Developing a Prediction Model for Money Spent at the Grocery Store

Kristen Keller

Introduction

In this study, we were interested in predicting the amount of money that a person spends at the grocery store per month using 22 available predictor variables. In the first section of this paper, nonlinear modeling techniques such as regression splines, LOESS, and smoothing splines are utilized to model the relationship between key predictor variables and the outcome variable. In the second section, model selection techniques such as best subset selection, LASSO, ridge regression, and stepwise regression are used to generate different models for predicting the amount of money spent at the grocery store from the given covariates.

Data Source

Data from the 2011-2012 National Health and Nutritional Survey (NHANES) were used. Every year the National Health and Nutrition Examination Survey is administered by the CDC to assess the health and nutrition status of adults and children living in the United States. Observations are collected on non-institutionalized citizen residents of the United States. A multi-level sampling scheme is used to select a representative sample of individuals from specific households within selected neighborhoods in selected counties. Each year over 5,000 individuals participate in the at-home interview portion of the survey, from which these variables are drawn.

In the 2011-2012 survey 9,756 individuals were selected, 9,338 of whom were examined. Data from the demographics, consumer behavior, dietary habits, food security, and physical activity portion of the questionnaire were used, along with information from the nutrition section. A list of the specific variables used in this study can be seen below (Table 1).

Table 1: Variables included from the NHANES 2011-2012 dataset

Section	Name	Description	Type	Notes
Demographics	RIAGENDR	gender	categorical	
Demographics	RIDAGEYR	age	continuous	topcoded at 80
Demographics	INDHHIN2	annual HH income	continuous	topcoded at 100,000
Demographics	INDFMPIR	ratio of family income to poverty	continuous	topcoded at 5
Demographics	DMDHHSIZ	number of people in HH	continuous	topcoded at 7
Demographics	DMDHHSZA	number of children <5 in HH	continuous	topcoded at 3
Demographics	DMDHHSZB	number of children 6-17 in HH	continuous	topcoded at 4
Consumer Behavior	CBD070	money spent on grocery store past 30 days	continuous	
Consumer Behavior	CBD090	money spent on nonfood items past 30 days	continuous	
Consumer Behavior	CBD110	money for food at other stores past 30 days	continuous	
Consumer Behavior	CBD120	money spent eating out past 30 days	continuous	
Consumer Behavior	CBD130	money spent on delivery past 30 days	continuous	
Dietary Habits	DBD895	# meals not prepared at home per week	continuous	
Dietary Habits	DBD381	# times per week get school lunch	continuous	adjust DBD895
Dietary Habits	DBD411	# times per week get school breakfast	continuous	adjudt DBD895
Dietary Habits	DBD905	# ready to eat meals in past 30 days	continuous	
Dietary Habits	DBD910	# frozen meals or pizzas in past 30 days	continuous	
Food Security	FSD032B	food didnt last because of money issues	categorical	
Physical Activity	PAQ605	vigorous physical activity at work	catgorical	recode combined PAQ

Section	Name	Description	Type	Notes
Physical Activity	PAQ620	moderate physical activity at work	catgorical	recode combined PAQ
Physical Activity	PAQ650	vigorous physical activity recreationally	catgorical	recode combined PAQ
Physical Activity	PAQ665	moderate physical activity recreationally	catgorical	recode combined PAQ
Nutrition	DR1TKCAL	calories in kcal in past 24 hours	continuous	
Nutrition	DR1TPROT	protein in gm in past 24 hours	continuous	
Nutrition	DR1TCARB	carbohydrates in gm in past 24 hours	continuous	
Nutrition	DR1TSUGR	sugars in gm in past 24 hours	continuous	
Nutrition	DR1TFIBE	fiber in gm in past 24 hours	continuous	

Partitoning of data

For this analysis, the assumption is made that missing values in the data are missing completely at random. As such, only observations with no missing values for the variables of interest are used. Of the 9,338 participants that were originally examined, only the 5355 will be used in this study. Prior to applying any modeling techniques, the data was split into a test set and a training set. A random split was used, allocating approximately 75% of the data to the training set and 25% to the test set. 4035 observations were allocated to the training dataset and 1320 to the test dataset.

Table 2: Summary statistics for continuous variables

	Train: Mean	SD	Min	Max	Test: Mean	SD	Min	Max
Age	31.78	22.65	1.00	80.00	30.10	22.19	1.00	80.00
Income to Poverty	2.42	1.68	0.00	5.00	2.43	1.66	0.00	5.00
Number in HH	3.60	1.63	1.00	7.00	3.65	1.62	1.00	7.00
Children Under 5	0.43	0.71	0.00	3.00	0.44	0.72	0.00	3.00
Children 6-17	0.94	1.11	0.00	4.00	0.95	1.13	0.00	4.00
Money: Grocery Store	438.68	304.45	0.00	5142.00	435.42	280.38	0.00	2571.00
Money: Food Other	94.82	166.95	0.00	2142.00	99.38	170.61	0.00	2142.00
Money: Eating Out	170.59	219.14	0.00	3000.00	170.21	213.63	0.00	2571.00
Money: Delivery	29.95	67.59	0.00	1000.00	29.40	72.30	0.00	1200.00
Money: Nonfood	35.48	68.17	0.00	1071.00	37.13	65.28	0.00	428.00
Not Prepped at Home	3.94	3.68	1.00	21.00	4.00	3.84	1.00	21.00
Ready to Eat	1.82	5.54	0.00	90.00	1.75	4.84	0.00	60.00
Frozen Meals	2.80	7.61	0.00	180.00	3.10	8.77	0.00	180.00
Fast Food	2.02	2.55	0.00	21.00	1.94	2.43	0.00	21.00
Food Didn't Last	0.26	0.44	0.00	1.00	0.27	0.44	0.00	1.00
Calories	2108.87	946.32	188.00	13687.00	2123.52	956.26	487.00	8359.00
Protein	77.37	39.54	2.97	387.37	79.34	38.84	10.42	289.06
Carbohydrates	267.04	124.31	3.80	1815.02	265.12	120.88	50.45	1150.35
Sugar	122.52	77.12	0.69	1048.48	118.89	71.82	3.08	668.24
Fiber	16.11	9.93	0.00	103.50	16.02	9.60	0.80	68.30

Summary statistics for the test and train datasets can be seen below. For most of the continuous predictors, the mean and standard deviation in the test set and the train set appear to be similar. All of the observations for each of the continuous variables appear to fall within a reasonable range (ex. no one claims to have eaten -3 or 50,000 grams of sugar). For each of the categorical variables, the proportion of observations in each category appears to be similar in the test and train datasets (Table 2, 3).

Table 3: Summary statistics for catagorical variables

	Train: Sum	Train: Proportion	Test: Sum	Test: Proportion
Male	2046	0.51	670	0.51
Food Didn't Last	1033	0.26	355	0.27
Physical Activity	2257	0.56	680	0.52
Income: 1	1210	0.30	381	0.29
Income: 2	918	0.23	308	0.23
Income: 3	686	0.17	238	0.18
Income: 4	1221	0.30	393	0.30

Statistical Analysis: Nonlinear Modeling Techniques

Three predictor variables of interest were chosen and nonlinear modeling techniques were applied to model the relationship between the response variable and each of the predictors. The predictors that were chosen were calories, income to poverty ratio, and money spent eating out.

First, natural cubic splines were fit to each of the predictor variables. For each variable, 5 different numbers of knots were tried. Specifically, models were fit containing 0, 1, 2, 3, and 4 knots. Originally, K-fold cross validation was used in an attempt to find the optimal number of knots, however the results of the cross validation did not seem satisfactory. There was very little difference between the models using 2, 3, and 4 knots for most of our predictor variables, so the number of knots chosen through cross validation varied wildly depending on the way the observations were randomized into the different folds for cross validation. Plots were created comparing the splines with different numbers of knots and these plots were used to make final decisions about the total number of knots. For calories, one knot appeared to be satisfactory. For income to poverty ratio, three knots appeared to be best. For money eating out, one knot also appeared to be sufficient (Figure 1).

Let it be noted that in all plots of the response variable against predictor variable used in the nonlinear modeling portion of the study, two observations with particularly high values for the response variables (more than $3000 spent at the grocery store in the past month) were excluded to reduce the scale of the y axis and make trends in the plots easier to visualize. These observations were still used to calculate the nonlinear fits, but they will not be shown in figures

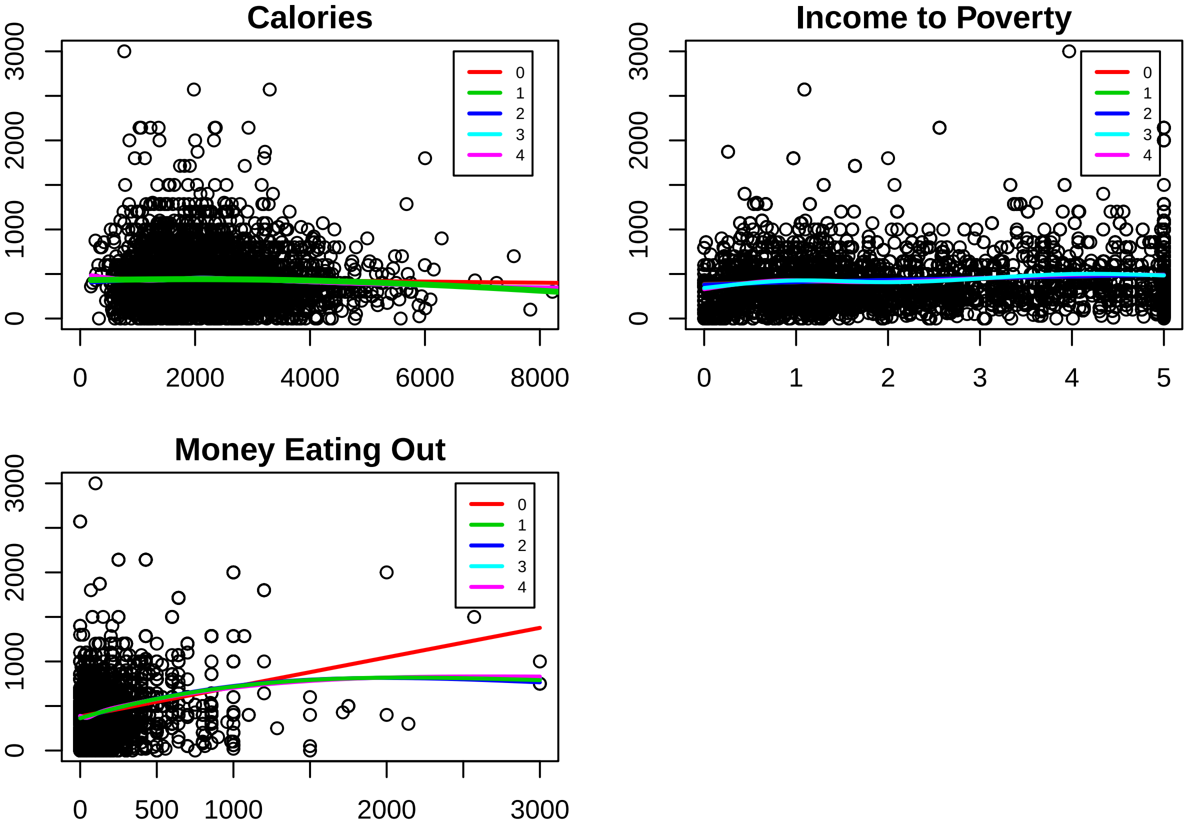

Figure 1. Natural cubic splines applied to three variables of interest. For each plot, the y axis represents money spent at the grocery store and the x axis represents the variable specified in the title of the figure.

Next, smoothing splines were fit to each predictor variable. A sequence of smoothing parameters ranging from 0.1 to 0.9 in steps of 0.1 was examined. Again, the smoothing parameters chosen using cross validation were not always the same parameters chosen based on a visual examination of the plots. In cases where the parameter choices based on cross validation and plot appearance were different, the latter was used. Based on cross validation, the best smoothing parameter for calories was one. The same parameter was chosen based on plot appearance. For income to poverty ratio and money spent eating out, the best smoothing parameters appeared to be 0.2 and 0.5, respectively. Based on chart appearance, the best parameters appeared to be 0.6 and 0.85, respectively (Figure 2).

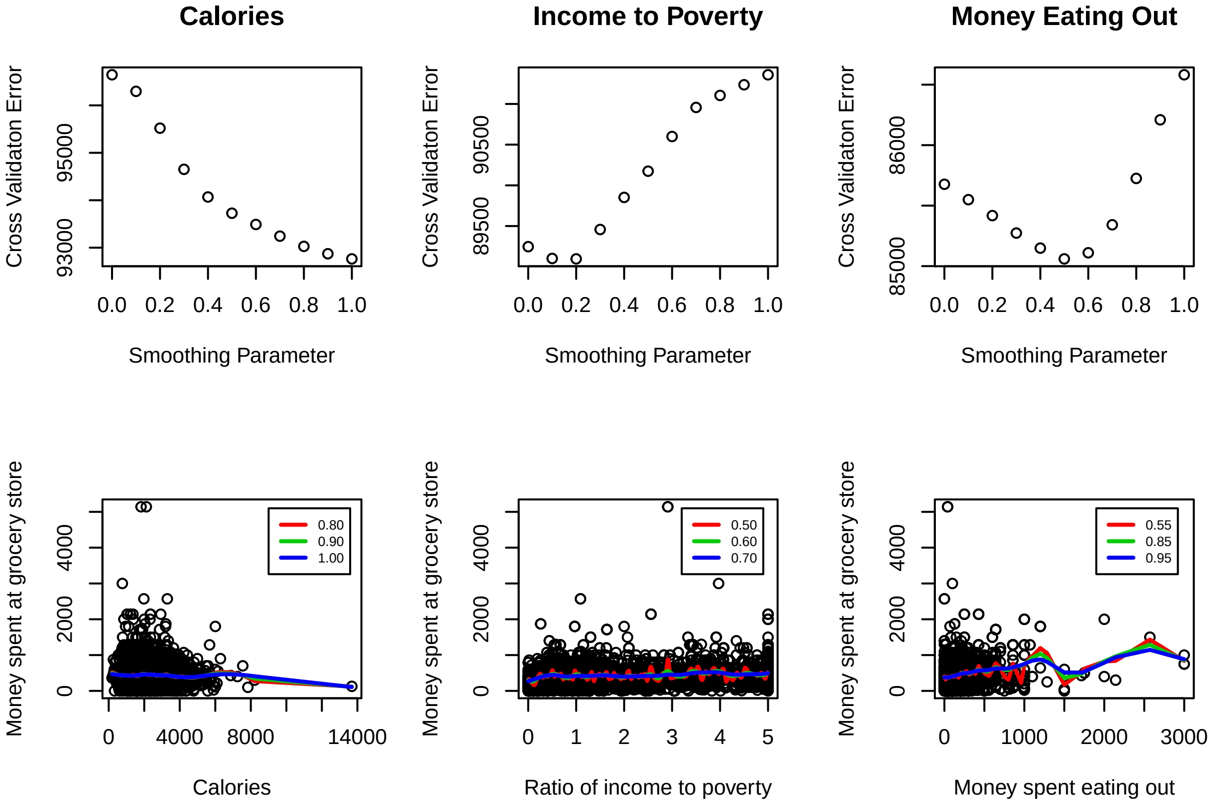

Figure 2. Smoothing splines applied to three variables of interest

Next, loess smoothers with spans of 0.2, 0.4, 0.6, and 0.8 were fit to the scatterplots of money spent on groceries against each predictor variable of interest. Using best judgement, 0.6, 0.4, and 0.4 were chosen as the smoothing parameters for calories, income to poverty ratio, and money spent eating out, respectively.

Overall, the different nonlinear modeling techniques gave fairly similar results. For the calories variable, the relationship with money spent on groceries appeared to be linear when regression splines were used. Smoothing splines and loess regression picked up some non-linearity in the relationship, but even based on these plots it appears that a linear term would have been sufficient for most applications. For the income to poverty variable, the relationship with money spent on groceries appeared to be approximately linear for loess regression, whereas some slight deviations from linearity were detected using regression splines and smoothing splines.

The differences between the methods were most clear when looking at the relationship between the variable representing money spent eating out and money spent at the grocery store. Using regression splines and smoothing splines, there appeared to be a zone for which the response variable was increasing with the predictor variable, then a zone for which the response variable leveled off. Using smoothing splines, the fit in the former zone was the same but the fit in the latter zone was different. Instead of leveling off, there was another dip, then a subsequent increase in the fitted values of the response variable. Test data could be used to help detect whether this pattern is truly there or whether it is just noise. Overall, it seems that the smoothing splines were most sensitive to slight deviations from linearity. This makes sense because the regression splines do have to adhere to the overall structure of polynomial regression and loess methods use data from surrounding neighborhoods to calculate the fit at each point, so sudden deviations at a certain point in the data do not have as much influence when using these methods.

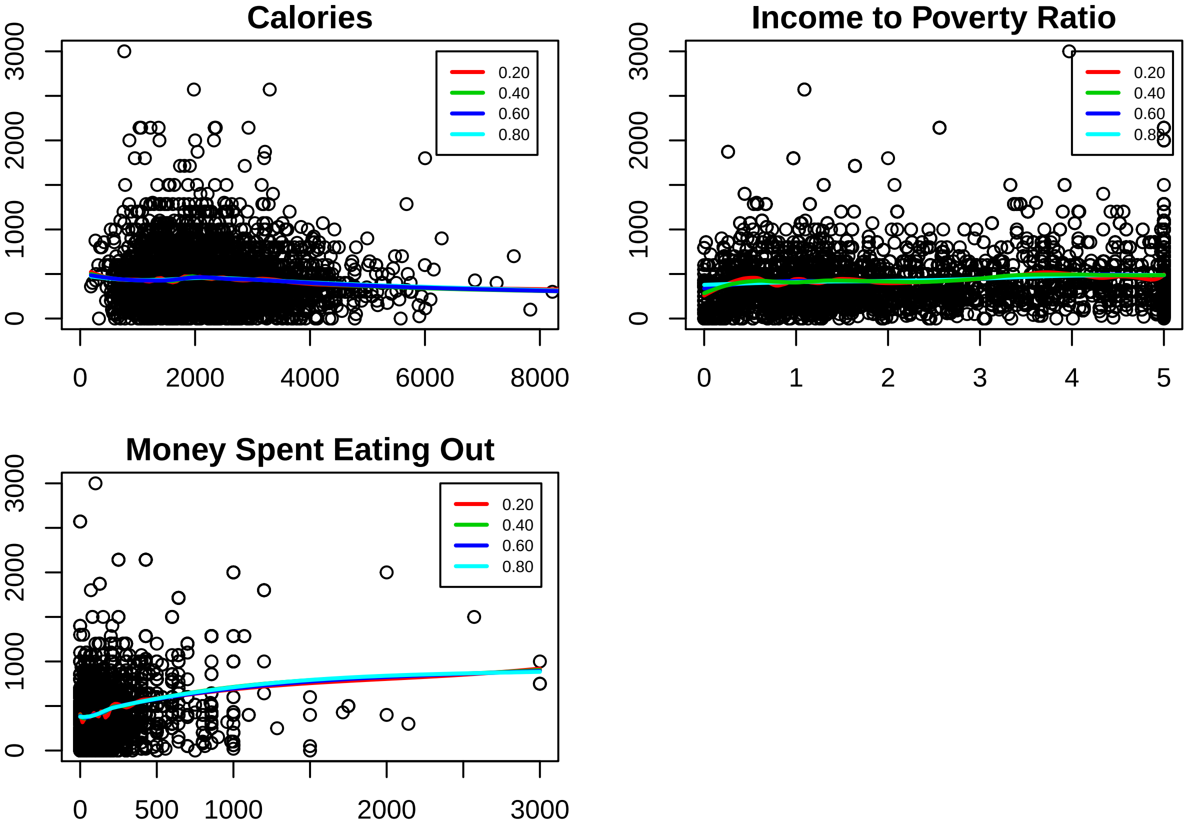

Figure 3. Loess smoothers applied to three variables of interest. For each plot, the y axis represents money spent at the grocery store and the x axis represents the variable specified in the title of the figure.

After nonlinear modeling approaches were applied to individual predictors, GAMs were fit to allow for the nonlinear modeling of multiple predictors at once. A selection of predictors that were designated as variables of interest prior to examining the data were used. These variables were income to poverty ratio, number in household, number of meals not prepared at home, calories consumed in the past 24 hours, and money spent on nonfood items. Calories consumed, income to poverty ration, and number of meals not prepared at home had linear relationships with the response variable. The number of people in the household had a slightly nonlinear relationship and money spent on nonfood items had a distinct nonlinear relationship with the resposnse variable.

Three main classes of GAMs were fit. In the first class, only the money spent on nonfood items variable was fit using nonlinear modeling techniques. In the second class, both the money spent on nonfood items and the number of people in the household variable were fit using nonlinear modeling techniques. In the final class of models, all five variables were fit using nonlinear modeling techniques.

For the class three models with all variables fit using nonlinear modeling techniques, there did not appear to be much difference between models that used regression splines, smoothing splines, or loess smoothers to fit these variables. After testing multiple models, it became apparent that it was not necessary for these models to be fit using nonlinear methods. There appeared to be only a slight advantage to fitting the number of people in household variable using nonlinear techniques. It did seem necessary to fit the money spent on nonfood items variable using nonlinear techniques. The final model that was arrived at used linear terms to fit calories consumed, income to poverty ration, number of meals not prepared at home, and number of people in household and fit money spent on nonfood items using a smoothing spline (Figure 4).

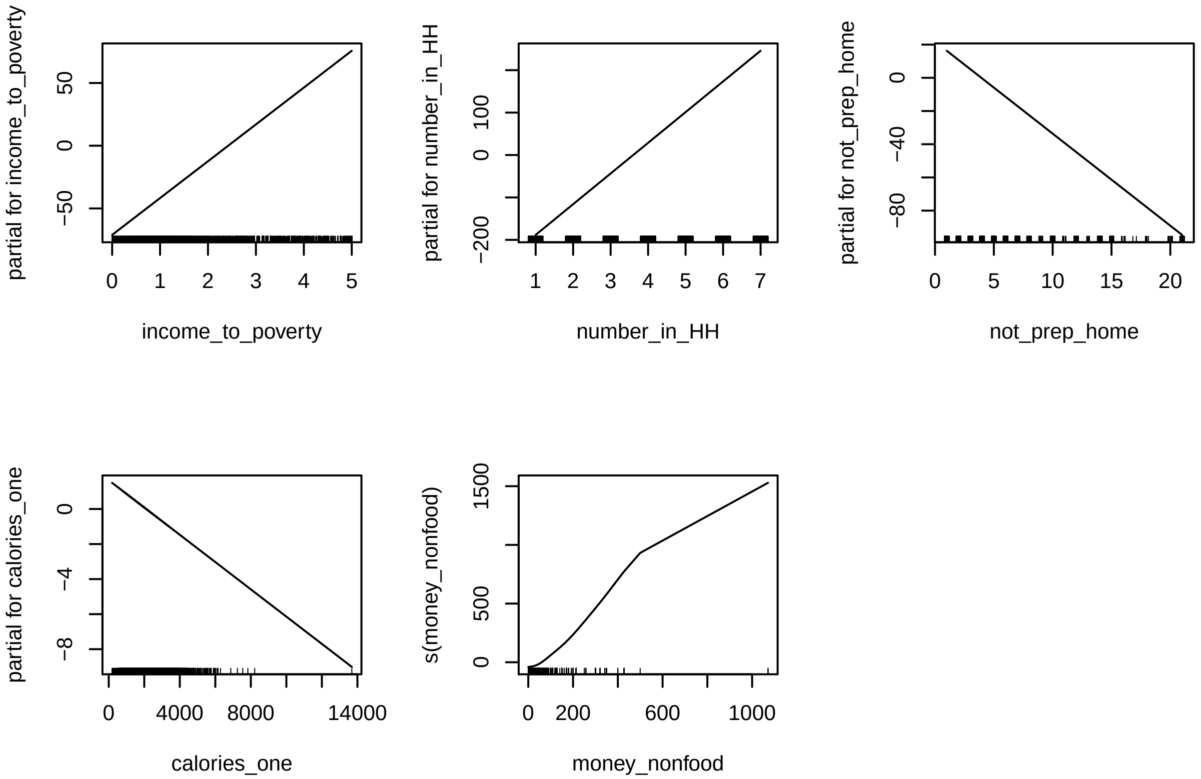

Figure 4. Plots of individual variables included in GAM against partial residuals.

It appears that most of the predictors in our dataset could be modeled more simply using a linear term or a simple transformation. There are no threshold values at which the relationship between a certain predictor variable and response variable suddenly changes. Many of the variables could be modeled using a linear term, but some others could use a log transformation or a square root transformation. For the most part, the variables that require this sort of transformation are variables relating to money spent at places other than the grocery store.

Statistical Analysis: Model Selection Techniques

In the next part of this study, model selection techniques such as best subset selection, LASSO, and stepwise regression were applied to the data in order to select the covariates that best predicted the amount of money spent at the grocery store. Prior to applying model selection techniques, transformations were performed on the predictor variables that displayed nonlinear bivariate relationships with the response variable. Loess smoothers were applied to help visualize the relationships. A square root transformation was applied to the variables representing the number of meals not prepared at home, the amount of money spent eating out, the amount of money spent on food at other stores, the amount of money spent on delivery, and the amount of money spent on nonfood items. A log transformation was applied to the variable representing the number of frozen meals eaten. Prior to the log transformation, one was added to each observation because many observations had an original value of 0. All further mentions of these variables will refer to the transformed versions of the variables.

The first model that was created was a model based on prior judgement. This model consisted of variables that we thought might be important predictors prior to examining the data. This model included the variables representing income to poverty ratio, number of people in the household, number of meals not prepared at home, the number of calories eaten the previous day, and the amount of money spent on nonfood items.

The next two models were created using best subset selection. For both models R^2 was used to decide between models of the same size. For the first model AIC was used to decide among models of different sizes and for the second model BIC was used. There were 15 predictors included in the final best subsets model created using AIC and 7 predictors in the model created using BIC.

The next 2 models were created using forward and backward stepwise selection. For both models, R^2 was used to choose among models of the same size and BIC was used to choose among models of different sizes. The first model included 7 predictors and the second model included 7 predictors. The same model was chosen using forward and backward stepwise selection. The final two models were achieved after running a LASSO and Ridge Regression. Cross validation was used to choose the tuning parameters for each. For the LASSO model a lambda value of 0.2625 was selected as having the lowest cross validation error and for ridge regression a lambda of 13.37 was selected. The LASSO model contained 21 predictors and the ridge regression mode contained all 22 predictors.

The lambda chosen for the LASSO model using the 1 se rule was 43.79. When this value was used, the final model contained only 4 predictors. This seemed like a more reasonable number of predictors as the previous selection methods selected only a few predictors to be in the model. This value of lambda was used moving forward. The lambda selected for ridge regression using the 1 se rule was also used. This value was 418.

Table 4: Coefficient values for each regression model

	Prior	Subsets (AIC)	Subsets (BIC)	Forward (BIC)	Back (BIC)	Lasso	Ridge
Intercept	74.762	92.285	109.113	109.113	109.113	218.161	248.839
Male	0.000	15.147	0.000	0.000	0.000	0.000	6.357
Age	0.000	0.000	0.000	0.000	0.000	0.000	-0.298
Annual Income	0.000	-36.744	-36.615	-36.615	-36.615	0.000	11.582
Income to Poverty	29.598	43.174	43.339	43.339	43.339	0.000	5.943
Number in HH	78.576	74.226	71.310	71.310	71.310	45.803	25.241
Children Under Five	0.000	-9.426	0.000	0.000	0.000	0.000	10.535
Children Under 16	0.000	22.945	25.130	25.130	25.130	1.861	26.241
Money Other Store	0.000	-3.597	-3.731	-3.731	-3.731	0.000	-0.740
Money Eating Out	0.000	7.571	7.424	7.424	7.424	2.240	3.550
Money Delivery	0.000	-1.553	0.000	0.000	0.000	0.000	0.757
Money Nonfood	16.177	15.631	15.564	15.564	15.564	8.541	7.411
Not Prep Home	-26.651	-40.292	0.000	0.000	0.000	0.000	-14.268
Ready to Eat	0.000	0.000	0.000	0.000	0.000	0.000	-0.008
Frozen Meals	0.000	-7.741	0.000	0.000	0.000	0.000	-4.359
Fast Food	0.000	0.000	0.000	0.000	0.000	0.000	-2.691
Food Didn't Last	0.000	0.000	0.000	0.000	0.000	0.000	2.264
Calories	0.001	-0.054	0.000	0.000	0.000	0.000	-0.003
Protein	0.000	0.615	0.000	0.000	0.000	0.000	0.047
Carbohydrates	0.000	0.288	0.000	0.000	0.000	0.000	0.010
Sugar	0.000	0.000	0.000	0.000	0.000	0.000	0.006
Fiber	0.000	0.000	0.000	0.000	0.000	0.000	0.413
Physical Activity	0.000	0.000	0.000	0.000	0.000	0.000	-3.328

Many of the model selection techniques chose similar models. All of the models created using BIC had the same final model, so the model for forward stepwise selection, backward stepwise selection, and the second best subsets model were all the same. The best subsets model that was chosen using AIC contained more than twice as many predictors as the models that were chosen using BIC. This makes sense as BIC puts a harsh penalty on models with a large number of predictors. The LASSO model contained a subset of the predictors from the models selected using BIC.

The model created using prior judgement appears to be a relatively good model. Four of the five predictors included in this model were also included in the models chosen using BIC. The coefficients for many of the important predictor variables that appear in most of the models appear to be similar across models that were not chosen using regularization methods. For example, number of people in the household has coefficient values ranging from 71 to 78.5 in the models not chosen using regularization and money spent on nonfood items has coefficients ranging from 15.5 to 16.2 (Table 4).

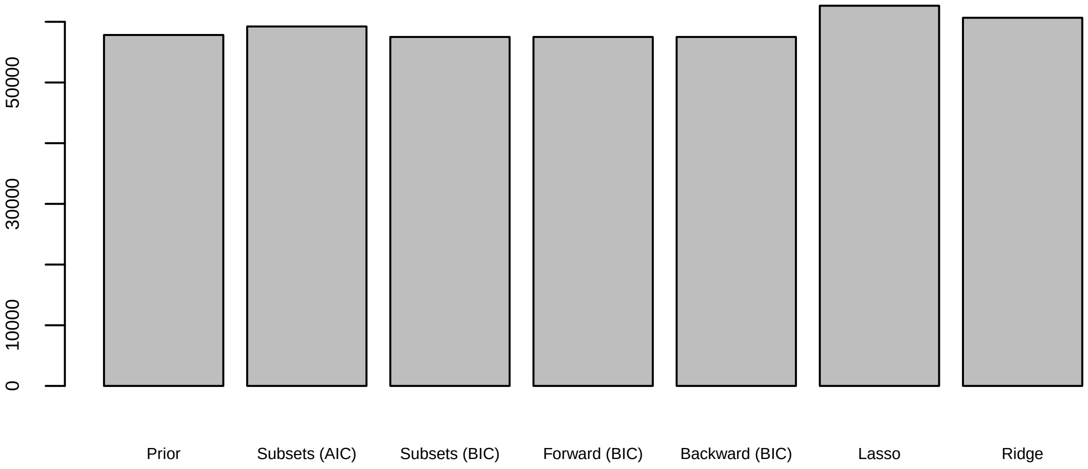

Figure 5. Test MSE for each model

The test dataset was used to compare predictive power among models. Test MSE gives us a way to compare true values in the untouched test dataset to the values predicted by the model fitted to the training dataset. The model chosen by best subsets selection and stepwise regression using BIC has the lowest test MSE of any of the models, followed closely by the prior judgement model and the AIC best subsets model. The ridge regression has a slightly larger MSE than the AIC best subsets model and the LASSO model has a slightly larger MSE than the LASSO model. If we were looking to choose the model with the best prediction power, we would chose the model that was selected using stepwise regresssion and best subsets with BIC as the selection criteria (Figure 5).

If we were looking for a model that was easily interpretable while still having good prediction power, most the modesl aside form the ridge regression model would work. The LASSO model might be less interpretable than the other models. While it does contain predictors that we would logically expect to have a strong relationship with the outcome variable, interpretation would be more difficult if we had to explain regularization methods to the audience. If one model had to be chosen, the model based on prior judgement would be best because it contains the fewest predictors of all the models while stil having similar predictive value s.

Discussion

There are a few limitations to our study that we would like to acknowledge. The first possible limitation is related to the method we used to partition or data into test and train sets. We used a random split to partition our data where. As such, our test data set and train data set were very similar. It may have been better to develop some kind of a nonrandom split or to look for data from a different study. If one of these other methods were used, the test MSE may have given a more accurate estimate of how each model preforms using unseen data.

The second possible limitation relates more to the first part of the of the study when nonlinear modeling methods were used to fit relationships between predictor variables and the response variable. Most of the predictor variables in our dataset had fairly linear relationships with the response variable. The first portion of the analysis may have been more interesting if a dataset with interesting nonlinear relationships was chosen.

The final limitation we will mention is related to our handling of missing data. The assumption that all missing values are missing completely at random is just not true. Even if it were true, we were still excluding a lot of potentionaly useful observations from our data due to missingness. It may have been better use methods that have been developed to deal with missing data in our analysis.