University of North Carolina Asheville
Journal of Undergraduate Research
Asheville, North Carolina
May 2020

Predicting Falls in Older People with the LASSO Method

Linna Brooks, Dino Dziergas
Mathematics-Statistics
The University of North Carolina Asheville
One University Heights
Asheville, North Carolina 28804 USA

Faculty Advisor: Jimin Lee

Abstract

Survival analysis can be used to estimate the lifespan of a particular population that is being studied. It can be used in a large number of fields that include, but are not limited to medicine, public health, biology, engineering, and marketing. Depending on if the individual has experienced the event of interest or not, the data can be censored or uncensored. It is important to be aware of this when conducting analysis. The Kaplan-Meier estimate incorporates this and shows the probability of surviving in a given length of time over many small time intervals. The data set used contains 301 individuals between the ages of sixty-five and ninety-six years old to predict the time until their first and second falls. Information such as walking speed, stride length, use of assistive devices, previous falls, and gate smoothness were used as explanatory variables. LASSO feature selection was used to reduce the dimensionality of the data set and reduce the amount of non-significant variables from the prediction. Once the variables were selected, a cox proportional hazard model was conducted.

1. Introduction

Tibshirani[4] introduced the least absolute shrinkage and selection operator, LASSO, method in 1996. LASSO contains properties of subset selection and ridge regression. Subset selection and ridge regression were used to improve ordinary least squares regression. Unfortunately, these methods have some drawbacks which include being extremely variable, reducing prediction accuracy, not setting coefficients in the model to zero, and not providing an easy model to interpret. As a result from this analysis, Tibshirani proposed the LASSO method. The LASSO method attempted to keep the better elements of subset selection and ridge regression while also shrinking some coefficients and setting others to zero. The introduction of this method provided a more accurate technique to coefficient selection. The LASSO method also avoided the explicit use of ordinary least squares estimates making it so the model does not suffer if the ordinary least square estimates behave poorly. The LASSO can change the signs of coefficients compared to least squares estimates. It can be difficult to obtain standard errors for the LASSO method due to the fact that it is non-linear and non-differentiable. The bootstrap method is one approach that can be used to assist this. Tibshirani conducted a brief example looking at prostate cancer data to compare the LASSO method, ridge regression, and subset regression. For the example conducted in the paper, the LASSO method and subset regression gave non-zero coefficients to the same three predictors. The paper also performed other simulations to compare LASSO to the non-negative garotte, best subset selection, and ridge regression. Using LASSO, the examples consistently performed well, where the other methods suffered in one way or another.

Tibshirani[4] looked at the LASSO method in a linear regression context. In the 1997 paper by Tibshirani[5], the LASSO method was applied to be used for variable selection in the Cox proportional hazard model, a model commonly used in survival analysis. Depending on a constraint, predictor variables are shrunk to zero making for a more interpretable final model. Due to the constraint having a smooth form the final model is more stable than stepwise and subset selection. In the example in the paper looking at lung cancer data, the LASSO method accurately predicted the

variables that had the dominant effect on the data being analyzed and provided a more accurate model compared to the stepwise method. In a simulation study with a few large effects the LASSO method outperformed the stepwise selection, picking the correct number of zero coefficients. In a simulation study with many small effects the LASSO method still outperformed the stepwise method.

Datta, Le-Rademacher, and Datta[2] used the LASSO method in their analysis of survival times of cancer patients. The paper performed a comparison of LASSO and partial least squares (PLS). It is pointed out that LASSO can be used for linear regression and for Cox's regression model with survival data. The use of microarray data in the study allows for analysis of the LASSO and PLS when the sample size is small compared to the number of covariates. Investigation of how to incorporate right censoring when performing survival analysis is also addressed in the paper. It is restated that the LASSO method fits a linear model via minimization of the error sum of squares that is subjected to a constraint specified by the user. Since it shrinks some coefficients in the model to zero it is a good method for variable selection. From the papers study, it was found that the LASSO method was more effective in selecting predictors that better explain the data than PLS.

LASSO is a valid method to build a model that best represents a dataset through its variable selection method. Being able to make the model easier to interpret and more accurate is the main reason why variable selection is used. Since the LASSO method puts a constraint on the sum of the absolute value of the model parameters and penalizes the coefficients that exceed the constraint, shrinking them to zero. The variables that have non-zero coefficients after shrinking are used, aiming to minimize prediction error. Eliminating irrelevant variables from a model that are not associated with the response variable reduces overfitting and makes the model easier to intercept. The LASSO method helps create a model with the most relevant variables in it.

In Schooten, et. al[3] principal component analysis (PCA) was used as a data reduction technique. It is important to note that the paper dealt with missing values in the data collection by inserting mean values, which resulted in a total of 301 participants with gait quality characteristics. Results showed 18 principal components which met their minimum criteria of an eigenvalue equal or greater to 1. Those 18 components explained 80.5% of the variance in the data. These results were rotated using a varimax rotation with Kaiser normalization. The principal components were then named based on the variables that weighed the heaviest on each factor. The factors were used as input for multivariate accelerated failure time (AFT) models with the response variables time-to-first fall and time-to-second fall. The AFT used the factor scores, which were introduced stepwise to the model with forward selection, until they ceased to contribute significantly at a significance level of 0.05.

2. Methodology

Feature selection is commonly used to build a model that best represents a dataset. Figuring out the most important features to describe the response variable is the first thing that should be done when conducting analysis. Feature selection determines a reduced number of independent variables that describe the dependent variable. Feature selection simplifies models making it easier to interpret, shortens data training times, and reduces overfitting. There are many different feature selection methods including LASSO, principal component analysis, and stepwise selection. The more variables included in a dataset, the more important feature selection becomes. It is difficult to determine which variables should be included in analysis when there are more variables. There are three types of feature selection methods; filter methods, wrapper methods, and embedded methods. LASSO is an embedded method.

The LASSO method puts a restraint on the sum of the absolute values of the model parameters and checks if the sum is less than a fixed value. The variables that are not less than the fixed value are shrunk to zero. Variables that are not shrunken to zero are used in the model. There is a tuning parameter, commonly referred to as λ, that controls the strength of the penalty term. When λ is zero, no variables are set to zero. The larger λ gets, the more variables are shrunk to zero.

Assuming that there is data $(x^i, y_i), i = 1, 2, \dots, N$, where $x^i = (x_{i1}, \dots, x_{ip})^T$ are the dependent or predictor variables and y^i is the response variable. Letting $\hat{\beta} = (\widehat{\beta_1}, \dots, \widehat{\beta_p})^T$, the objective of the LASSO is to solve equation (1).

$$\min \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum \beta_j x_{i,j} \right)^2 \right\} \text{ subject to } \sum_j |\beta_j| \leq t \qquad (1)$$

Where t is the upper bound for the sum of the coefficients. Letting X be the covariate matrix so $X_{i,j} = (x_i)_j$ and x_i^T is the ith row of X it can be written as equation (2).

$$\min \left\{ \frac{1}{N} \|y - \beta_0 1_N - X\beta\|_2^2 \right\} \text{ subject to } \|\beta\|_1 \leq t \tag{2}$$

Where $\|\beta\|_p = (\sum_{i=1}^{N} |\beta_i|^p)^{1/p}$ and 1_N is a N by one vector of ones. Since it is standard to work with variables that have been centered and covariates that are standardized, the formula can be rewritten as equation (3).

$$\min \left\{ \frac{1}{N} \|y - X\beta\|_2^2 \right\} \text{ subject to } \|\beta\|_1 \leq t \tag{3}$$

Finally, the Lagrangian form is written as equation (4).

$$\min \left\{ \frac{1}{N} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\} \tag{4}$$

The relationship between λ and t is inversed. As t goes to infinity, λ becomes zero and vise versa. When the optimization problem, equation (4), is minimized, some coefficients in the model are shrunk to zero, $\widehat{\beta}_J(\lambda) = 0$. The variables whose coefficients equal to zero are not included in the model.

The R package **glmnet** is used in the analysis. The glmnet package provides procedures for fitting the entire LASSO regularization path for linear regression models. The algorithm uses cyclical coordinate descent to find the minimum.

3. Data

Data from Schooten, et. al[3] was used in this analysis. The data consists of participants over the age of sixty-five with the oldest being ninety-six. This age group is important to consider in prediction of falls because they are the most at risk of falls associated with mortality. The data was collected using three methods: questionnaires, wearable accelerometers, and physical assessments. The accelerometer was to be worn at all times during the day throughout eight consecutive days to accurately measure physical inactivity from the participant. Things such as transportation, non-wearing, laying down, and sitting down were identified with the accelerometer manufacturers algorithm to eliminate false reads. These were averaged over the eight days to help identify daily physical activity for the participants.

Participants were found between March 2011 and January 2014 in Amsterdam. To be eligible for the study, participants had to be between the ages of sixty-five to ninety-six. They would also need to have the ability to walk a minimum of twenty meters with assisted devices if needed and have a mini mental state examination where they received a score between nineteen and thirty. The participants were recruited from surrounding general practitioners, pharmacies, residential care centers, training groups, and hospitals were given written informed consent. All protocols were approved by the medical ethical committee of the Vrije Universiteit Medical Hospital.

Accelerometer data for gait was logged for stride frequency, walking speed, and step length bases on ten second or longer bouts. Variability in stride frequency, stride length, and waling speed were measured as well as autocorrelation at the dominant period in the anteroposterior (AP), mediolateral (ML), and vertical (VT) direction of acceleration. Gait symmetry was logged at the harmonic ratio and gait smoothness was kept as an index of harmonicity. Other variables logged in all three directions of acceleration were width and magnitude of the dominant period and percentage of power below 0.7Hz.

Data collected from questionnaires and examinations were obtained during the visits regarding the accelerometer. Things such as age, gender, weight, height, and use of assisted walking devices like a cane were gathered. Validated questionnaires covered characteristics of fall risk that may have an effect on future fall change. Cognitive function, internal fear of having a fall, LASA risk profile, and depressive symptoms were measured. The risk profile consisted

of questions relating to education, independence in daily life, owning pets, frequent dizziness, alcohol use, and grip strength which was measured with a dynamometer[3].

The data had a fair amount of missing data points spread throughout with only 136 of the 301 total observations having full data collection. The data consisted of 153 female participants and 148 male participants. To deal with the missing data, mean values for each of the collected numerical variables were calculated and missing values were replaced by the averages based on gender. This is similar to how the missing data was dealt with in Schooten, et. al[3].

4. Results

To begin analysis, categorical variables for time until fall in months and whether the individual was censored were added to the dataset. To find the explanatory variables that are most relevant to predicting the response variable, time until a fall, the LASSO method was used. The R package **glmnet** was used to determine these variables. There were seventy-four total possible explanatory variables in the dataset. These variables were placed in a matrix while the dependent variable, time, was placed in a vector. The function glmnet was then used. This function fits a generalized linear model via penalized maximum likelihood. A survival object was then created for an input to the function. The glmnet package is able to perform LASSO regression and ridge regression, LASSO regression is what was used.

Running this code returns a sequence of different models corresponding to different values of λ. Figure 1 below shows a plot of the results from the function. Different values of λ are on the x-axis of the plot. The lines in the plot represent each of the seventy-four explanatory variables. Due to having seventy-four explanatory variables the plot is slightly difficult to read, but the plot provides a visual of which variables influence the response variable and to what extent. Analyzing Figure 1 the ninth, eighth, and fourth variables clearly have influence over time until a fall. These variables are, respectively, stride frequency, walking speed variability, and stride length variability. These variables are not necessarily going to be selected in the LASSO method though. Stride length variability steadily negatively affects hazard rate of failure (falls). Stride frequency and walking speed variability positively affect hazard rate of failure. The rest of the variables do not appear to have significant impacts on the model. This is hard to see though because of the high number of variables in the model.

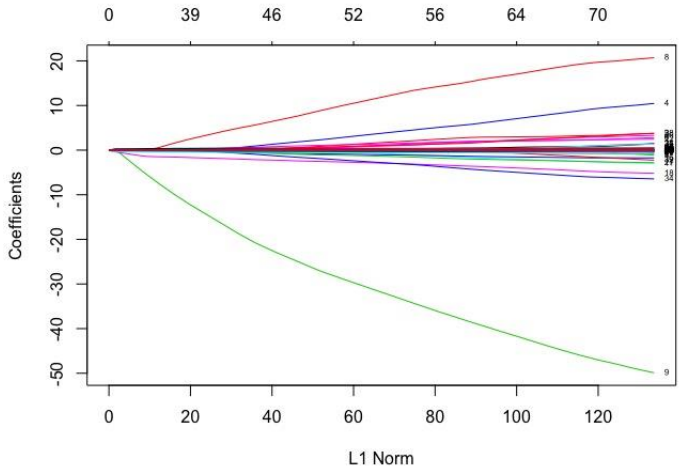

Figure 1. Glmnet of all explanatory variables.

Next, the function **cv.glmnet** was used to perform k-fold cross-validation for glmnet. It produces a plot and a value for λ that could be used for LASSO variable selection. The next goal is to pick a value of λ to use in the model. The inputs for this function are very similar to the inputs for the glmnet function. All the explanatory variables are taken in as a matrix, a survival object, and alpha of one indicating LASSO feature selection is to be used are the inputs for the function. The x-axis of the plot is the log of λ. The numbers at the top of the plot state how many explanatory variables are included in the model. The red dot's position along the y-axis tell the area under the curve (AUC) calculated when including the number of variables shown on the top of the x-axis. The AUC is used to help determine the best model for predicting time until a fall. The dashed vertical line on the plot indicates the lambda with the lowest mean squared error. This is the lambda most commonly used when performing LASSO feature selection.

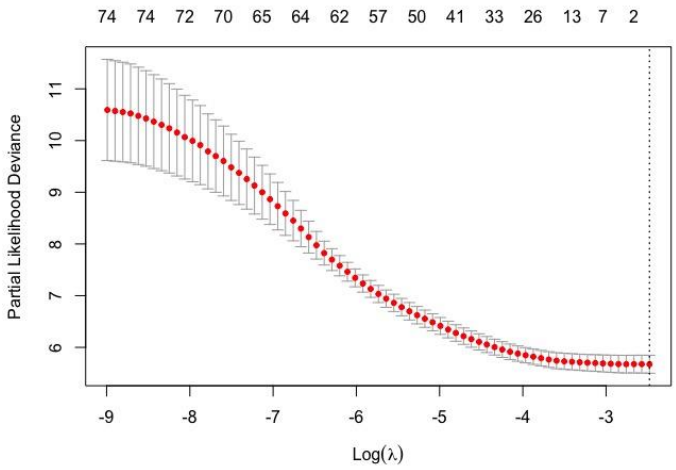

Figure 2. Cross-validation plot for λ.

The λ that resulted in the lowest mean squared error was 0.08383409. Now that a value of λ has been recommended, coefficients can be chosen with that λ. When a λ of 0.08383409 was used no variables were selected. Furthermore, analyzing Figure 1 and Figure 2 indicate that there is not much of a difference in the mean squared error with a λ selected equal to 0.04 versus the minimum λ. Using 0.04 as the selected λ, eleven variables were selected for the model. Those variables include root mean square ML, magnitude of dominant period in frequency domain VT, mean logarithmic rate of divergence per stride AP, categorical male gender, number of fall in the past six months, categorical if the individual frequently experiences dizziness, total LASA fall risk profile score, time on TMT, categorical use of a walking aid, median number of strides per locomotion bout, and duration of lying down. Figure 3 is a recreation of Figure 1 with the eleven selected variables.

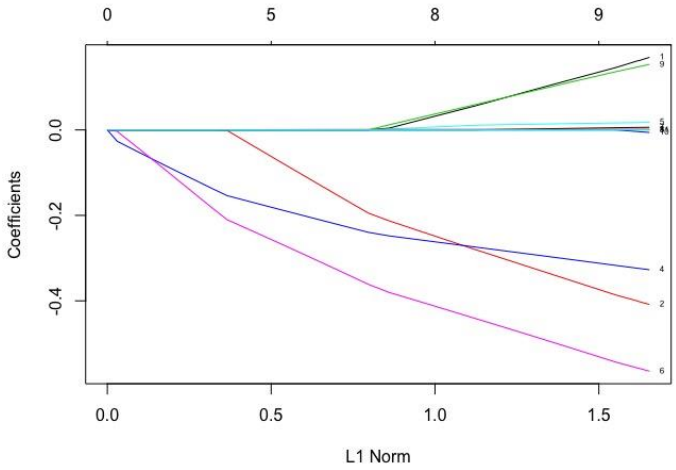

Figure 3. Glmnet of selected explanatory variables.

Frequently experiencing dizziness had the largest negative effects on hazard rate of failure along with being a male and the magnitude of dominant period in frequency domain VT. Use of a walking aid and the root mean square ML had the largest positive effects on hazard rate of failure. The remaining variables had a less significant impact on time until a fall. Since a cox proportional hazard model was used as the input to the function, a negative coefficient implies a better survival probability. So being a male and frequently experiencing dizziness means those individuals will have a higher survival probability.

After the eleven variables were selected, a correlation matrix was created to analyze the selected variables closer and better understand the data. Looking at the correlation coefficients between the selected variables the relationship between each other and time can be seen along with if it the relationship is positive or negative. Table 1 is the resulting correlation matrix with time until a fall and the eleven selected explanatory variables.

Table 1. Correlation Matrix for Time Until a Fall and Eleven Selected Variables

	TIME	RMS	MVT	MAP	MG	F6M	FED	LASA	TMT	WA	MSLB	DL
TIME	1											
RMS	-0.044	1										
MVT	0.068	0.23	1									
MAP	-0.033	-0.134	-0.649	1								
MG	0.094	-0.065	-0.048	0.096	1							
F6M	-0.063	-0.037	-0.093	0.001	-0.004	1						
FED	0.088	-0.071	0.006	0.004	-0.115	-0.1	1					
LASA	-0.031	-0.089	-0.087	0.042	0.024	0.515	-0.054	1				
TMT	-0.06	-0.116	-0.091	0.126	0.03	0.018	-0.031	0.006	1			
WA	-0.034	-0.081	-0.076	0.112	-0.037	-0.132	0.284	-0.111	-0.01	1		
MSLB	-0.024	0.261	0.104	-0.128	-0.091	-0.128	0.093	-0.124	-0.055	0.047	1	
DL	0.071	-0.036	-0.094	0.081	0.025	-0.036	-0.023	0.095	0.054	0.006	0.143	1

While almost all of the correlations are low, some of the highest correlations between time until a fall and the explanatory variables belong to the variables that had the largest effects on hazard rate of failure in Figure 3. Frequently experiencing dizziness and male gender are the two variables that are the highest correlated with time until a fall of the eleven variables selected.

Now that explanatory variables have been selected, survival analysis can be conducted. To start, a Cox proportional hazards regression model was fit with the selected variables. The R package **survival** was used with the coxph function. The function takes a formula with a response variable and explanatory variables. To conduct further survival analysis the packages **survival** and **survminer** were used. The function survfit creates a survival curve, also known as a Kaplan-Meier curve, from a previously fitted Cox model. A survival curve shows the proportion of the population in the data that survive until a point in time. The y-axis gives the proportion of individuals surviving ranging from one, everyone survived or one hundred percent survival probability, to zero, no one survived or zero percent survival probability. The x-axis gives the time after the start of observation. In this research that is zero to twelve months. A survival curve always starts with a one hundred percent survival percentage at time zero. It then decreases or remains the same. A survival curve can never increase. Survival curves are not smooth curves typically, they are curves with "steps" down each time there is a failure at the observation time. The survival probability is calculated by the number of individuals that survived divided by the number of individuals that are at risk. Individuals that have already had the event occur, dropped out, or have not reached that time interval yet are not counted as at risk. The precision of the estimates depends on the number of observations. The estimates on the left-hand side are more accurate than those on the right-hand side. Figure 4 is a survival curve for this research's model.

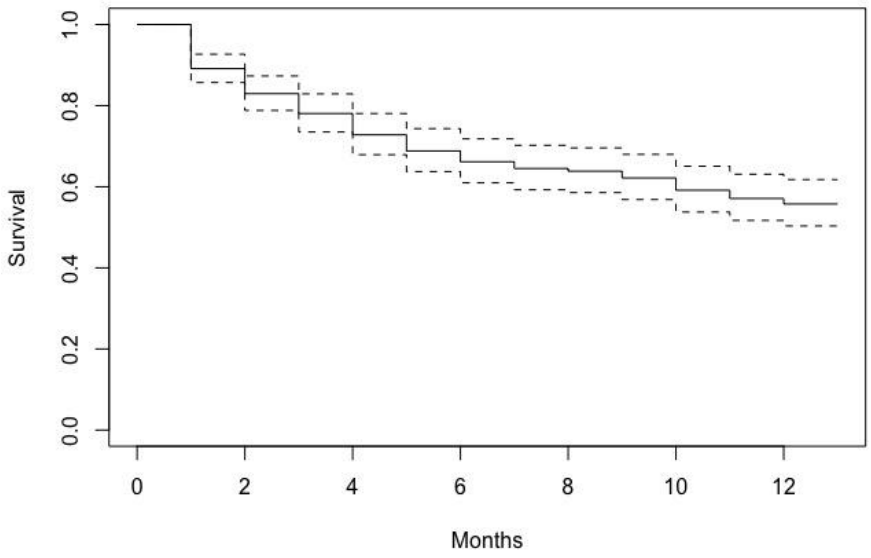

Figure 4. Survival curve.

The solid line illustrates the predicted survival curve based on the model. The dotted lines above and below the survival curve is a ninety-five percent confidence interval for the curve. Table 2 includes the survival probabilities for this survival curve along with the number of individuals at risk and number of individuals that experienced a fall at each time interval. The ninety-five percent confidence interval is also included. After one month, thirty-four individual experienced a fall. The survival probability was 89.1%. After six months, there were 104 total falls. The survival probability was 66.2%. After twelve months, there were 135 total falls. The survival probability was 55.8%. This means that 55.8% of individuals in the study did not experience a fall or dropped out. These individuals are censored.

Table 2. Survival Probabilities

Time	# at Risk	# of Falls	Survival %	Std. Err.	L 95% CI	U 95% CI
1	301	34	0.891	0.0178	0.857	0.927
2	267	19	0.830	0.0217	0.788	0.873
3	248	15	0.781	0.0240	0.735	0.829
4	233	16	0.728	0.0259	0.679	0.781
5	217	12	0.688	0.0270	0.637	0.743
6	205	8	0.662	0.0276	0.610	0.718
7	197	5	0.645	0.0280	0.592	0.702
8	192	2	0.638	0.0281	0.586	0.696
9	190	5	0.622	0.0284	0.568	0.680
10	185	9	0.591	0.0288	0.538	0.651
11	176	6	0.571	0.0290	0.517	0.631
12	170	4	0.558	0.0291	0.503	0.618

Due to categorical male gender being a selected variable by the LASSO method, separate survival curves for males and females is interesting to analyze. Furthermore, separating high and low gait-quality within males and females gives ability for in depth analysis of survival probabilities. Figure 5 shows four survival curves, two for females separating high and low gait-quality and two for males separating high and low gait-quality. Not surprisingly, males

and females with a high gait-quality have higher survival probabilities than those with low gait-quality. Males with high gait-quality have the highest survival probabilities for the four groups. This suggests that females with high gait-quality are more susceptible to falls than males with high gait-quality. This is seen by the curve for high gait-quality males being above that for high gait-quality females. After twelve months, high gait-quality males had a survival probability of 73.2% where high gait-quality females had a survival probability of 64.3%. The same pattern follows for males and females with low gait-quality. Low gait-quality females are more susceptible to falls than low gait-quality males. After twelve months, low gait-quality males had a survival probability of 58.9% where low gait-quality females had a survival probability of 47.3%. No matter the gait-quality, males and females see nearly a 10% difference in survival probability after twelve months.

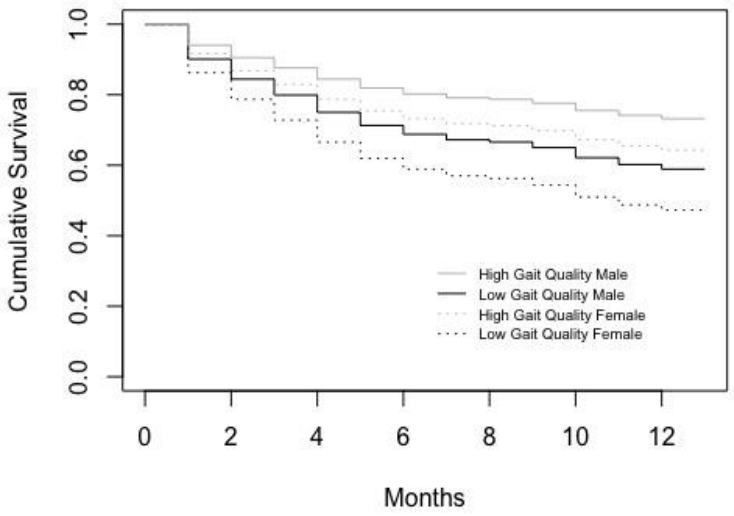

Figure 5. Survival curve separating high gait-quality and low gait-quality.

5. Conclusion

The LASSO method provides a more accurate approach for feature selection. Elements of subset selection and ridge regression are included in LASSO. The method puts a constraint on the sum of the absolute value of the model's parameters and penalizes the coefficients that exceed that constraint. Those coefficients are shrunk to zero and not included in the model. Elimination of variables that are deemed irrelevant reduces overfitting and makes the model stronger. Only relevant variables are kept in the model.

Missing data within the dataset was dealt with by taking the mean values for each of the collected numerical variables and replacing the missing values with the mean based upon gender. Once there was no more missing data, LASSO feature selection was performed. There were seventy-four variables in the dataset and the selected value for λ selected eleven variables to be included in the model. The chosen penalty estimate resulted in no variables selected. A λ equal to 0.04 was then selected. Once the variables were selected and analyzed, survival analysis was conducted. In the future, the research would like to consider employing the LASSO constraint in Principal Component Analysis. An overall survival curve was created using the model with the eleven selected variables. This curve gives information on survival probabilities over time intervals and the number of individuals still at risk of falling. After six months, 66.2% of individuals did not experience a fall. After twelve months, just over half of the individuals included in the study experienced a fall.

With a categorical for male gender being selected as a relevant variable, looking at a survival curve separating males and females was relevant in the analysis. To expand this, males and females were split into high and low gait-quality. The gait-quality classification was determined by finding the median value of each of the non-binary selected features from the LASSO method. The values on the upper end of the median were classified as high gait-quality whereas the

values less than the median were classified as low gait-quality. This allowed for a survival curve separating males and females within high and low gait-quality specifications giving a clearer look at the time until first fall. It was found that males had a higher survival probability than females in the same gait-quality classification. In fact, they were more likely to not have a fall by approximately 10%.

6. Acknowledgements

The authors wish to express their appreciation to Jimin Lee for her assistance conducting this research.

7. References

1. Fonti, Valeria. “Feature Selection Using LASSO.” VU Amsterdam, 2017.
2. Datta, Susmita, Jennifer Le-Rademacher, and Somnath Datta. “Predicting Patient Survival from Microarray Data by Accelerated Failure Time Modeling Using Partial Least Squares and LASSO.” Biometrics 63, no. 1 (March 2007): 259–71. https://doi.org/10.1111/j.1541-0420.2006.00660.x.
3. Schooten, Kimberley S. Van, Mirjam Pijnappels, Sietse M. Rispens, Petra J. M. Elders, Paul Lips, Andreas Daffertshofer, Peter J. Beek, and Jaap H. Van Dieën. “Daily-Life Gait Quality as Predictor of Falls in Older People: A 1-Year Prospective Cohort Study.” Plos One 11, no. 7 (July 7, 2016): 1–13. https://doi.org/10.1371/journal.pone.0158623.
4. Tibshirani, Robert. “Regression Shrinkage and Selection Via the Lasso.” Journal of the Royal Statistical Society: Series B (Methodological) 58, no. 1 (1996): 267–88. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.
5. Tibshirani, Robert. “The Lasso Method For Variable Selection In The Cox Model.” Statistics in Medicine 16, no. 4 (1997): 385–95. https://doi.org/10.1002/(sici)1097-0258(19970228)16:4<385::aid-sim380>3.0.co;2-3.