Regression Analysis

Joseph Francia and Nicholas Saber

Abstract

Least-squares regression is by far the simplest and most well-known form of regression. Why? Well, for one, obtaining the solution to least-squares regression is fairly simple mathematically. One just needs to be able to do some matrix math and derivatives. More importantly, however, the solution to least-squares regression has the smallest variance of all other unbiased estimators. To sum it all up, the solution to least squares regression is unbiased, has minimal variance, and has a nice, clean analytical solution. So why use any other form of regression?

It actually turns out that least-squares regression is usually not optimal if we are interested in predicting values of new observations, given an existing datset we have. Why might this be the case? As we mentioned earlier, least squares regression provides us the estimate with the least variance amongst all other unbiased estimators. However, this unbiased estimator actually has more variance than many other biased estimators. This excess variance in the solution for least-squares regression usually results in least-squares not being the optimal regression to use for prediction. As a result, in this project we are going to explore other forms of regression in order to see which type of regression is the best at prediction.

Introduction

Our intention for this project is to evaluate the predictive power of the following five regressions: Least-Squares regression, Ridge regression, Lasso regression, Principal Component regression, and Partial Least-Squares regression. We will be using a dataset where the response variable is an individual's credit card balance and the explanatory variables are an individual's characteristics (gender, income, age, etc.). Before we ran any sort of regression on the dataset, we first engaged in some exploratory data analysis by making histograms, boxplots, and summary statistics for the variables in our dataset. After this exploratory data analysis, we turned our attention to evaluating the performance of each regression.

We evaluated the performance of each regression by testing to see which

algorithm did the best job of predicting an individual's credit card balance, given an individual's characteristics (age, income, gender, etc.). We used built in cross-validation functionality in R functions in order to divide the data into test and training sets. We then compared the mean squared error between our five regressions. Before we get into the specifics of the regression methodology though, lets first take a look at our dataset.

Data

Before running any of these regressions, we first did some exploratory data analysis (EDA). Because we have both quantitative and qualitative explanatory variables in our dataset, our EDA for each variable would differ, depending on whether or not the variable was quantitative.

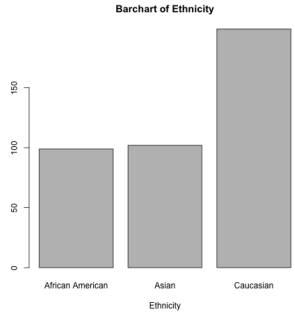
For the quantitative variables, I created images, and we also computed summary statistics. More specifically, we computed each quantitative variable's standard deviation, interquartile range, range, and median. In addition, we also computed a histogram and a boxplot for each quantitative veariable. Finally, I created a correlation plot and a regression object between all the quantitative variables as well. I've shown the correlation plot for the quantitative variables below:

	Income	Limit	Rating	Cards	Age	Education	Balance
Income	1.000	0.792	0.791	-0.018	0.175	-0.028	0.464
Limit	0.792	1.000	0.997	0.010	0.101	-0.024	0.862
Rating	0.791	0.997	1.000	0.053	0.103	-0.030	0.864
Cards	-0.018	0.010	0.053	1.000	0.043	-0.051	0.086
Age	0.175	0.101	0.103	0.043	1.000	0.004	0.002
Education	-0.028	-0.024	-0.030	-0.051	0.004	1.000	-0.008
Balance	0.464	0.862	0.864	0.086	0.002	-0.008	1.000

As for the qualitative variables, I created a different set of images and computed a different set of summary statistics. First, I created a frequency table for each qualitative variable. Next, I created a barchart for each qualitative variable. As an example, I've attached both a barchart and frequency table below for the Ethnicity variable, an explanatory variable.

	Counts	Proportion
African American	99	0.248
Asian	102	0.255
Caucasian	199	0.498

Figure 1: Barchart for Ethnicity Variable



After doing our exploratory data analysis, we divided our data into training and test sets. The idea behind training and test sets is that we will run our regression on the training data set, and we will then test the accuracy of our regressions on our test data set. I did this by first creating a sequence of numbers, ordered 1:400. Then, I used the *sample* function in order to shuffle this ordered sequence. I then took the first 300 values of this shuffled sequence, and I marked these values as indices corresponding to our training set. The final 100 values of this shuffled sequence, meanwhile, corresponded to the indices in our test set.

Methods

After dividing our data into training and test sets, we're now ready to run our regressions. Before we start, however, I want to give a brief theoretical overview of each regression method.

Least squares regression is the most basic form of regression. The solution to least-squares regression minimizes the sum of squared residuals. While this solution is unbiased, it turns out that this solution tends to have larger amounts of variance.

Thus, we turn to modifications of least squares regression: Ridge Regression and Lasso Regression. The solutions to both of these regression problems are biased estimators that have lower variance than the least-squares solution. In Lasso regression, the solution is the beta vector that minimizes both the sum of squared residuals (bias) and the **absolute** norm of the beta vector (model complexity). In Ridge regression, meanwhile, its solution is the beta vector that minimizes both the sum of squared residuals and the **squared** norm of the beta vector. By penalizing and shrinking the length of the beta vector, both Ridge and Lasso Regression are considered shrinkage methods. By finding a solution that minimizes **both** variance (model complexity) and bias (residuals), these biased estimators can perform better from a prediction standpoint than the least-squares estimator.

There are, however, more alternatives to least-squares regression. Lets take a look at Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR). These methods are known as dimension-reduction methods. Unlike shrinkage methods, PLSR and PCR first reduce the column/dimension size of the dataset by summarizing the dataset with fewer columns. The columns of this reduced dataset consist of linear combinations of columns from the original dataset. Then, a model is fit onto these newly formed columns. It turns out that these shrinkage models can actually outperform least-squares regression. The difference between PLSR and PCR is how they go about forming the columns of the reduced/summarized dataset. PCR forms the columns of its reduced dataset by finding linear combinations of its predictors that best represent the original dataset. PLSR, meanwhile, is similar to PCR, but it also incorporates the response variable into its computation. In other words, when forming its summarized dataset, PLSR takes into account how well each predictor column explains the response variable.

Analysis

Once my partner and I felt like we had a solid understanding of these regression methods, we felt like we were ready to run these regressions on our dataset. One of us was going to work on Partial Least Squares Regression, Ordinary Least Squares Regression, and Ridge Regression. Meanwhile, the other would work on Lasso Regression and Principal Component Regression.

Despite the differences between each regression, the formatting and workflow for each regression script remained quite similar. We would begin by first running a specified regression on the training set. We would use built-in cross-validation functionality in R's regression functions to find the optimal regression parameter. Afterwards, we would take our model and our optimal parameter and use them to predict values from our test dataset. We would then calculate the model's mean squared error.

Amidst all of this, we would also save certain images, computations, and R objects that we thought might be informative. For instance, for each regression, we saved a plot, which showed how prediction accuracy varies with a regression's parameter. We also saved each regression's estimated beta coefficients, which came from running a regression and its optimal parameter on the whole dataset.

In order to run all these regressions, my partner and I heavily relied on two key packages in R: **glmnet** and **pls**. The **glmnet** package contained functions that helped with our shrinkage regressions (Lasso and Ridge) while the **pls** package contained functions that helped with our dimension reduction regressions (PLSR and PCR).

Results

Before we look at regression performance, lets take a look at how the estimated regression coefficients differ for each regression:

	Coefficients for PLS Regression	Coefficients for Least Squares	Coefficients for Ridge Regression	Coefficients for PCR	Coefficients for Lasso Regression
Income	-0.602	-0.595	-0.545	-0.598	-0.552
Limit	0.675	1.060	0.674	0.958	0.925
Rating	0.670	0.279	0.614	0.382	0.368
Cards	0.042	0.053	0.043	0.053	0.045
Age	-0.021	-0.007	-0.030	-0.023	-0.017
Education	-0.007	-0.012	-0.002	-0.007	0.000
GenderFemale	-0.047	-0.010	-0.037	-0.023	0.000
StudentYes	0.902	0.920	0.752	0.926	0.888
MarriedYes	-0.021	0.000	-0.055	-0.019	0.000
EthnicityAsian	0.068	0.026	0.000	0.037	0.000
Ethnicity Caucasian	0.004	0.013	-0.022	0.022	0.000

As one can see with the above table and plot, Lasso Regression provides us a sparse estimator. In other words, Lasso Regression provides a solution in which multiple beta coefficients are zero. Although Ridge does not provide a sparse solution, it does tend to shrink the beta coefficients towards zero. This is why both Lasso and Ridge are called shrinkage methods.

Thanks to the above visual, it's also easy to tell that there are four variables that have a strong relationship with the response variable (Credit Card Balance). These variables have beta coefficients that are large in magnitude and they are Income, Limit, Rating, and Student. The Limit, Rating, and Student variables are strongly positively correlated with Credit Card Balance while the Income variable is significantly negatively correlated with Credit Card Balance.

Now that we have a better understanding of each of these regression methods, lets look at the data in order to see how each regression performed:

Figure 2: plot of chunk unnamed-chunk-2

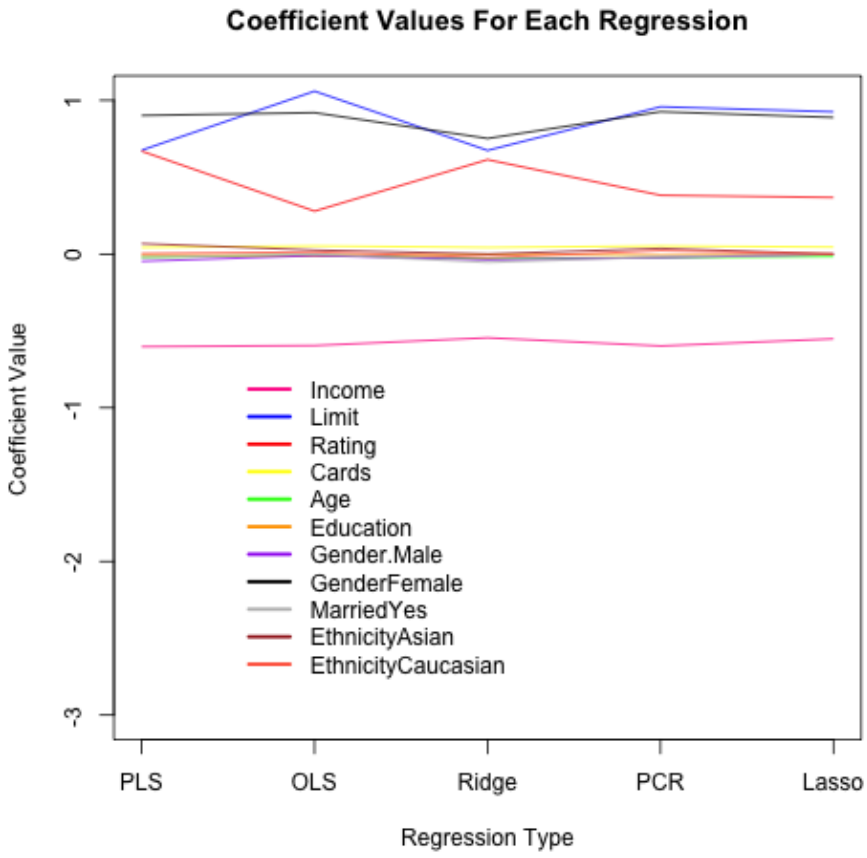


Figure 3: Coefficient Plot

Conclusion

As one can see in the MSE table, least squares is the best performing regression from a prediction perspective. While least-squares regression seems like it would be the best regression method to use for prediction purposes, this is probably only the case for this dataset. Ridge and Lasso, thanks to their penalty on model complexity, actually usually outperform least-squares prediction in prediction. As for the dimension-reduction regression methods, this table indicates that principal component regression is significantly inferior to partial least squares regression. This is likely because, unlike principal component regression, partial least squares regression takes into account how each explanatory variable interacts with the response variable.

In short, while it's surprising how well least squares performed in predicting the Balance variable, least squares' excellent performance may only be particular to this dataset. We would recommend that in future prediction projects, aspiring data-scientists use a variety of regressions for prediction in order to see which one is optimal.