduction in MSPE.

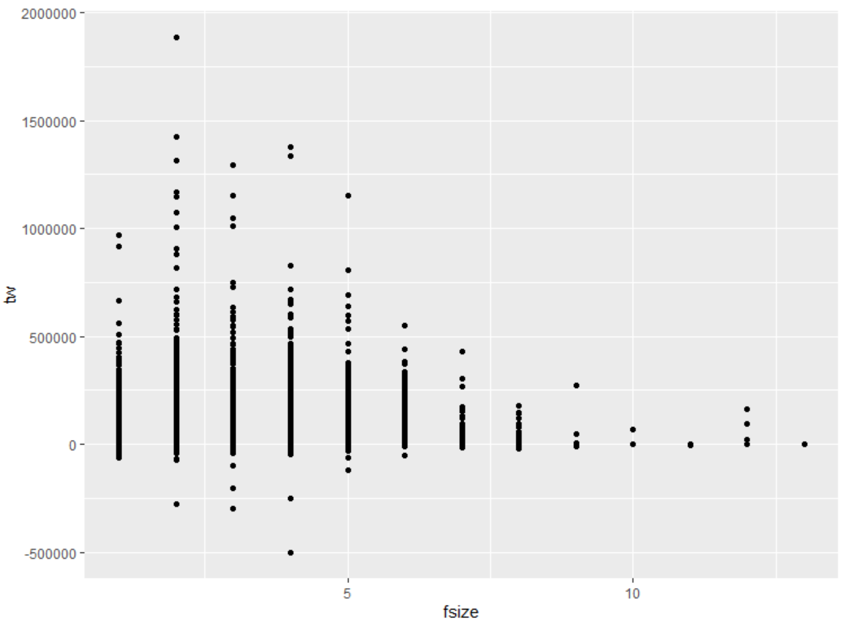


MSPE from K-fold CV (tw~hval)	
Linear	6583142765
Cubic Splines	6500465022

Now incorporating the transformations into the full dataset and discarding the original hval column, conducting a 10-fold cross validation using all regression options shows the following result, evincing its predictive accuracy in comparison with the baseline model.



10-fold Cross Validation				
Reg Type	Ridge	Lasso	Forward Stepwise	Backward Stepwise
Baseline	1830395990	1731124909	1730501083	1730501083
GAM with hval splines	1792015500	1731875678	1728833800	1728833800

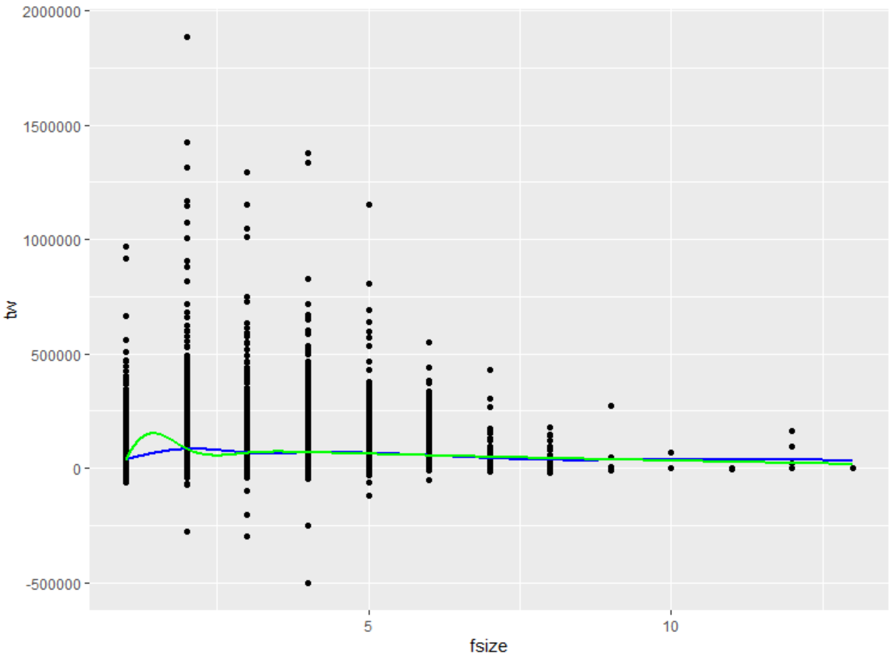


tw ~ fsize

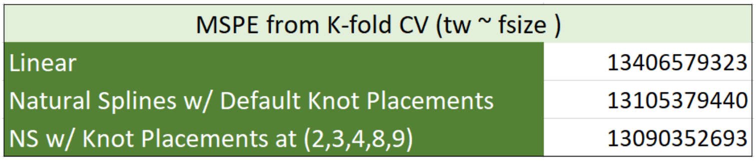
The graph depicting the relationship between total wealth and family size exhibits a conspicuous non-linear configuration, standing out as one of the most prominent instances within the data. This observation aligns with intuitive reasoning, as exceedingly large family sizes impart an immense economic strain on the household, while more affluent families are better positioned to support multiple children. Since higher values of tw (tw>1000000) are focused on the lower ranges of fsize, these observations are not likely to be due to chance and therefore serve as good predictors of the model. Hence, these observations will not be removed for the consequent analysis.

Given the scarcity of observations in the upper fsize ranges, the usage of natural splines is more appropriate for this instance than ordinary cubic splines. A 10-fold cross validation is utilized to ascertain the optimal degree of freedom for the natural splines. A double loop is constructed for this purpose. The outer loop iterates through various degrees of freedom, while the inner loop traverses the 10 partitions of the dataset.

It is evident that fsize is a discrete predictor with very limited range, spanning from 1 to 13. Hence, the selection of the number of knots, which is equal to the degree of freedom in the context of natural splines, should be exercised with extreme precaution, as too many knots would overfit the data. Luckily, the degree of freedom leading to the lowest MSPE is found to be 5.



However, the default natural splines function in R selects the knots arbitrarily at equidistant locations, producing a result that leaves more to be desired (shown in green above). This approach fails to capitalize on the advantages of splines, which enable the placement of additional knots in regions exhibiting nonlinear fluctuations, while employing fewer knots in relatively flat regions within the data. Following a series of knot placement experiments, knot placements at 2, 3, 4, 8, and 9 (shown in blue) yield the best MSPE results. Another 10-fold cross validation is conducted to compare the potency of the custom knot placement against the default knot placement. As shown below, the



MSPE from K-fold CV (tw ~ fsize)	
Linear	13406579323
Natural Splines w/ Default Knot Placements	13105379440
NS w/ Knot Placements at (2,3,4,8,9)	13090352693

Unfortunately, when the splines basis functions are included along with other predicator variables, though, the overall model yielded a higher MSPE despite the spline basis functions' superior predictive performance when tw is regressed on fsize alone. This outcome suggests that the fsize splines shall not be incorporated into the final model.

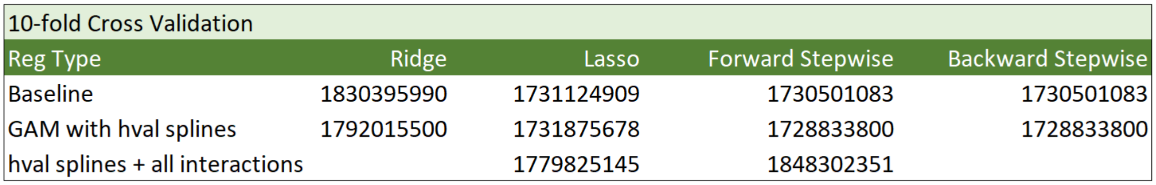
Interaction Terms

Currently, our existing model lacks any interaction terms. Nonetheless, it is worth considering that certain covariates may not be independent of one another, potentially leading to a situation where the effect of one covariate influences the effect of another. By incorporating interaction terms, we can effectively capture these complex dependencies and enhance the model's explanatory power and predictive performance.

Given the substantial number of covariates in the dataset, visualizing and analyzing the complex relationships between these covariates and the outcome variable becomes arduous and time-consuming. To streamline the process, a for loop can be employed to systematically create all possible two-way interaction terms and add them to the dataset.

Recall that the original dataset has 18 variables in total. After removing hs and hequity to prevent perfect multicollinearity, the dataset now has 16 variables. Deducting the outcome variable, the dataset has 15 features, or covariates. 15 features create 14 + 13 + … + 1 = 105 total possible two-way interaction terms. Luckily, the dataset has an enormous number of observations to support the analysis.

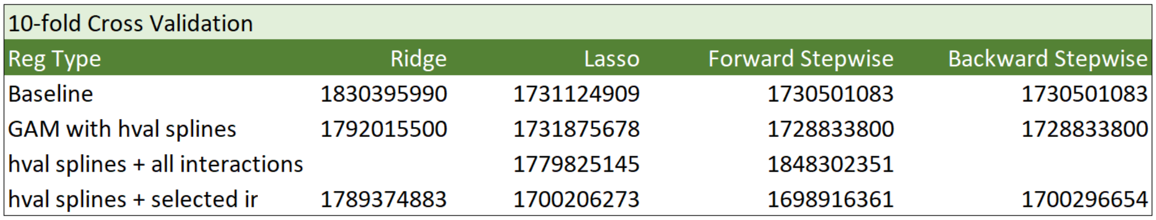
Given that the 105 interaction terms are not analyzed prior to the regression, an assumption can be made that many of them will likely be inconsequential predictors for tw. Consequently, ridge regression and backward stepwise regression are deemed unsuitable for our analytical approach. Instead, the focus is on utilizing lasso and forward stepwise regression techniques, renowned for their proficiency in feature selection.



10-fold Cross Validation				
Reg Type	Ridge	Lasso	Forward Stepwise	Backward Stepwise
Baseline	1830395990	1731124909	1730501083	1730501083
GAM with hval splines	1792015500	1731875678	1728833800	1728833800
hval splines + all interactions		1779825145	1848302351	

The result of the cross validation is expected, as too many covariates, even after some feature selection, tend to overfit the data. However, the regressions are valuable in providing a filtered list of interaction terms, upon which a manual sorting process can take place. Interaction terms will be selected from the list provided by forward stepwise regression, which yielded a lower MSPE.

Extracting only the most statistically significant interaction terms, marked “***” by the forward stepwise regression, and re-running a cross validation yields a favorable outcome.



10-fold Cross Validation

Reg Type	Ridge	Lasso	Forward Stepwise	Backward Stepwise
Baseline	1830395990	1731124909	1730501083	1730501083
GAM with hval splines	1792015500	1731875678	1728833800	1728833800
hval splines + all interactions		1779825145	1848302351	
hval splines + selected ir	1789374883	1700206273	1698916361	1700296654

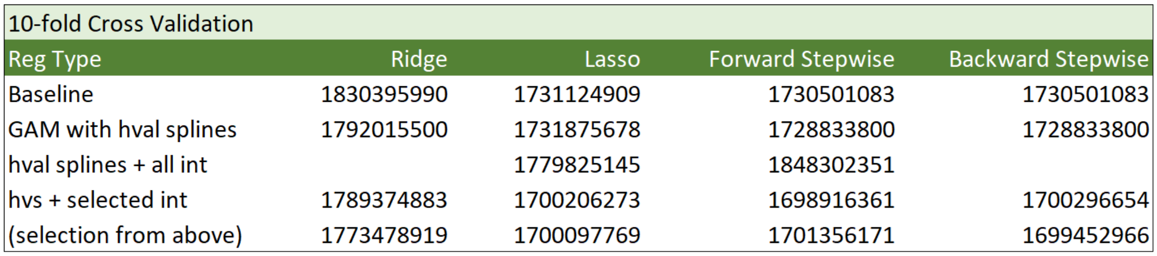
Finalizing the Models

Until this juncture, we have executed various models on the complete model encompassing all raw covariates, except those that underwent transformations. It is noteworthy that ridge regression lacks any capability for feature selection. This places ridge regression at a substantial disadvantage when contrasted with lasso and forward stepwise regression.

Given the presumption of overfitting within the extant model, the removal of irrelevant predictors not only confers benefits to ridge, but also yields advantages for lasso and forward stepwise regression by curtailing the MSPE.

After excluding all covariates omitted by the forward stepwise regression, a novel dataset is formulated. The application of the same four types of regression on the newly created dataset engenders a conflicting result. Ridge regression is now more accurate but lasso and forward stepwise regressions drop in their predictive accuracy.

After evaluating the MSPE across all models through K-fold cross-validation, it can be inferred that the most favorable outcome is achieved by applying forward stepwise regression to the dataset with hval transformations and the selected interaction terms.



10-fold Cross Validation				
Reg Type	Ridge	Lasso	Forward Stepwise	Backward Stepwise
Baseline	1830395990	1731124909	1730501083	1730501083
GAM with hval splines	1792015500	1731875678	1728833800	1728833800
hval splines + all int		1779825145	1848302351	
hvs + selected int	1789374883	1700206273	1698916361	1700296654
(selection from above)	1773478919	1700097769	1701356171	1699452966

Conclusions

In conclusion, this paper aimed to construct a predictive model for total wealth (tw) in US dollars using data from the 1991 Survey of Income and Program Participation (SIPP), with potential applications in business and policymaking. The dataset, containing 7933 observations of households with employed individuals aged 25-64, encompassed a diverse range of covariates that could influence total wealth. The study undertook a thorough examination of variables and addressed multicollinearity concerns, laying the foundation for subsequent analyses.

The initial baseline model, assuming linear relationships between covariates and total wealth, was established using various regression techniques. A comprehensive comparison of ordinary least squares (OLS) regression, forward/backward stepwise regression, ridge regression, and lasso regression was performed through K-fold cross-validation. Among these, forward and backward stepwise regressions exhibited similar predictive performance, while ridge and lasso regressions yielded less favorable results.

Acknowledging the potential non-linearity of relationships, the study introduced generalized additive models (GAM) by applying spline transformations to specific covariates. This approach led to refined models that better captured the intricacies of the data.

Furthermore, the exploration of interaction terms among covariates enriched the model's explanatory power. Forward stepwise regression, known for its feature selection ability, was employed to select statistically significant interaction terms. Incorporating these terms into the model through cross-validation produced a more refined and predictive outcome.

In essence, this study demonstrated the effectiveness of tailored modeling approaches, including transformations, interaction terms, and feature selection, in enhancing the predictive accuracy of wealth prediction models. The paper's findings contribute to a better understanding of the complex relationships between covariates and total wealth.

Limitations

Several limitations of this study should be addressed. This project is completed as a final project for a summer session course, which, unfortunately, is ran under an extremely tight time constraint. If more time is given, more nonlinear relationships can be explored and are likely to provide a more accurate prediction of tw.

The methodologies used in this paper derives only from well-established methods in the data science industry. If more time is available, a more complex machine learning model can be built on top of the four regression packages using methods such as queries.

Also, given the time constraints, I created a code that is limited to this study and requires constant manual maneuver and on-the-fly analysis and decisions regarding the data. It is not applicable to other studies.

This paper is also limited to the course materials of the summer session course. There is only a limited selection of methods:

- Regression: OLS, Lasso, Ridge, Forward/Backward Stepwise
- Nonlinearity: Polynomials, Step Functions, Cubic Splines, Natural Splines

This model can definitely be improved by numerous other methods. Poisson distribution, for example, might be beneficial to the analysis of discrete variables such as age and educ.