OCCIDENTAL COLLEGE
COMPUTER SCIENCE DEPARTMENT

Computer Science 349: Machine Learning
Project No. 2: Los Angeles Bike Share

Dr. Kathryn Leonard

Shasta Clokey

Los, Angeles, CA, October 3, 2017

Description of Data and Research Question

The goal of this research inquiry is to convince the Los Angeles city council that a bike share program implemented in this city would be exceedingly popular. To do this I will analyze linear regressions on data from a bike share program in the netherlands to determine which factors(weather,time of year, etc.) affect the number of riders who rent bikes on any given day. Using predictor elimination techniques(which I will describe below), I will identify the set of predictors which give the most accurate forecasts for number of rentals. I will then use this model to determine how many riders will use the bike share program on average Winter, Spring, Summer, and Autumn days, accounting for holidays and non-holidays individually. After showing high predicted levels of use in Summer, Autumn, and Winter, I will suggest a plan for a Los Angeles bike share program that will be highly utilized and financially beneficial for the city.

Multiple Linear Regression

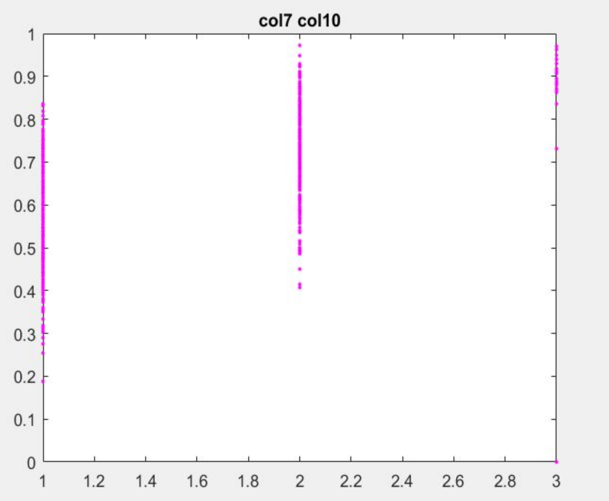
The goal of multiple linear regression is to find a p-dimensional linear correlation between the p many independent variables of a dataset and the dataset's dependent variable. To do this, we subtract each predicted value using the model from the true output of the dataset and add up the square all of these "residuals" to calculate what is appropriately named the "Residual Sum of Squares"(RSS). By varying the coefficients in the multiple linear regression and examining the critical points of the RSS, with a little help from linear matrix algebra we can determine values for coefficients, B1...Bp, which allow the regression model to best fit the given dataset. By looking at one minus the RSS divided by the TSS(rss away from the mean rather than each individual outcome), we can see the proportion of the variance in the model which can be attributed from the predictors(R^2 statistic). Although the RSS and R^2 statistics are accurate metrics to define "fit" in some circumstances, they have a large common flaw. Adding more predictors to a system will always increase the RSS and the R^2. Therefore, we need to be clever and create an adjusted r^2 and RSS which account for the increase in variables to give an accurate representation of "fit". Using this method in combination with the techniques to eliminate extraneous predictors which I will describe below, we will be able to accurately predict the number of bike rentals on any day given measurements for that day such as temperature, and wind speed.

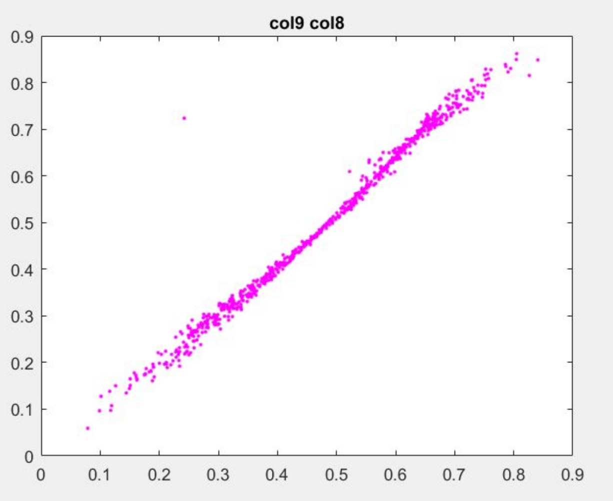
$$\mathbf{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad \mathbf{TSS} = \sum_{I=1}^{n}(y_i - \bar{y})^2 \qquad R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

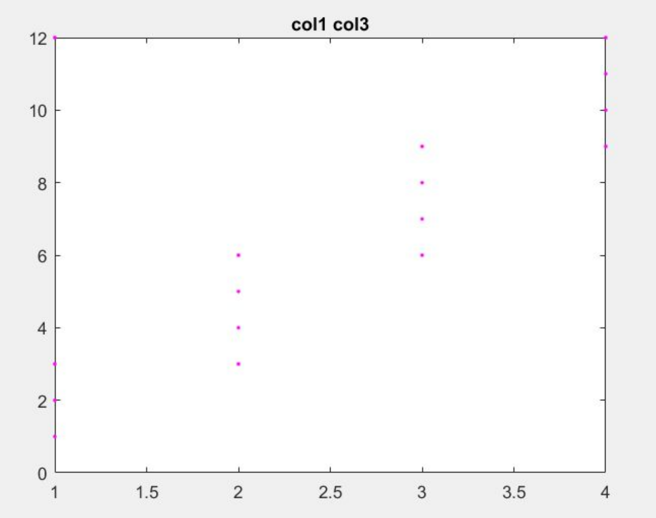
$$\mathrm{Adj}R^2 = 1 - \frac{\mathrm{RSS}/(n-p-1)}{\mathrm{TSS}/(n-1)} \qquad \mathrm{RSE} = \sqrt{\frac{1}{n-p-1}\mathrm{RSS}}$$

Graphical Search for Collinearity

Having two predictors in a dataset that are linearly related causes problems due to redundancy. If two predictors are linearly related, then they will likely predict outcomes based on the same root cause, effectively doubling the predicted value. This is clearly an undesirable result if one is trying to make accurate predictions, so in this section I will discuss how to eliminate collinear features. To best utilize human intuition, I have created visual representations of the predictor data by plotting each set of features against each other set. After examining each of the plots, I have determined that the weather(7)/ humidity(10), temperature(8) and atem(9), and season(1)/ month(3) are each linearly correlated. The linear relationships between these features are visually clear and also logical. It makes sense that humidity would depend on the weather, that two different ways to measure temperature would have a strong linear correlation, and that seasons would be related to months. Before determining which features to eliminate, I will discuss the two other methods I used to find extraneous predictors.



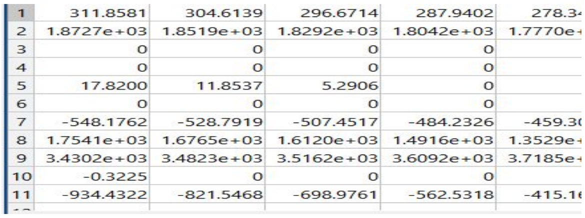




Lasso Technique for Eliminating Predictors

The lasso technique for eliminating extraneous predictors functions by adding on a sum of the absolute value of the regression coefficients all multiplied by a variable lambda. By varying this constant, lambda, we can influence the regression model to make the coefficients of the extraneous predictors go to zero. Using the lasso technique, I tested a range of different values for lambda and examined the values of the coefficients to determine which of them went to zero around the same value of lambda. I found that, by lambda = 150, the month, holiday, weekday, working day, and humidity predictors all went to zero. This made me consider them as potential extraneous features. To verify which of these predictors to eliminate, I tested the feature set using one more Regularization technique.

Lasso: $$\mathrm{RSS} + \lambda \sum_{j=1}^{p} |\beta_j|$$

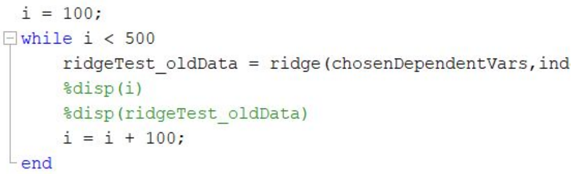


1	311.8581	304.6139	296.6714	287.9402	278.3
2	1.8727e+03	1.8519e+03	1.8292e+03	1.8042e+03	1.7770e
3	0	0	0	0	
4	0	0	0	0	
5	17.8200	11.8537	5.2906	0	
6	0	0	0	0	
7	-548.1762	-528.7919	-507.4517	-484.2326	-459.3
8	1.7541e+03	1.6765e+03	1.6120e+03	1.4916e+03	1.3529e
9	3.4302e+03	3.4823e+03	3.5162e+03	3.6092e+03	3.7185e
10	-0.3225	0	0	0	
11	-934.4322	-821.5468	-698.9761	-562.5318	-415.1

Ridge Technique for Eliminating Predictors

The ridge regression technique ads the sum of the regression coefficients squared all multiplied by a variable lambda to optimize the values of the coefficients in order to provide the most accurate model possible to predict outcomes. By varying lambda, we can cause the value of the less important coefficients to get very small, allowing the important coefficients to dominate in the prediction results. To determine a value of lambda that was reducing the extraneous coefficients and maintaining those which were important, I performed ridge regressions using values of lambda ranging from 100 to 500 and examined the values of the coefficients for each differing value of lambda. I determined that all of the coefficients started to reduce after a value for lambda of about 300, so I chose lambda = 300 to provide adequate distinction between important and extraneous coefficients. Using this lambda value, I determined that the insignificant predictor coefficients according to the ridge regression were month, weekday, and working day.

Ridge regression: $$\mathrm{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2$$



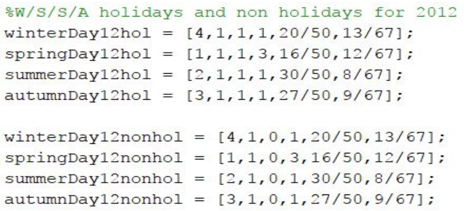
```
i = 100;
while i < 500
    ridgeTest_oldData = ridge(chosenDependentVars,ind
    %disp(i)
    %disp(ridgeTest_oldData)
    i = i + 100;
end
```

Justification of Final Predictor Set

Now that we have used three separate techniques to eliminate extraneous predictors, we can establish a final set of predictors to be used to forecast the number of bike rentals on any given day. Because colinearity is very disruptive in linear regression, I will eliminate predictors humidity(10), temperature(8), and month(3). Because predictors month(3),weekday(5), and working day(6) were found to be extraneous by both ridge and lasso, I will eliminate all of these three predictors. I will not eliminate predictor holiday(4) because it is not reinforced in the ridge regression or the graphical elimination scheme. Ultimately, my final predictor set will include season, year, holiday, weather, atemp, and wind speed as predictors. Using this predictor set, I performed a final multi regression and computed that the adjusted R^2 statistic was .78. Given that this means 78% of the variation in the output is described by the predictors, I have come to the conclusion that this final model can be used to make a judgement on whether or not a bike share program would be successful in Los Angeles. I will use this model to forecast the total number of riders on any given day.

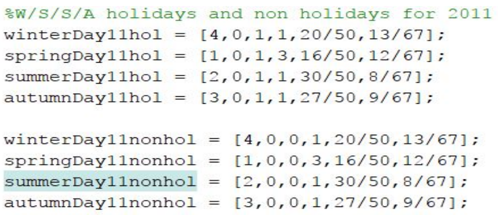
Predictions for Average Seasonal Days

Using my final model, constructed sets of predictor values for average days in Los Angeles at different points in the year to generate a range of predictions which I will be able to use to judge the usefulness of a bike share program in this city. I predicted that on holidays during the winter, approximately 4200 people used the bike share service. I predicted that on holidays during the spring, approximately 1100 people used the bike share service. I predicted that on holidays during the summer, approximately 4800 people used the bike share service. I predicted that on holidays during the autumn, approximately 4800 people used the bike share service. I predicted that on non-holidays during the winter, approximately 5000 people used the bike share service. I predicted that on non-holidays during the spring, approximately 2000 people used the bike share service. I predicted that on non-holidays during the summer, approximately 5500 people used the bike share service. I predicted that on non-holidays during the autumn, approximately 5500 people used the bike share service. Each of these predictions was determined using the following “average” Wi/Sp/Su/Au days for holidays/non-holidays.



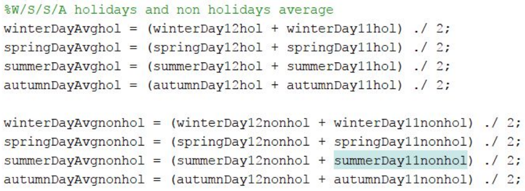
```
%W/S/S/A holidays and non holidays for 2012
winterDay12hol = [4,1,1,1,20/50,13/67];
springDay12hol = [1,1,1,3,16/50,12/67];
summerDay12hol = [2,1,1,1,30/50,8/67];
autumnDay12hol = [3,1,1,1,27/50,9/67];

winterDay12nonhol = [4,1,0,1,20/50,13/67];
springDay12nonhol = [1,1,0,3,16/50,12/67];
summerDay12nonhol = [2,1,0,1,30/50,8/67];
autumnDay12nonhol = [3,1,0,1,27/50,9/67];
```



```
%W/S/S/A holidays and non holidays for 2011
winterDay11hol = [4,0,1,1,20/50,13/67];
springDay11hol = [1,0,1,3,16/50,12/67];
summerDay11hol = [2,0,1,1,30/50,8/67];
autumnDay11hol = [3,0,1,1,27/50,9/67];

winterDay11nonhol = [4,0,0,1,20/50,13/67];
springDay11nonhol = [1,0,0,3,16/50,12/67];
summerDay11nonhol = [2,0,0,1,30/50,8/67];
autumnDay11nonhol = [3,0,0,1,27/50,9/67];
```



```
%W/S/S/A holidays and non holidays average
winterDayAvghol = (winterDay12hol + winterDay11hol) ./ 2;
springDayAvghol = (springDay12hol + springDay11hol) ./ 2;
summerDayAvghol = (summerDay12hol + summerDay11hol) ./ 2;
autumnDayAvghol = (autumnDay12hol + autumnDay11hol) ./ 2;

winterDayAvgnonhol = (winterDay12nonhol + winterDay11nonhol) ./ 2;
springDayAvgnonhol = (springDay12nonhol + springDay11nonhol) ./ 2;
summerDayAvgnonhol = (summerDay12nonhol + summerDay11nonhol) ./ 2;
autumnDayAvgnonhol = (autumnDay12nonhol + autumnDay11nonhol) ./ 2;
```

Argument for Los Angeles Bike Share Program

Given the high volume of projected riders regardless of season, I propose that the city install a self serve, electronically locked bike share system that allows bikers to rent bikes directly from their phones and return them to any of the hub stations which will be strategically placed around each sub-city in Los Angeles. The data clearly shows that we can expect 4000 to 5000 riders every day in every season except spring. If the bike share program charges even as low as a dollar an hour to rent out bikes, this program would generate four to five thousand dollars in passive income for the city every day. Even in the spring, when ridership drops to one to two thousand riders due to poor weather conditions, the average passive income would still generate thousands of dollars in revenue. Because the system would require no paid attendants due to being self-serve, and because the cost of maintenance would be minimal in comparison with the income, the city would be able to profit greatly from a city wide bike share. Discounting the costs of maintenance, the bike share program could generate a yearly income of $1,365,000($4500 for each day in the summer/autumn/winter, $1500 for each day in the spring). This money could be used to build infrastructure, fund educational programs, clean the public beaches, feed the thousands of homeless, or fund any of the other programs on the endless list of Los Angeles city projects. Ultimately, the installation of a bike share program in the city of Los Angeles would be a profitable investment which would also greatly benefit the community.