Chloe Zeller

Machine Learning

Project 2: Bike Sharing

**Introduction**

In this paper, I argue that the city of Los Angeles should consider implementing a bike sharing program. I considered data from a successful bike share program in Washington D.C. [1] Using multiple regression techniques, I created 5 models of the data. My regressions indicate that a bike sharing program would be very successful in Los Angeles. For this reason, I encourage the city of Los Angeles to consider implementing a bike share program.

*Data*

The original dataset that I used spans two years of the bike share program in Washington D.C. It has 11 explanatory variables to describe the given day (e.g. month, temperature, humidity, precipitation etc.) and 3 response variables: the number of bikes that were *casually* checked out on that day, the number of *registered* bike users that checked out bikes on that day, and the *total* number of bike checks on that day.

*Process*

From the original dataset, I derived four different multiple regressions. Then, I created six different data points to describe average weather patterns in Los Angeles. Four of the data points represent average days in the winter, spring, summer and autumn in Los Angeles. I created these by finding average weather patterns during each of the seasons in Los Angeles. For example, to find my winter data point, I chose a day in December, and found the average wind speed in Los Angeles during December, the average humidity, temperature etc. Then, using the model I had created from the original data set, I found a value for the predicted number of bikes checked out, given those conditions. The other two data points were a “peak” day and a “low demand” day. The “peak” day was a day with conditions most suitable for checking out a bike (i.e. holidays, weekends, sunny, warm, low wind speed etc.). The “low demand” day had poor weather conditions (i.e. raining, cold, high humidity), conditions that I guessed would not correlate with many bike checkouts. As with the seasonal data, the peak and low demand data was also compared with the model, and the predicted number of bike checkouts was found.

Not all regressions are equal. I found that using a multiple regression with a lasso regression was the most successful at eliminating non-helpful features, maximizing the adjusted R2, and maximizing the number of bikes checked out. Each regression will be explained in the next section.

**Feature Selection**

We used four different methods for creating a multiple regression of the bike share data. We began by conducting a baseline regression, where all 11 explanatory variables were taken into account. The original explanatory variables were:

* Year (which of the two years the given data was from)
* Month
* Holiday (whether or not it was a holiday)
* Weekday (which day of the week)
* Working Day (whether or not it was a work day)
* Weather Situation (scale from 1-4, 1 being nice weather, sunny and clear, and 4 being precipitation, heavy cloud cover etc.)
* Temperature
* ATemperature
* Humidity
* Wind Speed

We also used the third of the three response variables: total number of bikes checked out on that day. I decided to use this response variable because it is representative of the other two, and if I had chosen one of the other two, I would have not been able to make generalized conclusions about how many bikes would be checked out on a given day.

*Original Regression*

The original regression simply takes into account all 11 explanatory variables, and the one response variable, and forms a regression. From this regression, we found an adjusted R2 of 79.7%. In other words, 79.7% of the variance seen in the data could be explained by the model that was created. I then created the six data points I described above and used my model to predict the number of bikers expected. I calculated the predicted number of bikers for each season, and then averaged them to arrive at a mean number of checkouts of 3170 bikes for any given day. Then, from my peak day and low demand day data, I found the predicted number of checkouts. On a peak day, we would expect around 5883 checkouts per day, and on a low demand day, we would expect 2908 checkouts. These results can be seen in the table below.

|  |  |
| --- | --- |
| **Adjusted R2** | 0.797 |
| **Predicted Number of Checkouts** | 3170 |
| **Predicted Number of Checkouts: Peak Day** | 5883 |
| **Predicted Number of Checkouts: Low Demand Day** | 2908 |

|  |  |
| --- | --- |
| **Confidence Intervals** | **Difference** |
| 997.390241101656 1940.61588807416  402.272202005059 617.278194571320  1912.72720711260 2168.67959620402  -72.5103552126088 -5.44877360923214  -913.688143757391 -124.295718734718  37.0629487383827 101.061483859595  -21.0127726045967 261.726751030619  -764.835252655642 -457.138763562288  -726.866574573370 4784.69878152367  452.877145533006 6693.67143127267  -1635.31848230348 -402.404660133832  -3453.36462255948 -1661.77365317775 | |  | | --- | | 943.225647 | | 215.0059926 | | 255.9523891 | | 67.0615816 | | 789.392425 | | 63.99853512 | | 282.7395236 | | 307.6964891 | | 5511.565356 | | 6240.794286 | | 1232.913822 | | 1791.590969 | |

After I had completed the regression, I plotted each explanatory variable against the residuals that I calculated. This produced 11 graphs, one for each variable. These graphs constitute **Figures 1-6** in the **Appendix**.

*Removing Collinear Features*

Simply looking at the 11 types of explanatory variables above allows us to draw conclusions that at least some of them must be collinear or correlated. To find out which ones were correlated with one another, I plotted each of the explanatory variables against each other, and I looked for incidence of a linear relationship. I did not measure Pearson’s correlation coefficient in measuring correlated features, but rather just relied on my own observation of linear relationships in the graphs I had created. I figured that any linear relationship that was strong enough to be *seen* would be evidence enough to eliminate it, so I decided to not spend time calculating correlation coefficients.

After I plotted each feature with all the other features, I noticed three clear instances of collinearity, which were also clearly intuitively related:

* Season and Month (features 1 and 3)
  + There are a discrete set of months associated with each season, so it makes sense that they were totally correlated.
* Temperature and ATemperature (features 8 and 9)
  + Because Temperature is the temperature in Celsius divided by 41 and ATemperature is the feeling temperature in Celsius divided by 50, the only difference between them is a few degrees and the division factor.
* Weather Situation and Humidity (features 7 and 10)
  + Because weather situation rating was based on cloud cover and precipitation (which are heavily correlated with how humid it is on a given day), it makes complete sense that these two would be correlated.

The plots for these three relationships can be found in the **Appendix** as **Figures 12-14.** Thus, after conducting this analysis, I chose to exclude features 3, 7 and 9 (month, ATemperature and weather situation). After I had reduced the set of explanatory variables, I performed another regression on just the non-collinear features and the response variable. The adjusted R2, mean number of checkouts, and mean number of checkouts on peak and low demand days are recorded in the table below.

|  |  |
| --- | --- |
| **Adjusted R2** | 0.778 |
| **Predicted Number of Checkouts** | 2939 |
| **Predicted Number of Checkouts: Peak Day** | 6265 |
| **Predicted Number of Checkouts: Low Demand Day** | 4427 |

|  |  |
| --- | --- |
| **Confidence Intervals** | **Difference** |
| 1329.01855587196 2269.98605074373  352.078237617236 482.459072167295  1880.18579225543 2147.86405692374  -948.881413283779 -123.900940854032  22.3824580290208 89.0604381579206  -64.5736370974036 230.793866236718  5144.46715588382 5920.12111048091  -3094.16413189903 -2109.91445113170  -4371.22392793653 -2564.15332442024 | |  | | --- | | 940.9674949 | | 130.3808346 | | 267.6782647 | | 824.9804724 | | 66.67798013 | | 295.3675033 | | 775.6539546 | | 984.2496808 | | 1807.070604 | |

It’s relatively surprising that the Adjusted R2 does not increase between the original model and this model, and even more surprising that it actually decreased. This means that less of the variance is explained in this model than the original one, even though we removed features that were confounding. However, this is a very small difference, and it is quite possible that the original model is low bias, high variance, and is over fit to the data to begin with.

That said, there is a higher predicted number of checkouts for the peak and low demand days, but not for the average number of checkouts.

After I had completed the regression, I plotted each explanatory variable against the residuals that I calculated. This produced 8 graphs, one for each variable. These graphs constitute **Figures 15-22** in the **Appendix**.

*Lasso Regression*

*Using Original Explanatory Variables*

Another regression that I wanted to compare to the original regression was one that used a lasso technique for reducing my betas, or the coefficients to my multiple regression. In the first iteration of this technique, I used the entire feature set of all 11 explanatory features. The values for Adjusted R2, and the average expected number of checkouts given each condition can be found in the table below.

|  |  |
| --- | --- |
| **Adjusted R2** | 0.784 |
| **Predicted Number of Checkouts** | 3445 |
| **Predicted Number of Checkouts: Peak Day** | 5845 |
| **Predicted Number of Checkouts: Low Demand Day** | 2680 |

|  |  |
| --- | --- |
| **Confidence Intervals** | **Difference** |
| 686.393702961611 1430.88794575803  330.723273260874 458.410386367779  1934.10788702069 2195.93022689052  49.8114350217031 331.862262226053  -890.907910812744 -647.799072424542  5341.60011734419 6211.09659028756  -2816.78918117423 -1071.47530459850 | |  | | --- | | 744.4942428 | | 127.6871131 | | 261.8223399 | | 282.0508272 | | 243.1088384 | | 869.4964729 | | 1745.313877 | |

Again, our Adjusted R2 is a bit lower than that of the original variable, but not enough to cause alarm. The process for choosing variables to eliminate in this case was a bit more subjective because it was important to not choose a lambda too small (and risk not excluding enough variables, causing low bias, high variance) or a lambda too large (and risk excluding too many variables, causing high bias, low variance). Once the betas were returned from running the regression, I chose a point when five columns had gone to zero, and then found the corresponding lambda used (lambda = 99.3504). The features I chose to exclude were month, holiday, weekday, temperature, and humidity. This would indicate that these features have less to do with whether or not someone is more or less inclined to check out a bike. In my analysis, I also performed *another* lasso regression after my initial lasso, just to see if there were any other features that should be excluded. Using a lambda = 158.19, the next column to go to zero was column 6, or wind speed. I chose to not do another regression excluding this column because it did not go to zero until very late, and it required a relatively high value of lambda, I risked overfitting the model.

After I had completed the regression, I plotted each explanatory variable against the residuals that I calculated. This produced 6 graphs, one for each variable. These graphs constitute **Figures 23-28** in the **Appendix**.

*Using Reduced Non-Collinear Explanatory Variables*

Intuitively, it made sense to perform lasso regression on the set of features that had already been cleaned for collinearity, which I also completed. The process was the exact same as the first lasso regression, just with a different set of explanatory variables.

|  |  |
| --- | --- |
| **Adjusted R2** | 0.784 |
| **Predicted Number of Checkouts** | 3467 |
| **Predicted Number of Checkouts: Peak Day** | 5357 |
| **Predicted Number of Checkouts: Low Demand Day** | 6683 |

|  |  |
| --- | --- |
| **Confidence Intervals** | **Difference** |
| 1856.76809116005 2707.77097675473  561.298711307253 743.222286438376  1997.03270439085 2390.21021309429  -1300.01383218355 -116.979971567074  15.8995201058885 114.540332494855  -4872.62637800911 -2264.69009174897 | |  | | --- | | 851.0028856 | | 181.9235751 | | 393.1775087 | | 1183.033861 | | 98.64081239 | | 2607.936286 | |

The columns that eventually went to zero were columns 5, 6, and 7, which correspond to working day, temperature and humidity. The corresponding lambda was equal to 188.98. I’m hesitant to trust this regression because the average number of expected checkouts in the low demand day is quite high in this case, much higher than it should be. This indicates that something went wrong in the model I created, perhaps the wrong features were chosen, or too many were excluded, maybe a mixture of both. It’s strange because the model still accounts for relatively the same amount or variance as the other models do, it’s just a poor predictor of expected number of checkouts on a low demand day, and is therefore not likely representative of the data we would expect to see in Los Angeles. I decided to take a closer look at the relative slopes of the coefficients I had chosen, to see if there was something obviously wrong with the model. The remaining coefficients were:

* Season = 652.26
  + This is not overwhelmingly surprising. Los Angeles maintains a more temperate climate over the course of the seasons, but it definitely begins the year in colder weather, and it slowly gets warmer.
* Year = 2193.62
  + This is probably the oddest of all the coefficients. It seems that the Washington D.C. data presented a major difference between the two years of the pilot. The positive slope indicates that the bike share program was significantly more popular in its second year, perhaps through word of mouth or increase in popularity.
* Holiday = -708.5
  + This is also relatively surprising, if there I a holiday, bikes are actually less likely to be checked out. This would indicate that at least for Washington D.C. users, they are much more likely to use bikes on workdays, possibly for a commute.
* Weekday = 65.22
  + As the week continues and you approach the weekend, it makes sense that more bikes would be rented out. It’s a bit surprising, however, that this coefficient was this useful, I hypothesized that it would have been excluded much earlier on.
* Wind Speed = -3568.66
  + This is highly intuitive, as wind speed increases, people will not want to go out biking.

Suffice it to say, this result is still confusing, and currently not attributable to anything in particular. Additional probing could be helpful, especially testing how flipping binary coded features will affect the value (i.e. swap the value of holiday from 0 to 1 or vice versa, depending on the particular variable).

After I had completed this new lasso regression, I plotted each explanatory variable against the residuals that I calculated. This produced 5 graphs, one for each variable. These graphs constitute **Figures 29-33** in the **Appendix**.

*Ridge Regression*

*Using Original Explanatory Variables*

In order to figure out which features were to be removed using ridge regression, there was a much more involved process. I tested values for lambda between 100 and 400, incrementing 100 each time. I tracked how the betas changed as lambda increased. The betas that decreased the most between increasing lambdas were eventually excluded from the data set. Setting lambda equal to 400 was the most fruitful in this process. The features that I removed were year, weekday, and working day (2, 5 and 6). I also considered the p values associated with each of these features, which I found in the original multiple regression. The values found for R2 and the average number of predicted checkouts are in the table below.

|  |  |
| --- | --- |
| **Adjusted R2** | 0.518 |
| **Predicted Number of Checkouts** | 4069 |
| **Predicted Number of Checkouts: Peak Day** | 4911 |
| **Predicted Number of Checkouts: Low Demand Day** | 2023 |

|  |  |
| --- | --- |
| **Confidence Intervals** | **Difference** |
| 2662.91057460703 4007.47510166382  312.477372029777 643.793567492340  -79.0460341577462 24.2466967387631  -1219.88497003032 -47.1105802861937  -699.694686019262 -228.322032425926  -1546.62164268590 6936.50284026176  -1490.85256339698 8117.05309242738  -3292.01909940603 -1414.31727064793  -4577.84718605549 -1819.76223876168 | |  | | --- | | 1344.564527 | | 331.3161955 | | 103.2927309 | | 1172.77439 | | 471.3726536 | | 8483.124483 | | 9607.905656 | | 1877.701829 | | 2758.084947 | |

It’s clear that ridge regression is not the best way to estimate which betas should be eliminated by reading the Adjusted R2 value for this regression. It accounts for 20% less variance in the data than any of the other regressions account for. This indicates that the wrong variables were excluded. Removing these particular variables was actually detrimental to the model’s ability to describe the data. It could also be that ridge regression is just not the best method for reducing the feature space.

Per usual, after I had completed the regression, I plotted each explanatory variable against the residuals that I calculated. This produced 8 graphs, one for each variable. These graphs constitute **Figures 34-41** in the **Appendix**.

*Using Reduced Non-Collinear Explanatory Variables*

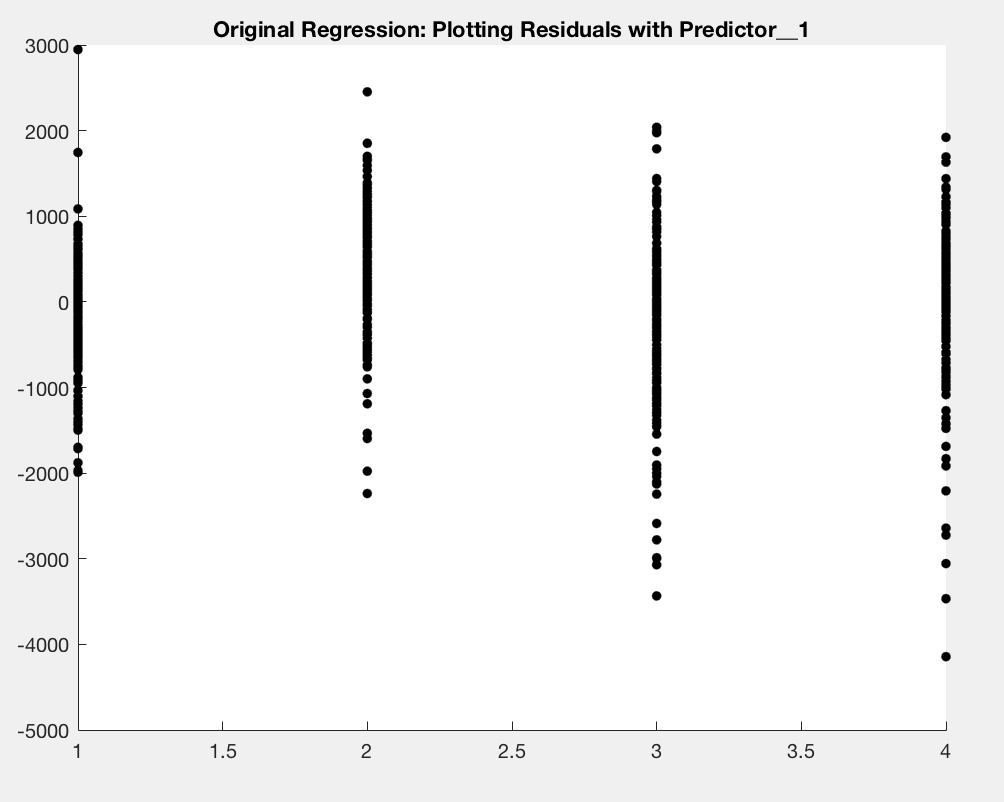
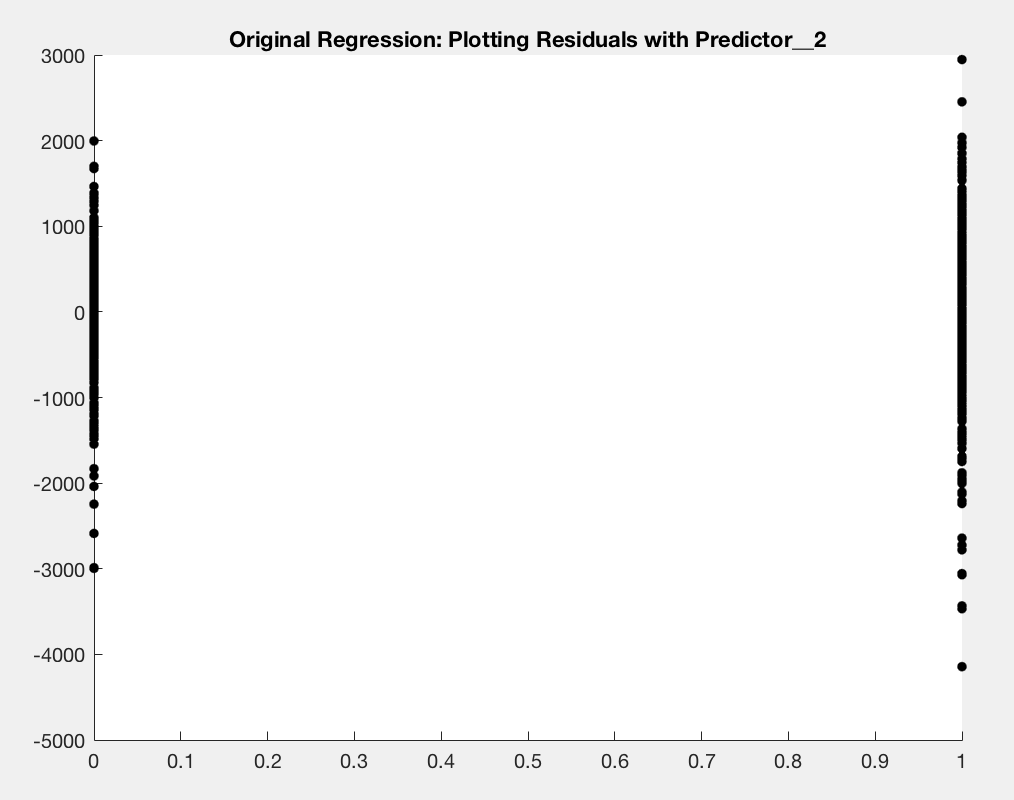
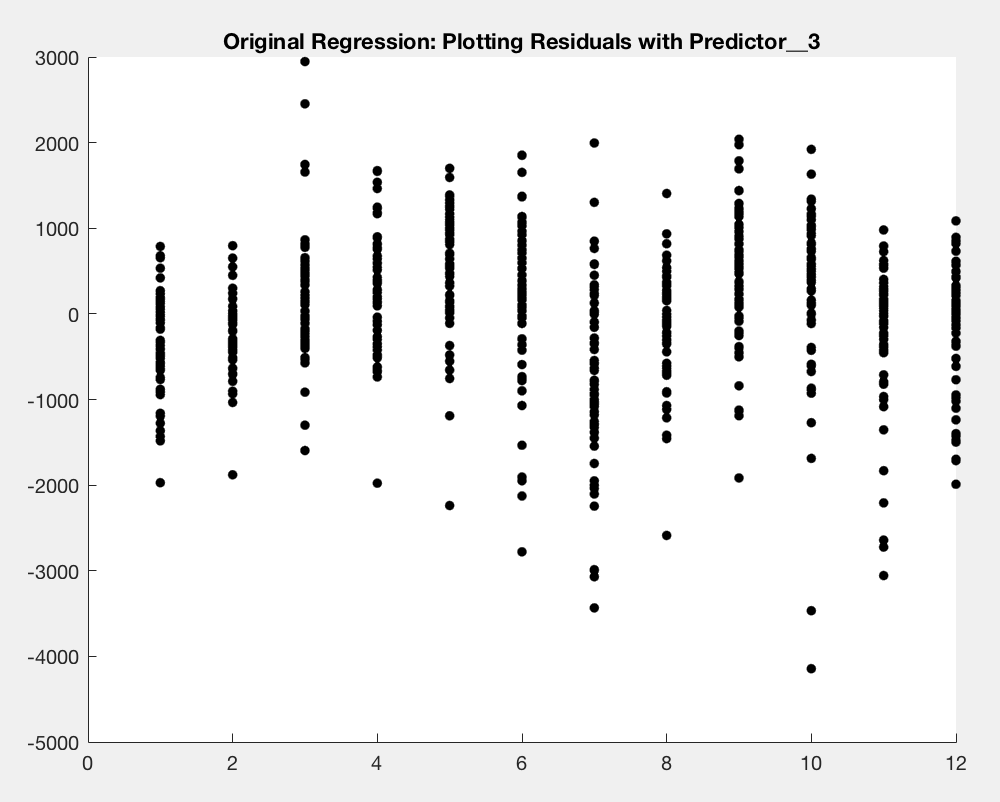
I did complete an analysis using ridge regression on the set of non-collinear features. However, when I varied the lambda significantly between different iterations of the algorithm, my betas remained constant. I took this as a strong indication that the algorithm was not minimizing any additional betas because they were all found to be sufficiently explanatory. It was for this reason that I did not do any additional analysis on this particular regression, there was nothing that would be found here that would differ from the findings in the *Removing Collinear Features* regression.

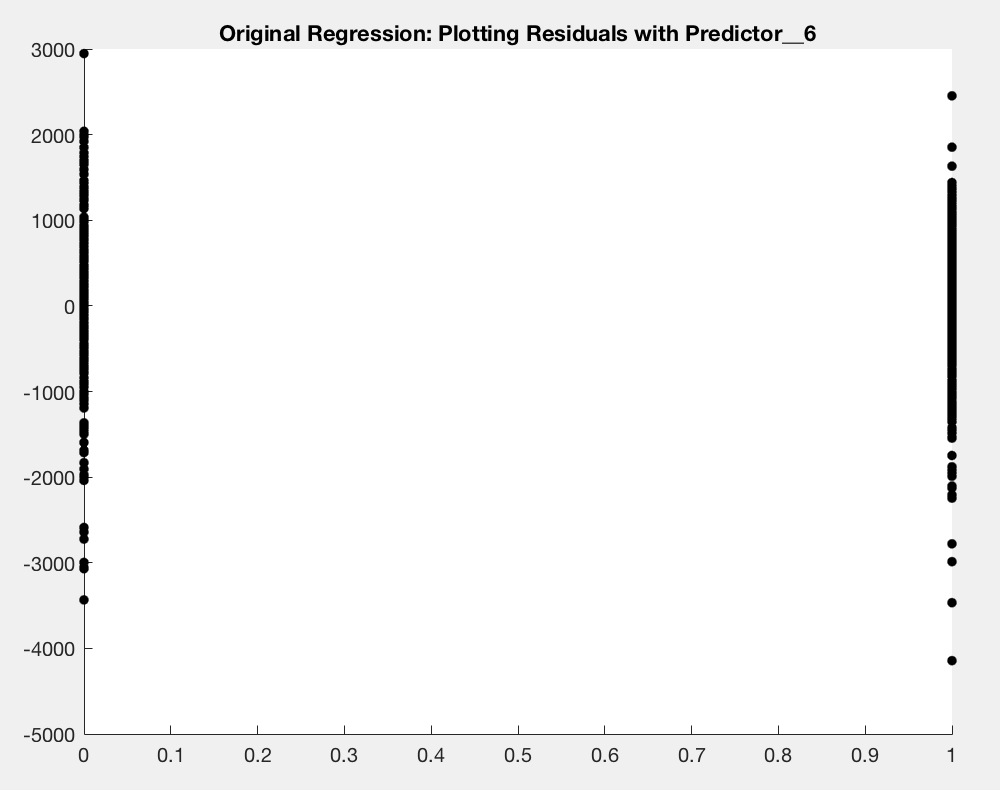
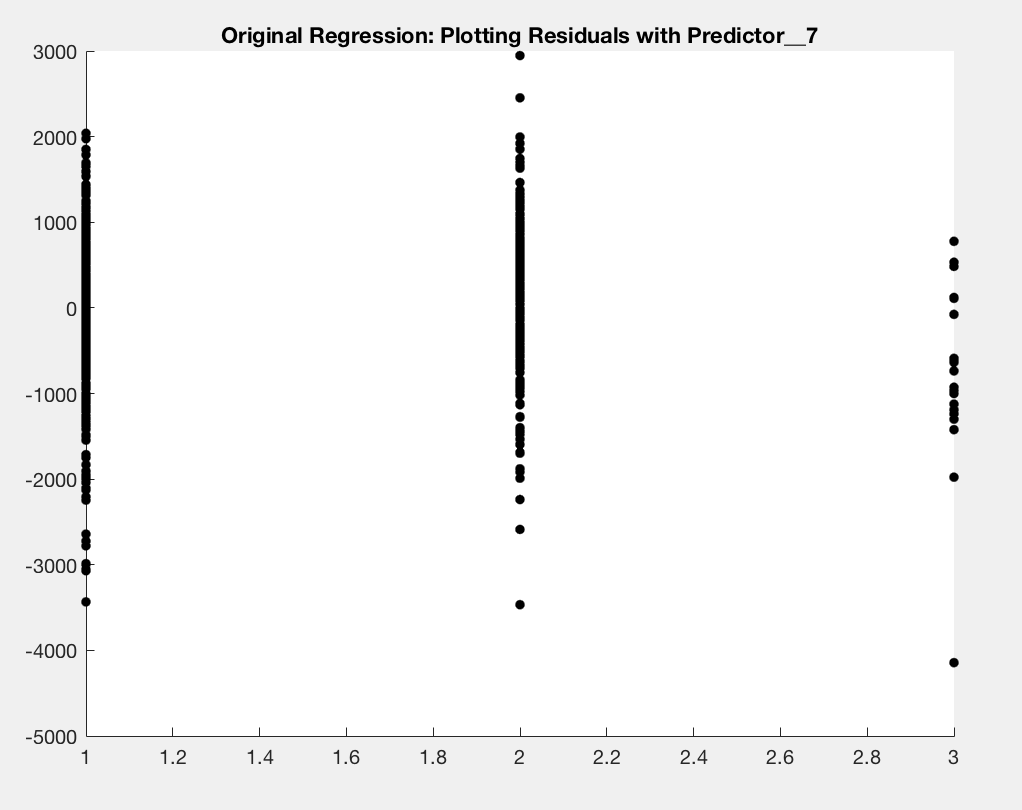
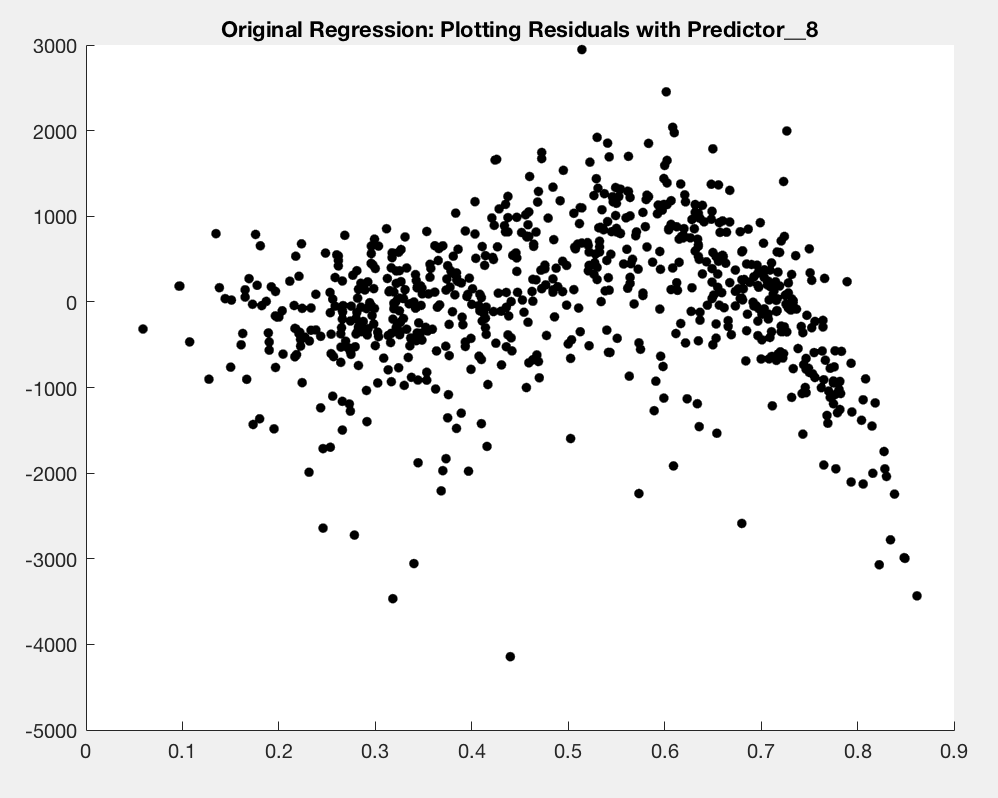
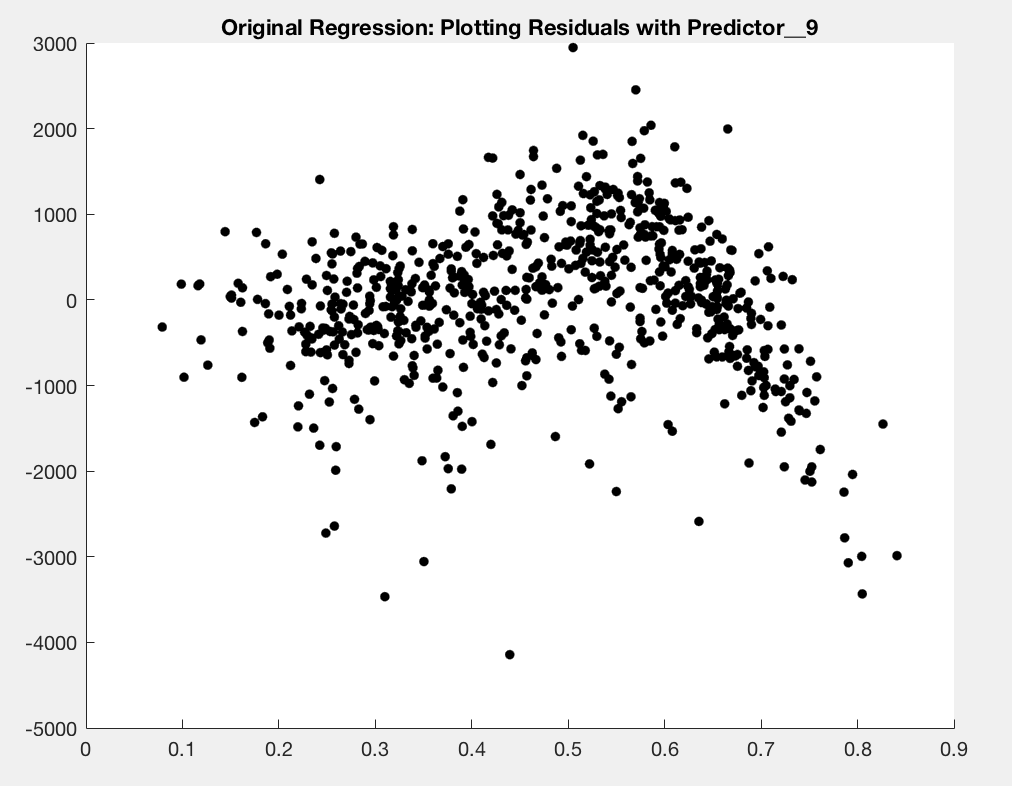
**Results**

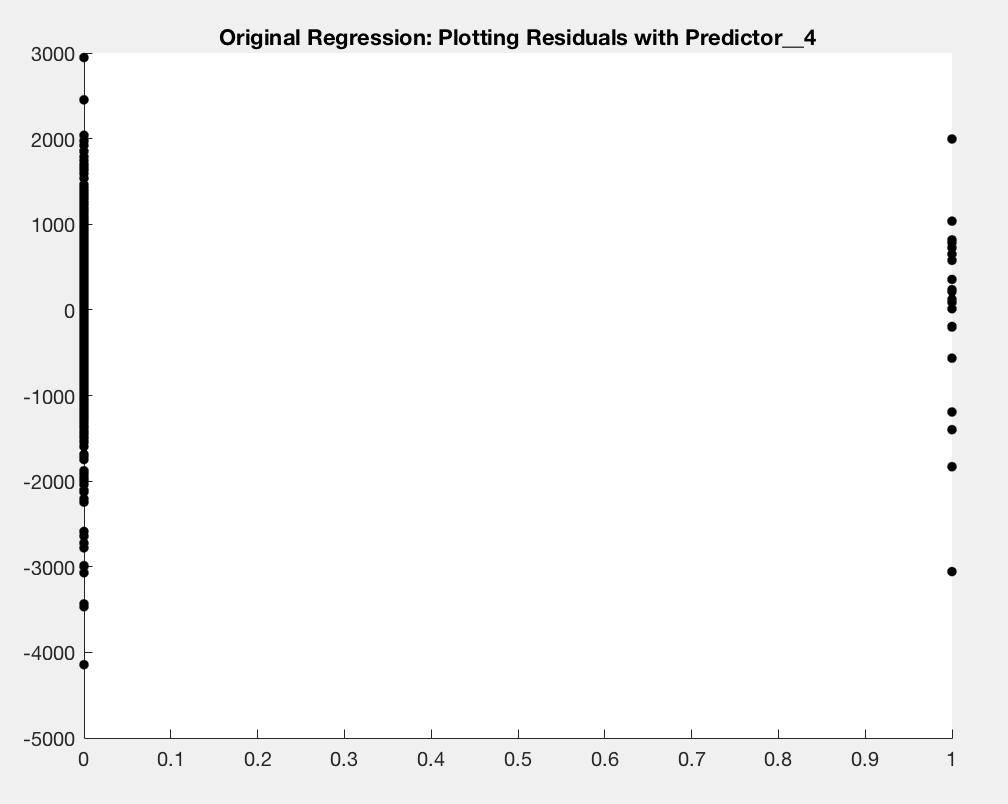
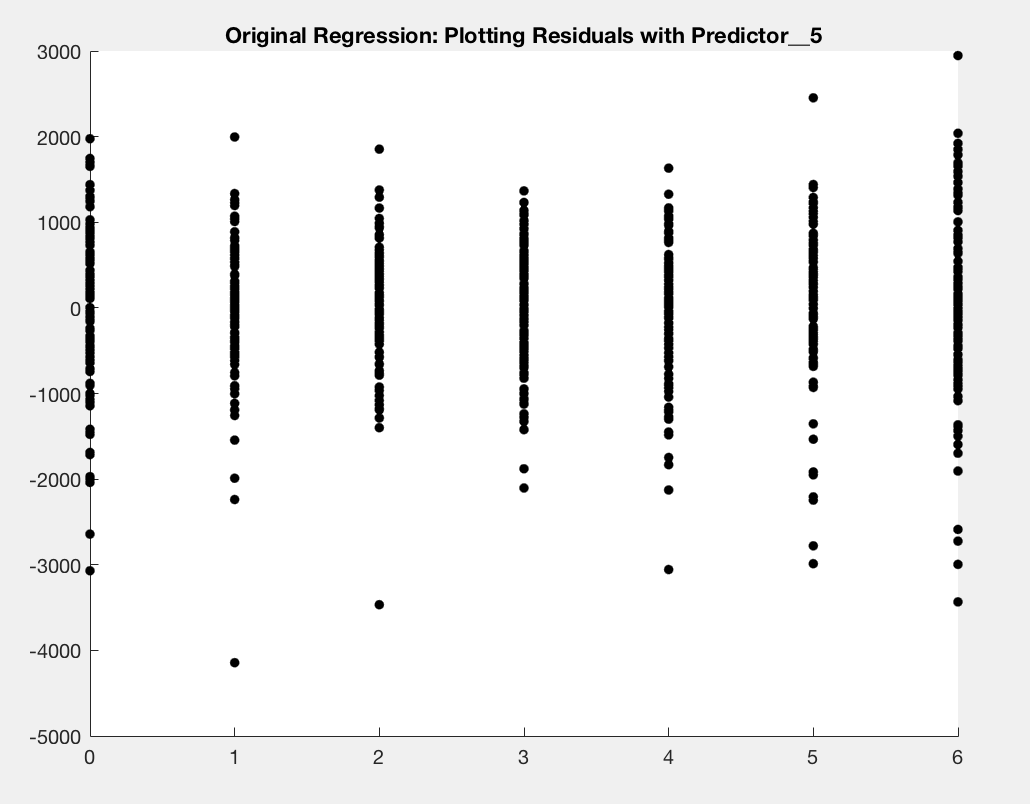
There is a strong indication that bike sharing in Los Angeles would be a great success! Across the board, due to warm weather, clear days, and an absence of foul weather, Los Angeles provides an excellent climate for renters to use a bike sharing program. In order to make a clearer estimate of whether or not Los Angeles should be chosen as a *permanent* location for a bike sharing program requires more research analysis. A feasible next step would be to try a pilot program, phasing it in during the early spring so that ideally, it grows in popularity over the spring, summer, and fall, which are the nicest seasons and consequently the seasons where most bike rentals would occur. The pilot should begin with around 3000 bikes. This will be substantial to cover the high demand days, with the possibility of all bikes getting checked out. 3000 is also a relatively small number compared to our estimates, but I think it is appropriate to mitigate installation costs of a pilot project. After data has been collected on the pilot project, the city can consider a more widespread integration of bikes.

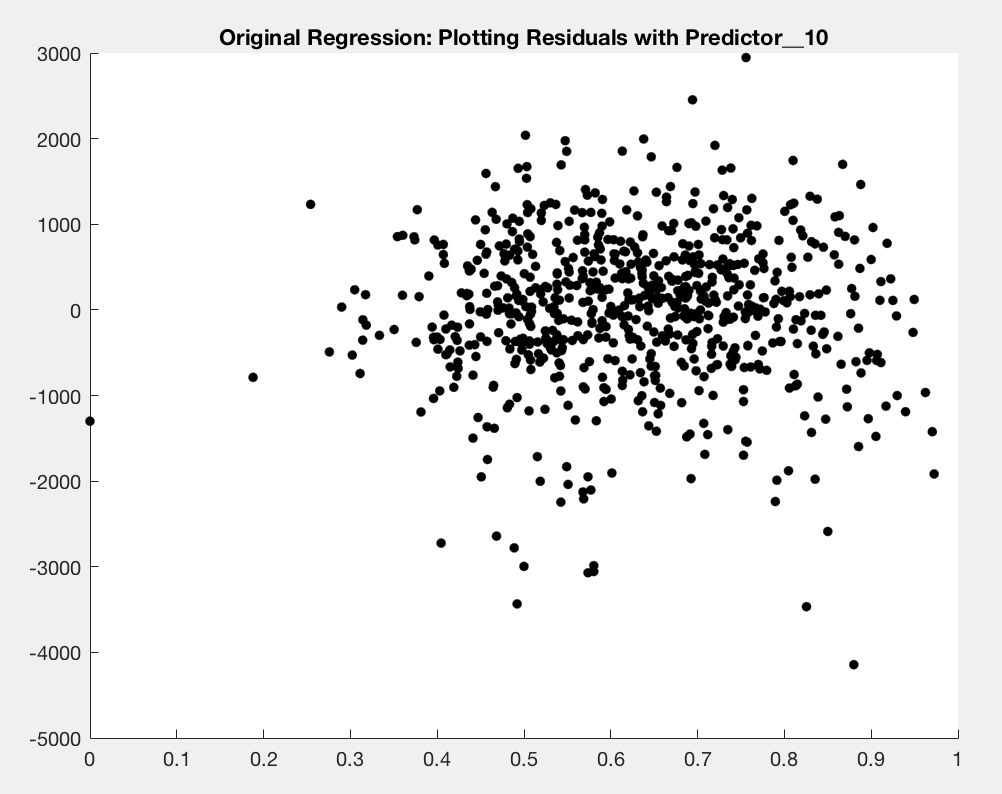
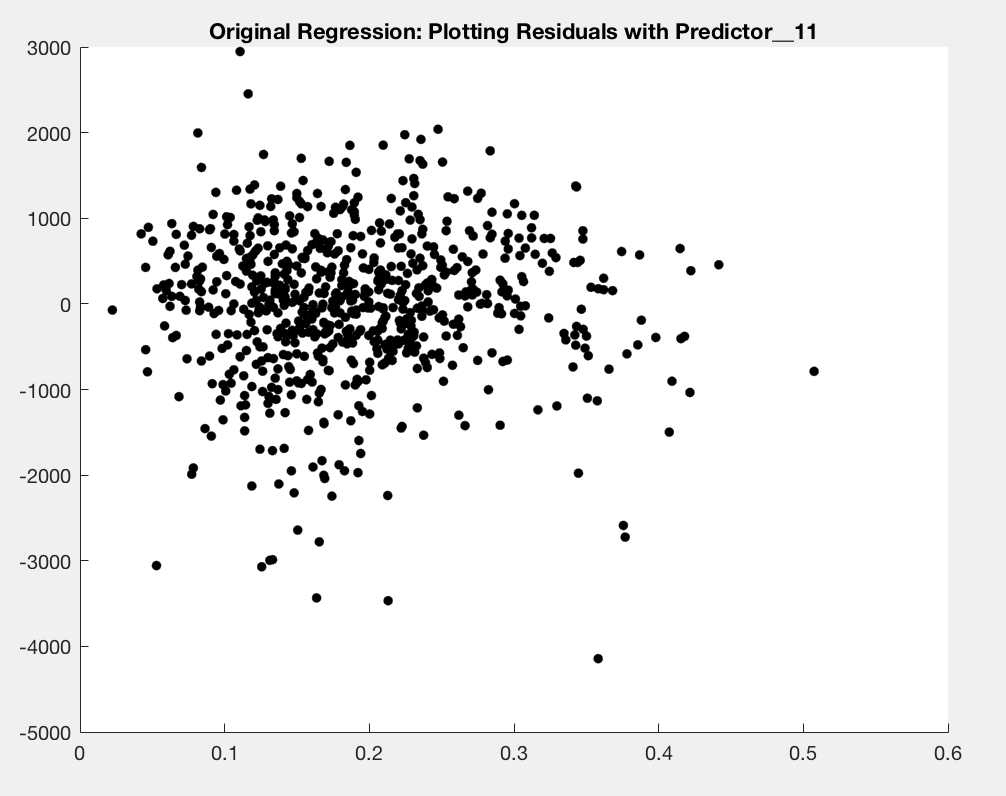
**Appendix**

**FIGURES 1-11**

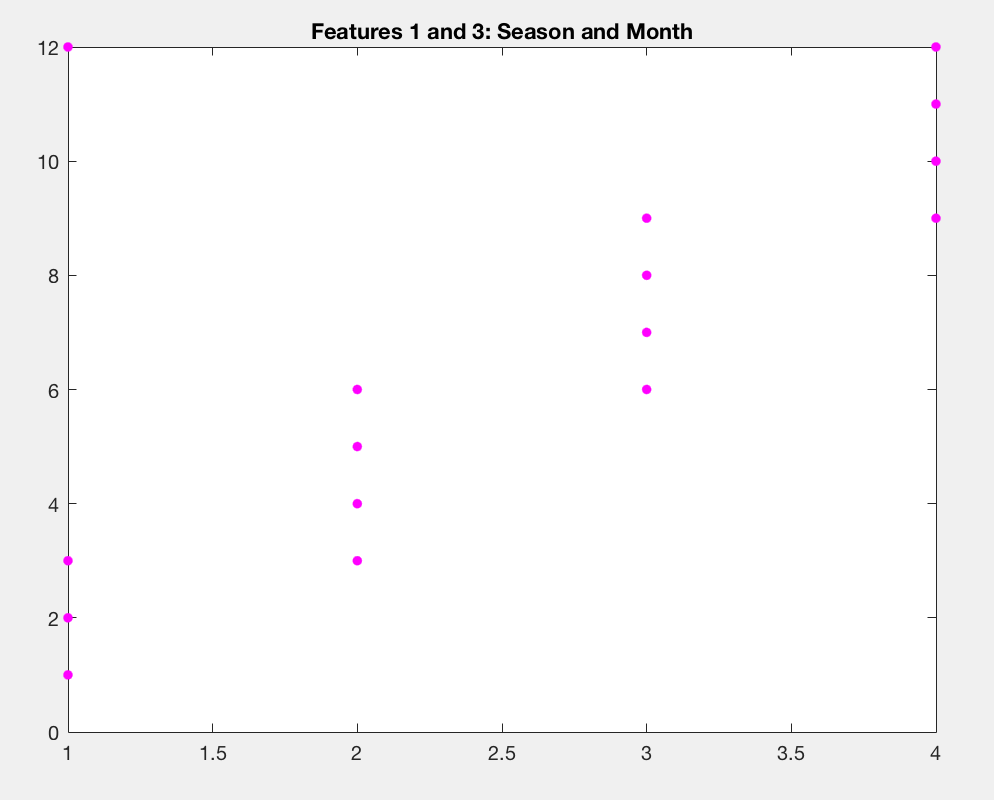


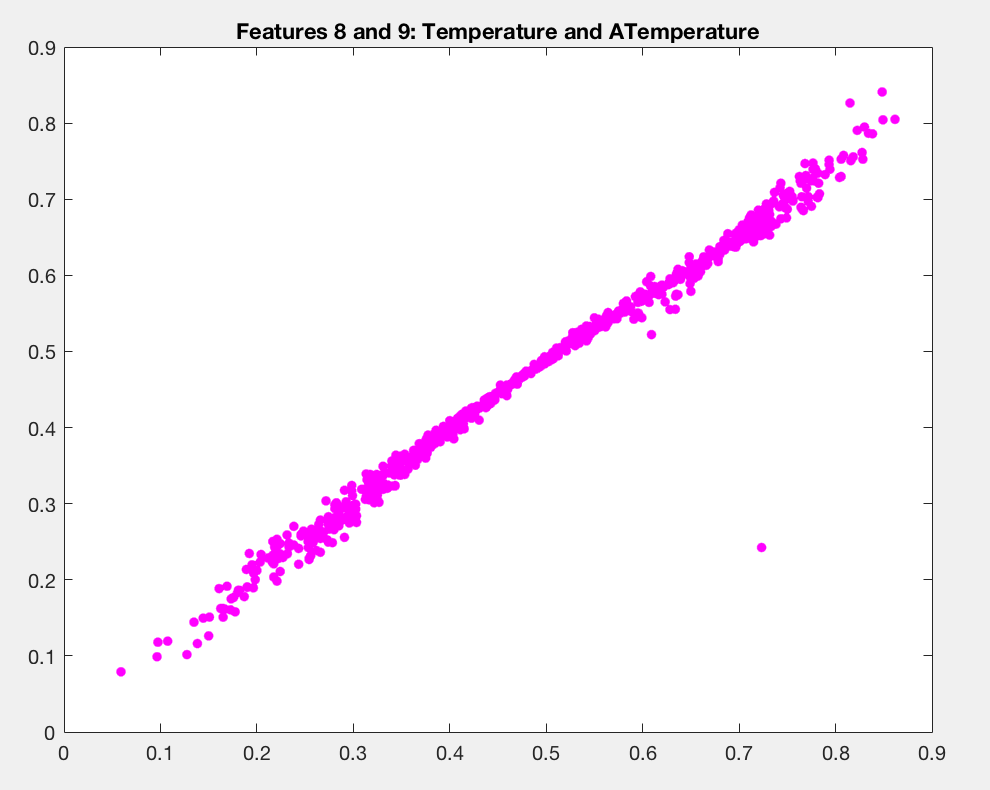
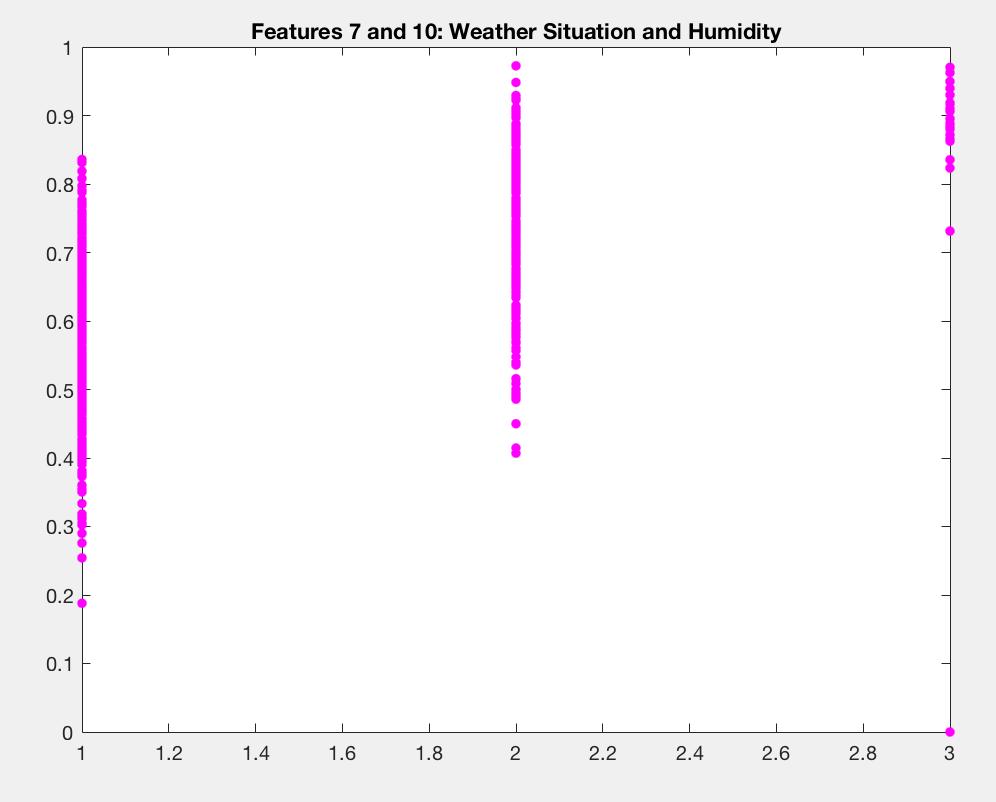




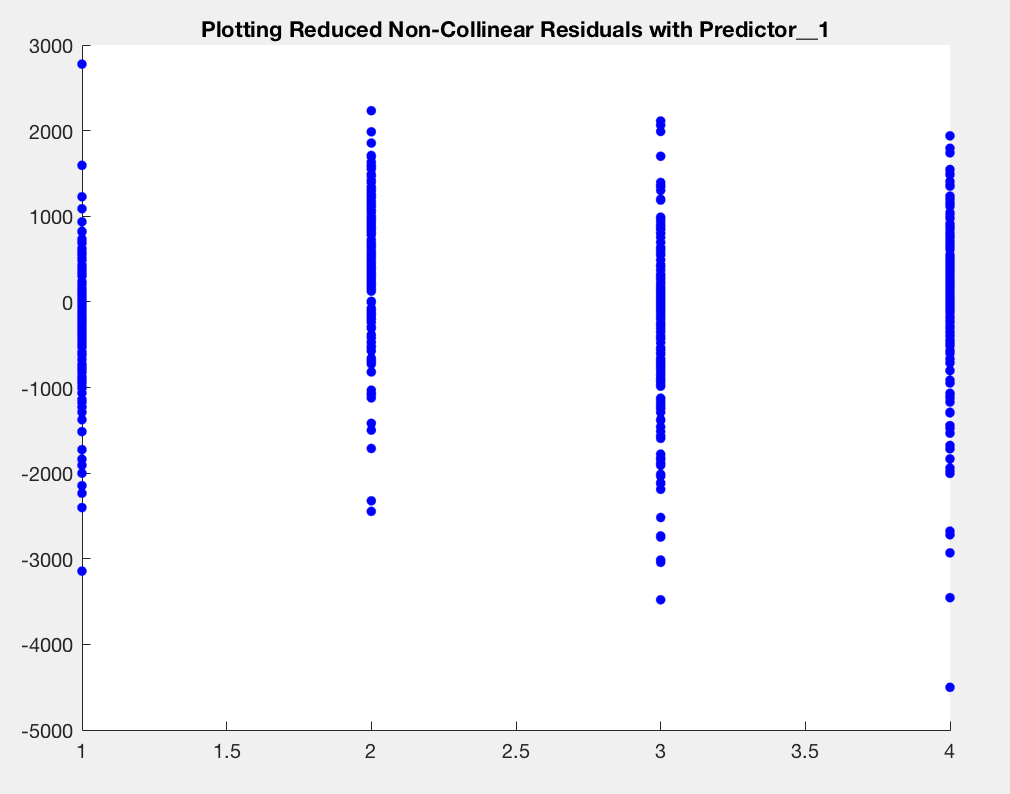
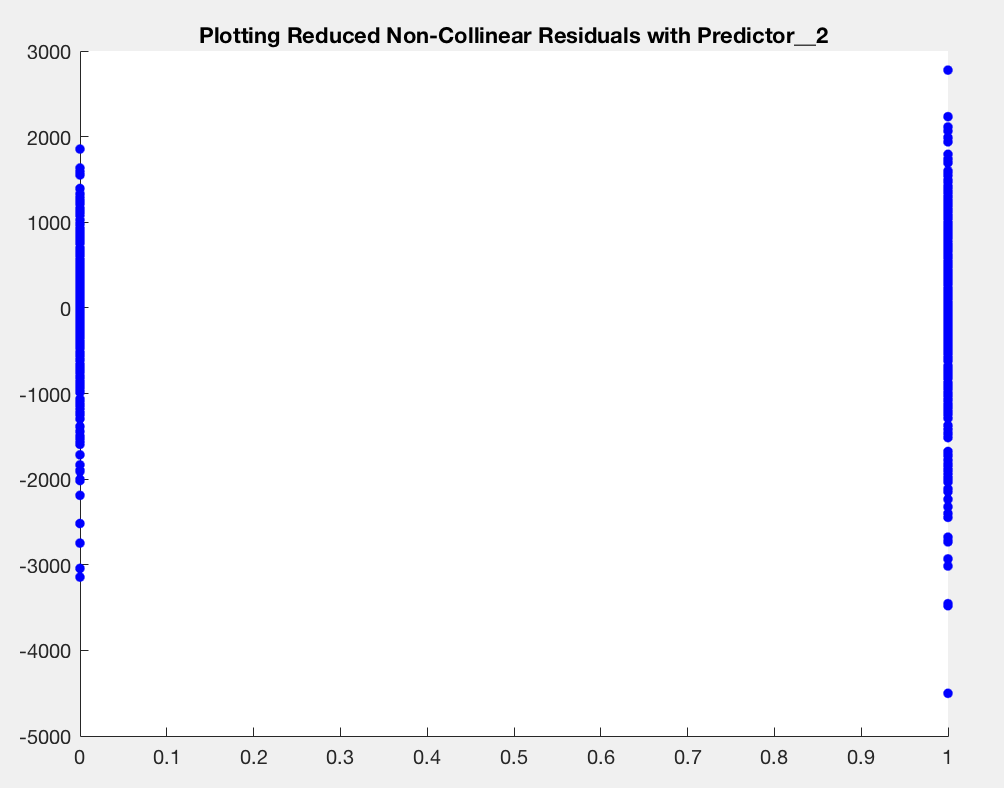
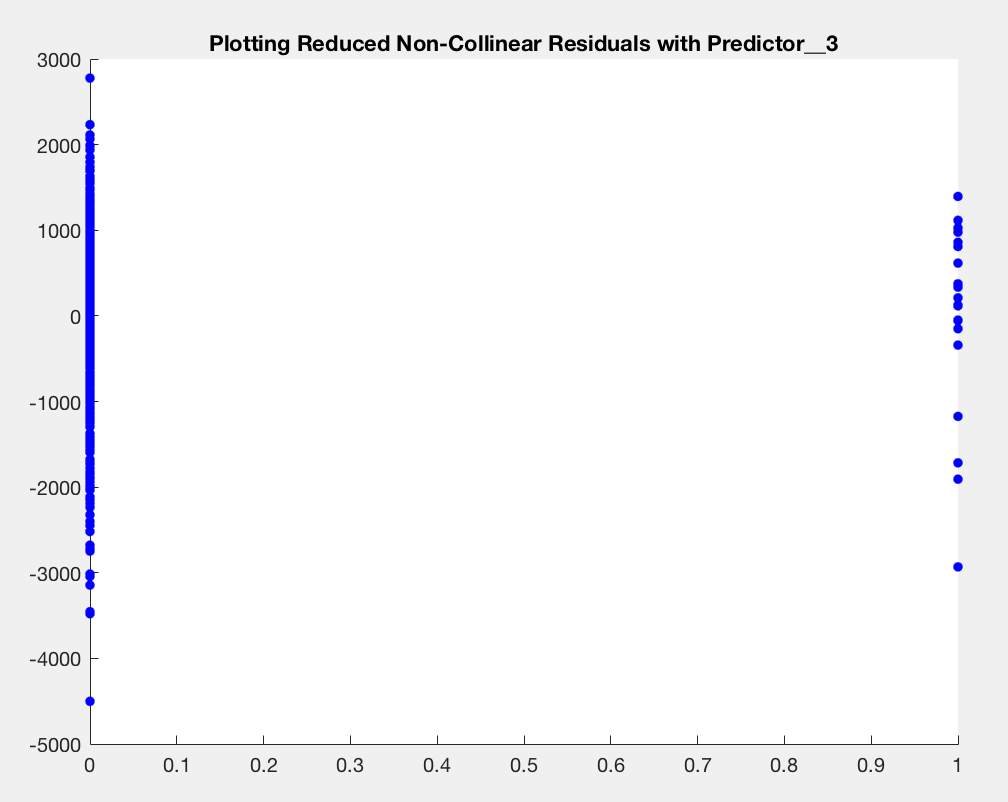


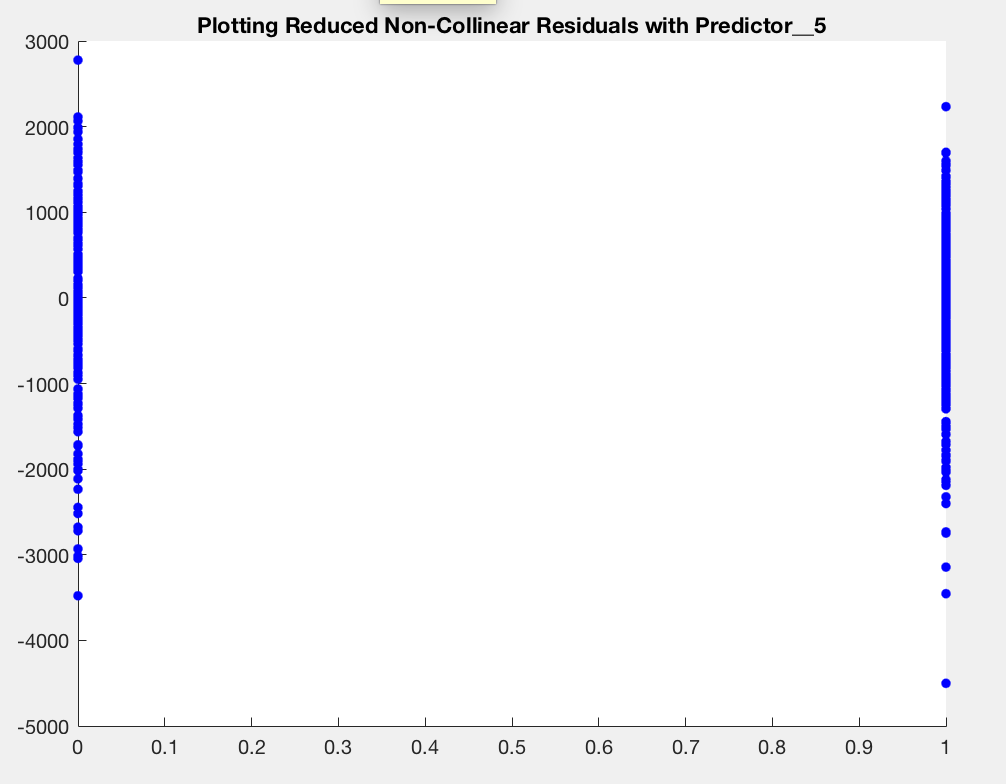
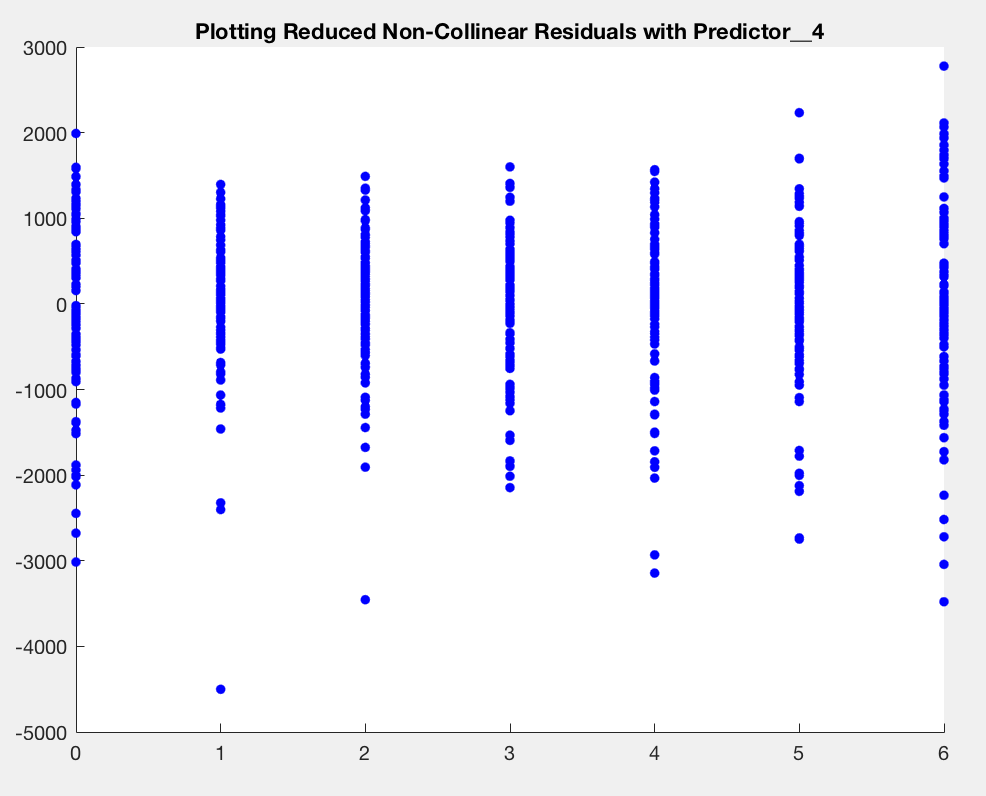
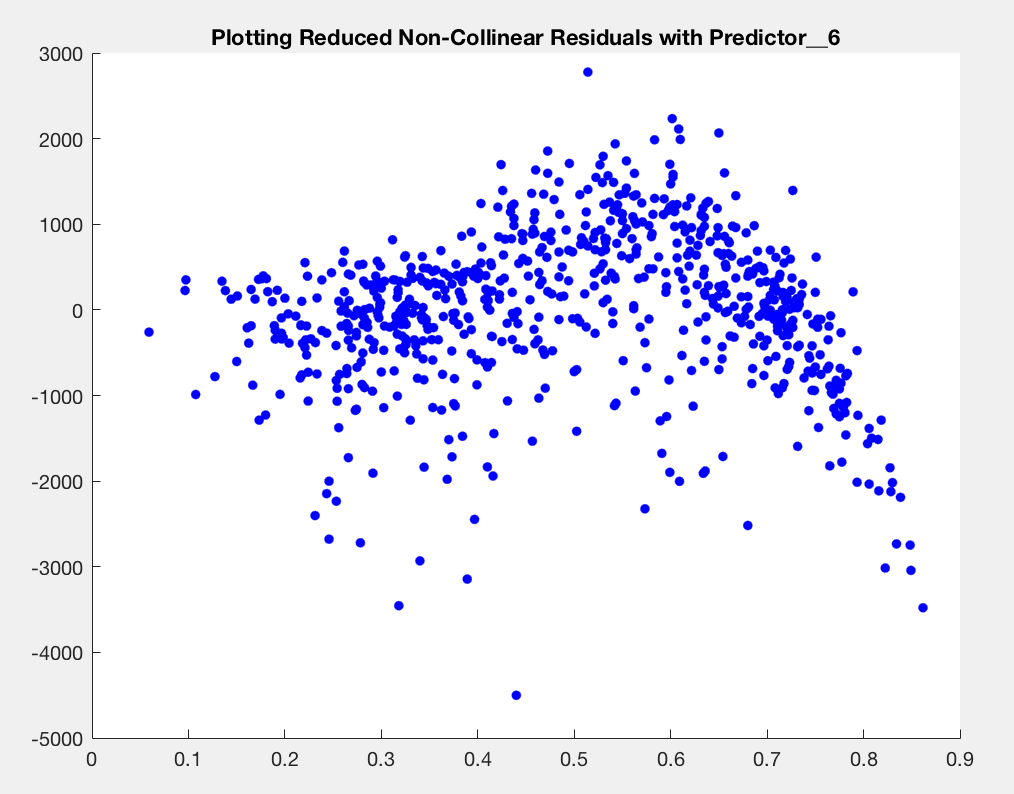
**FIGURES 12-14**

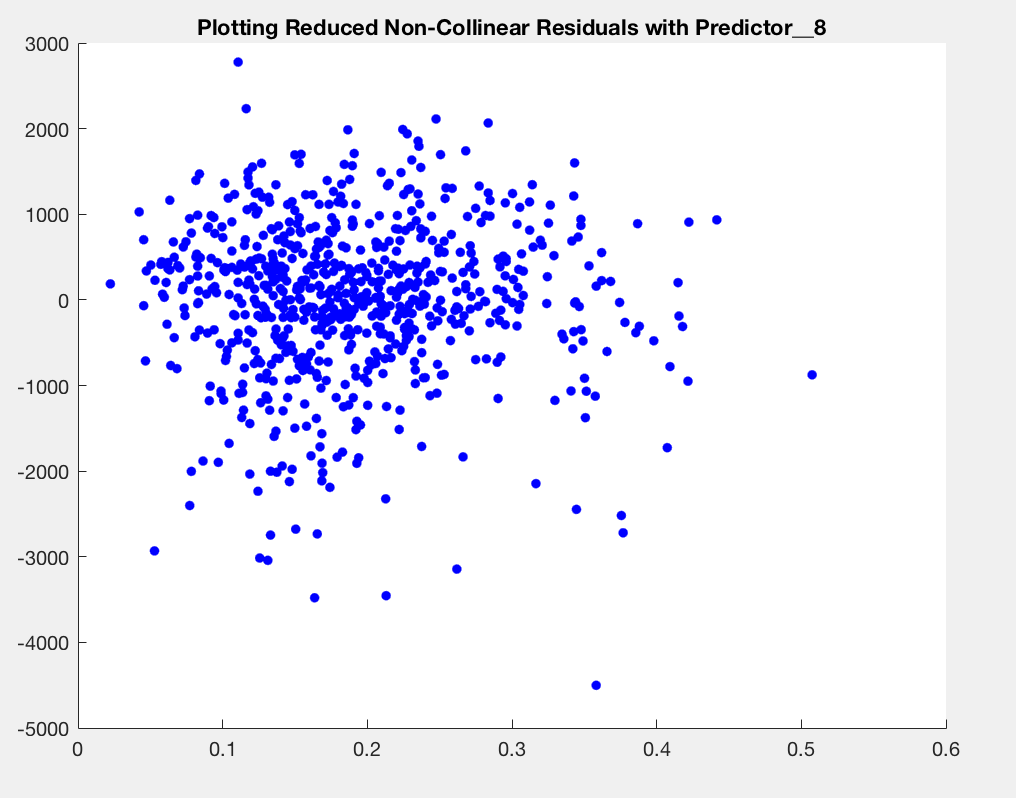
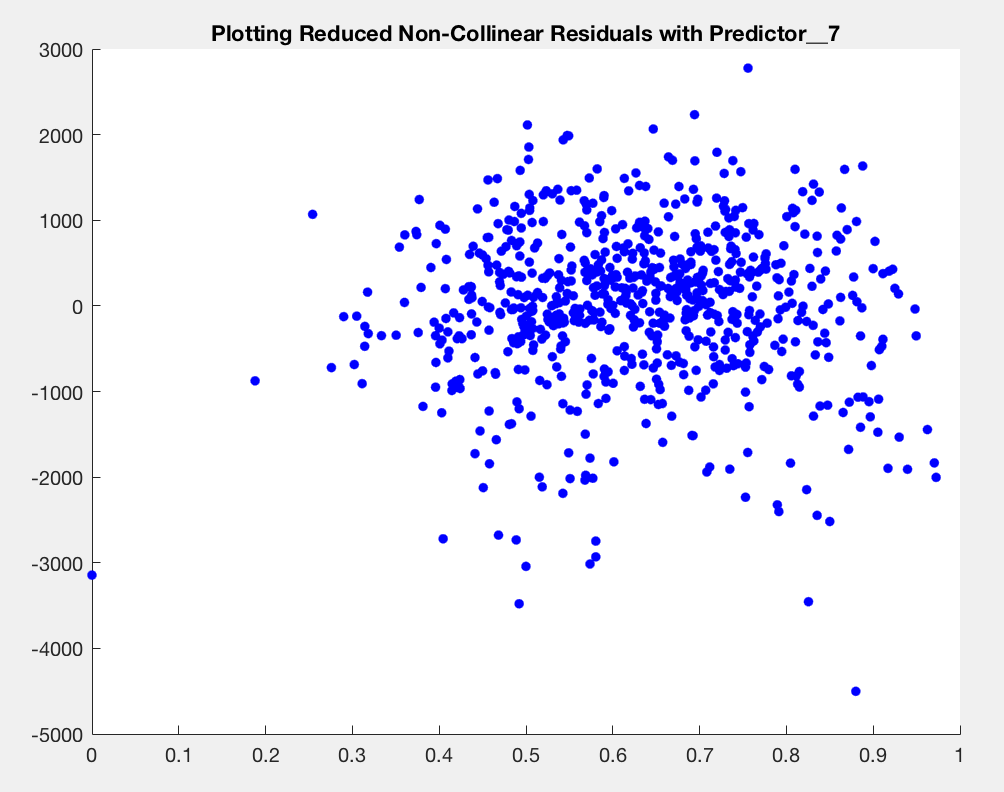




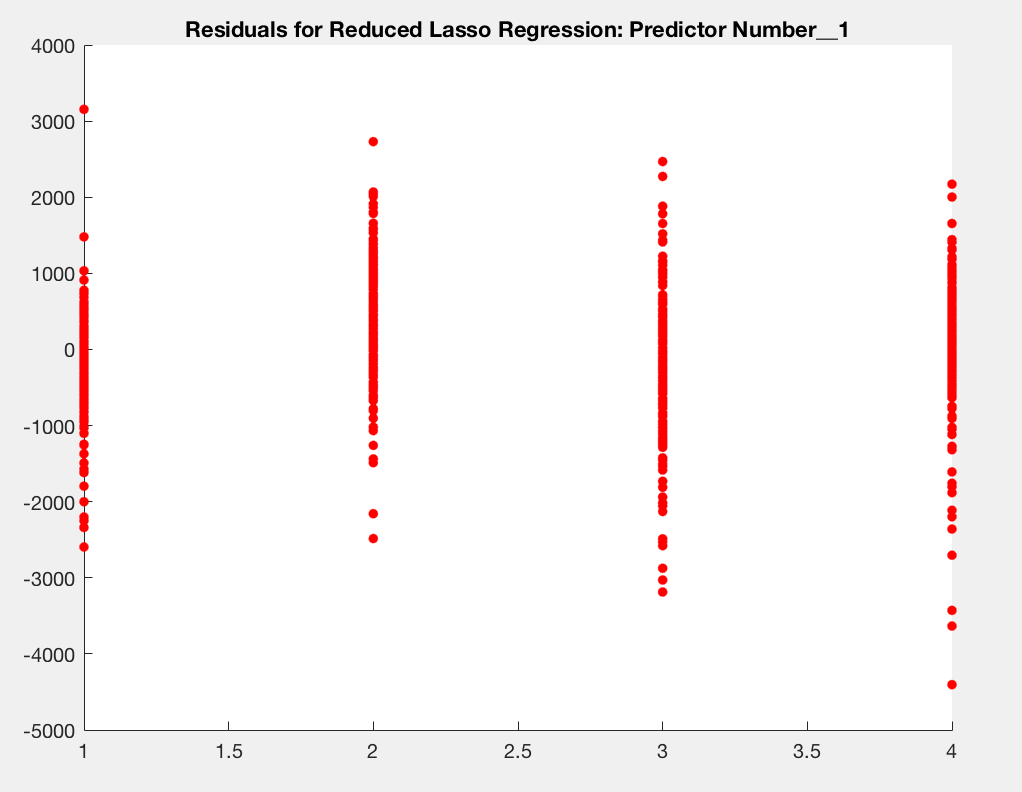
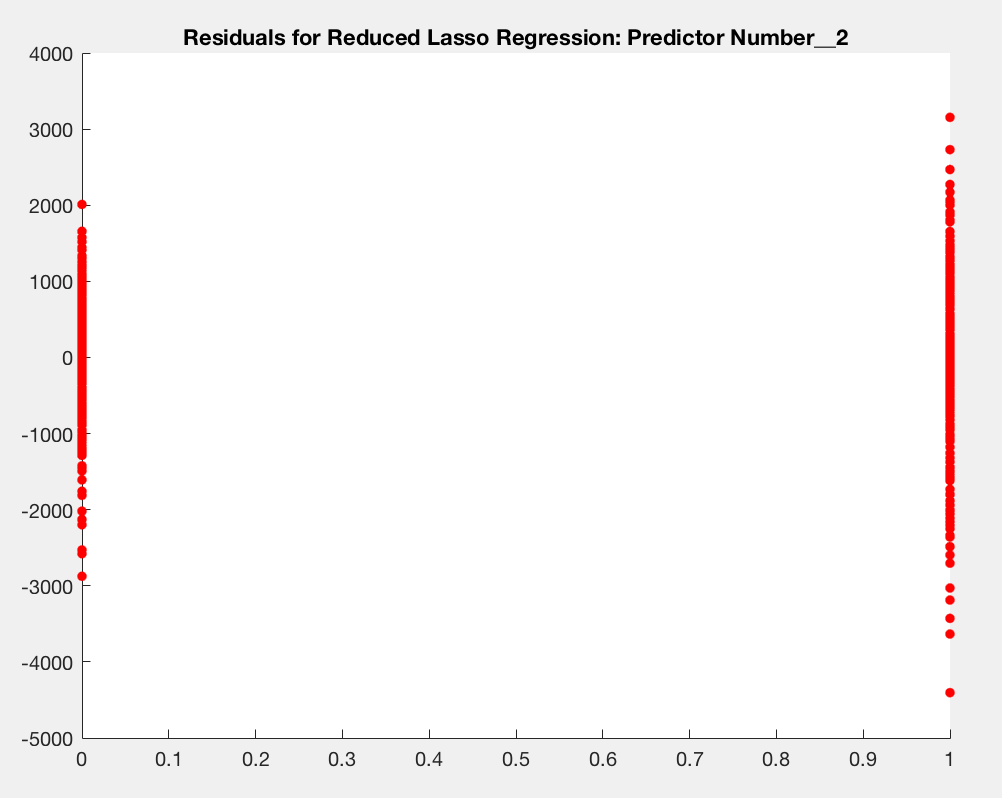
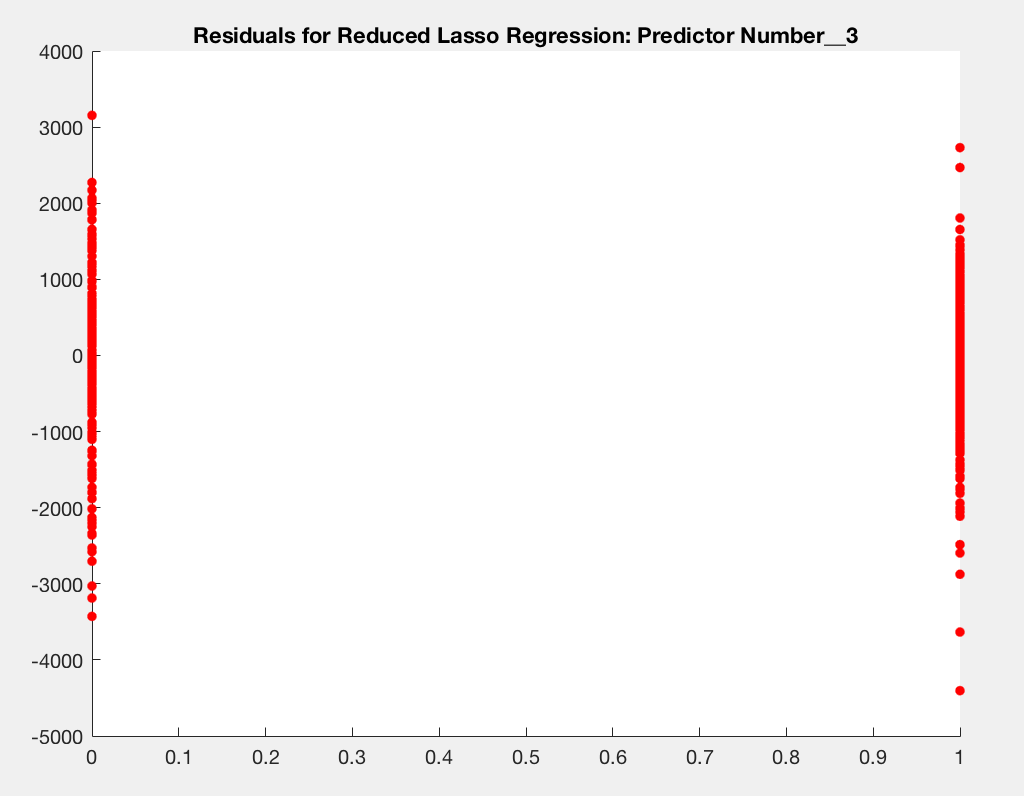
**FIGURES 15-22**

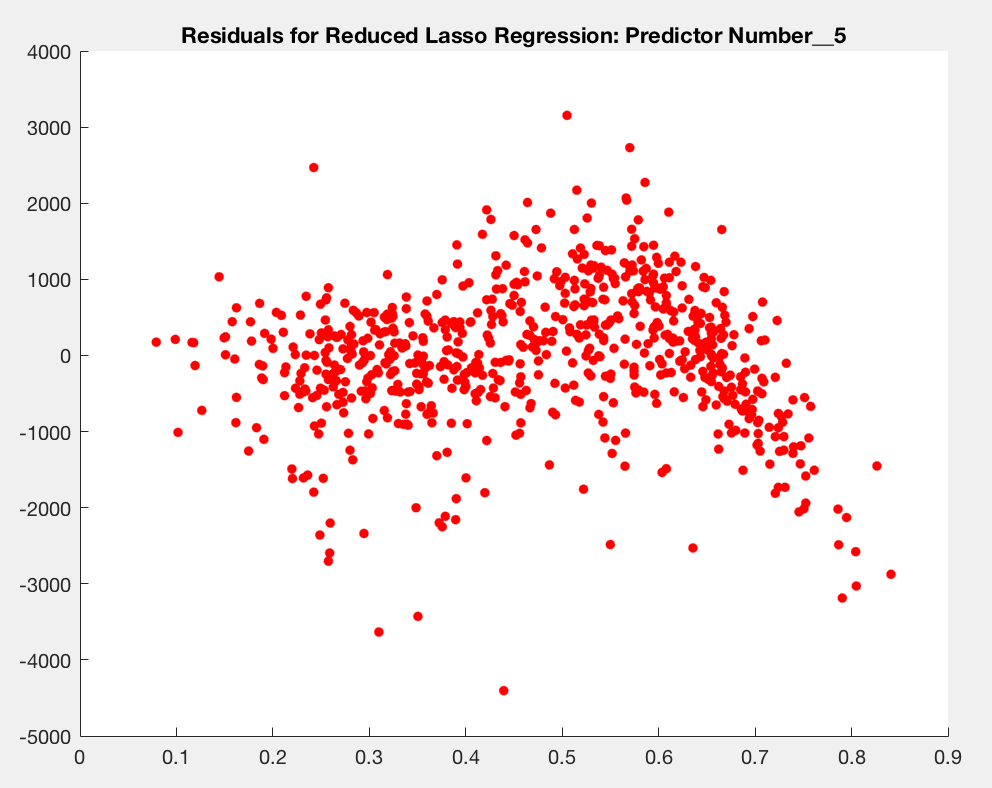
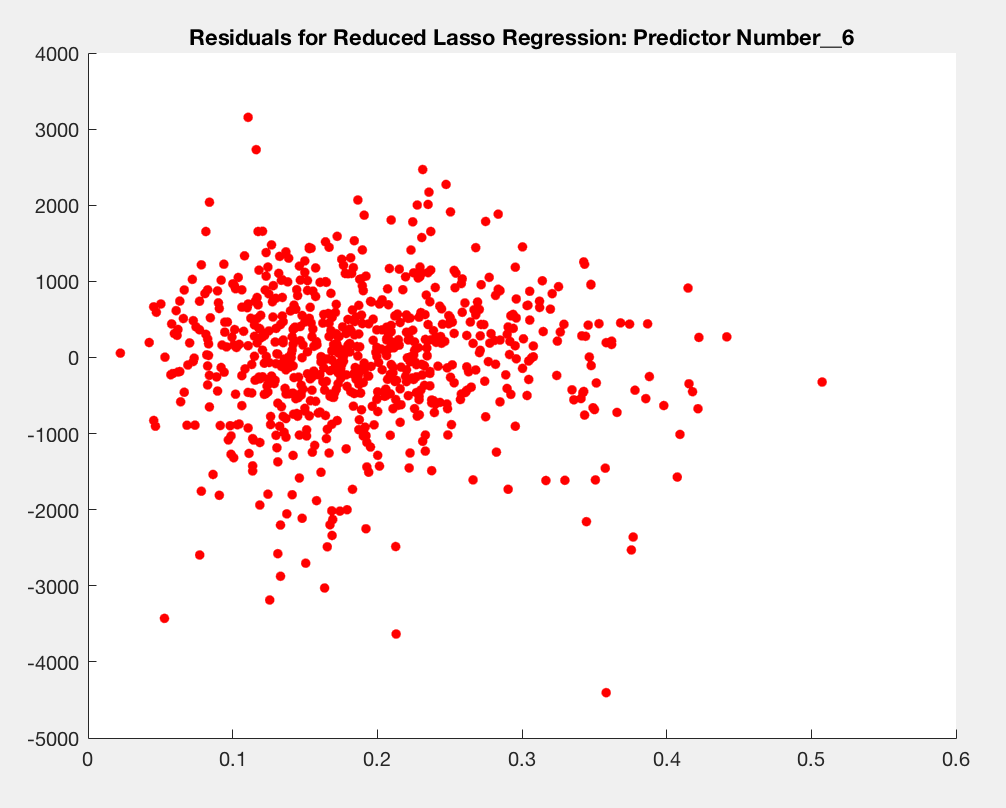
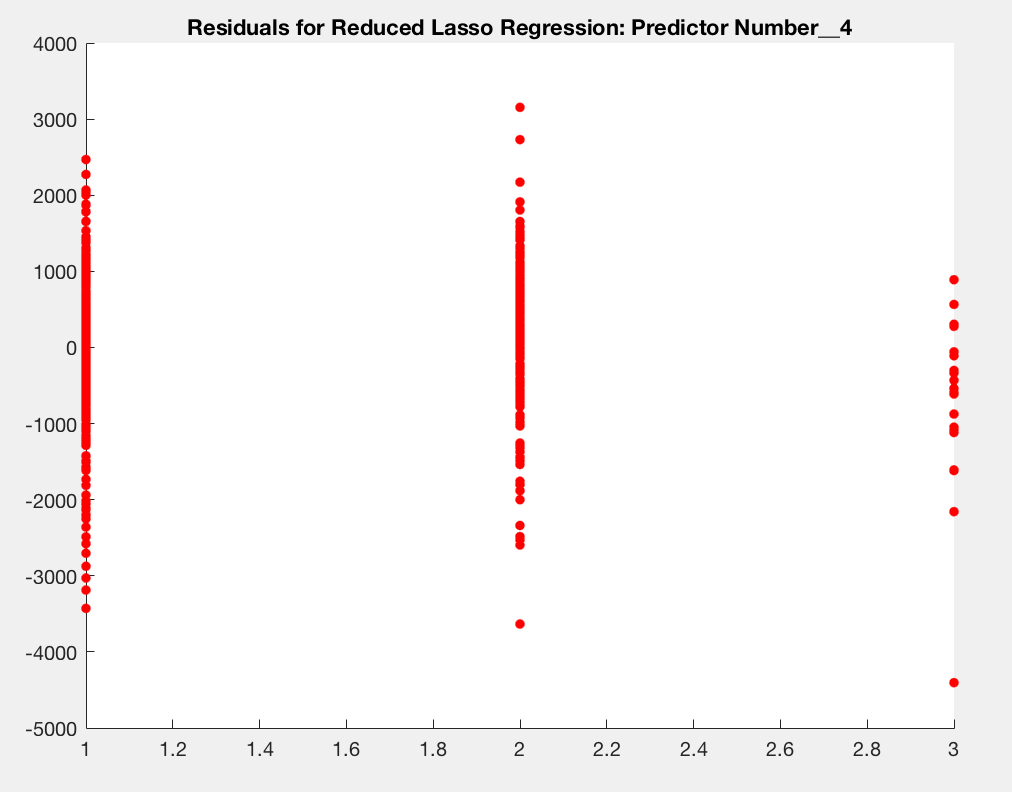




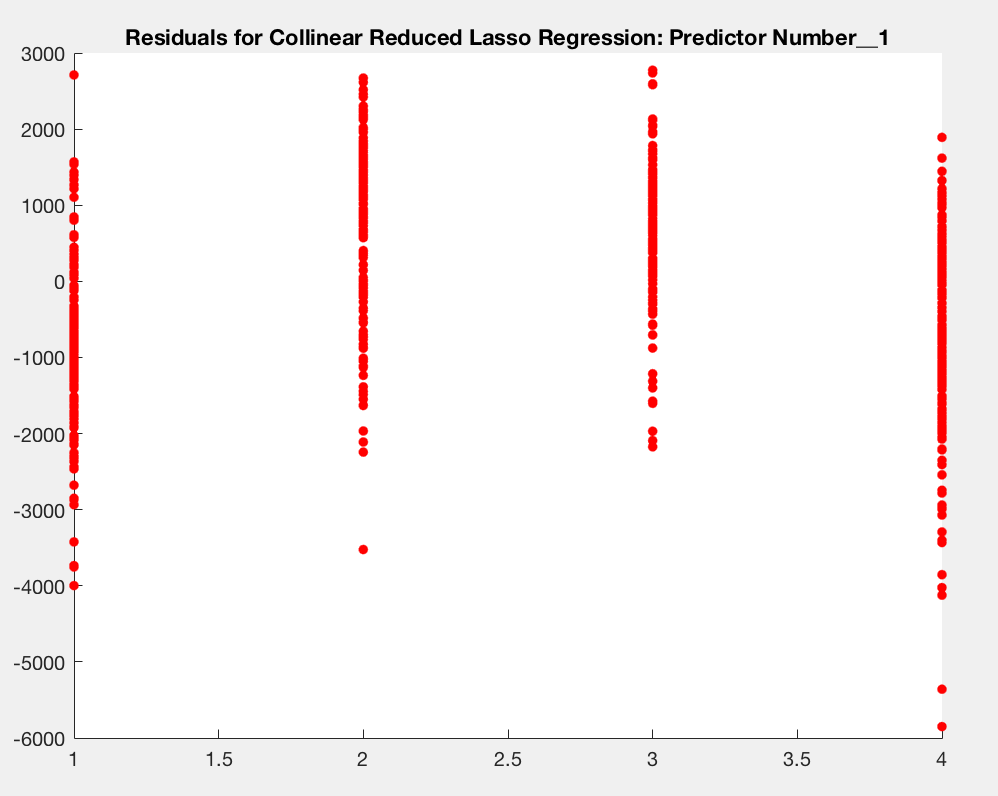
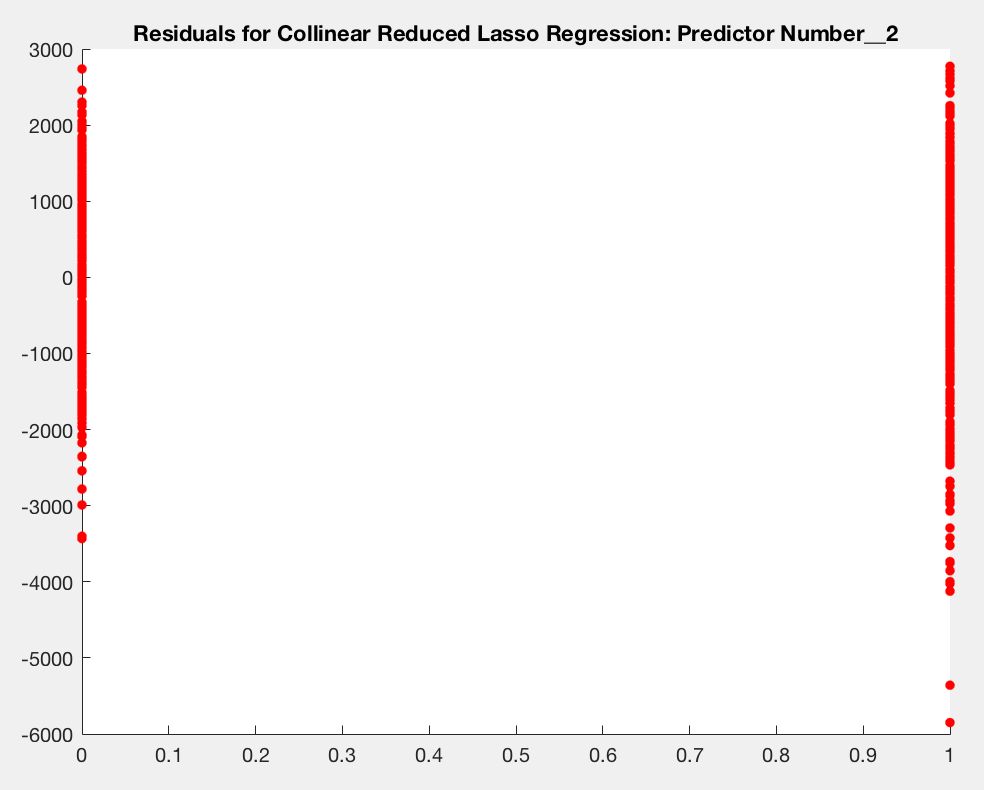
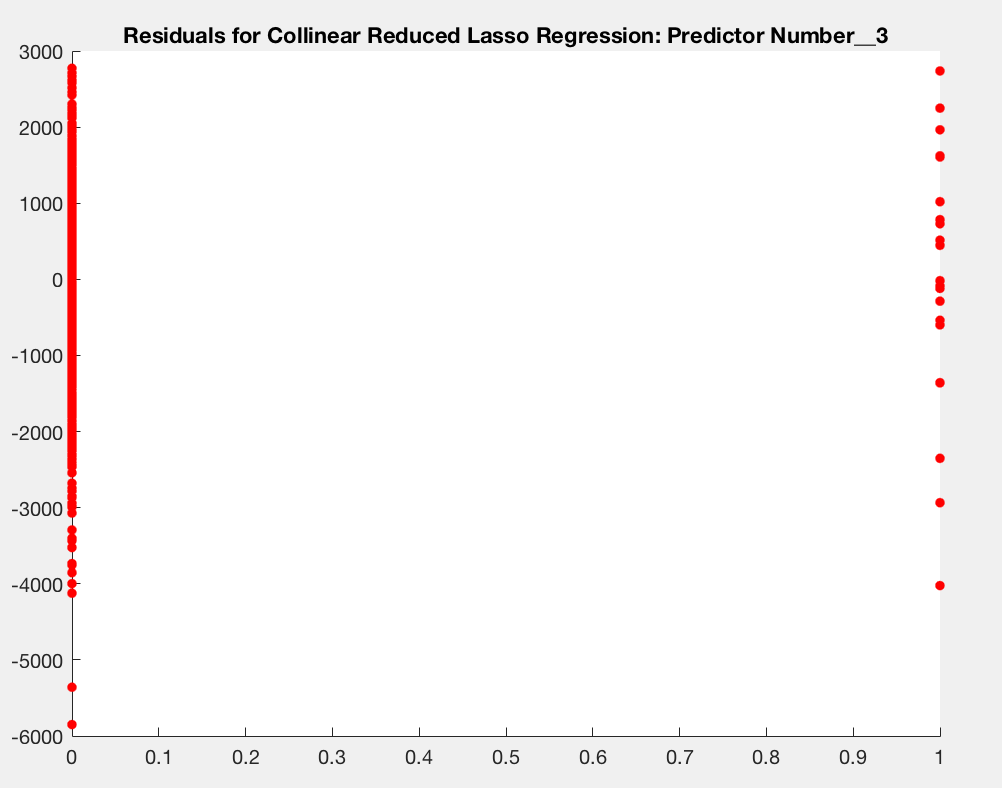


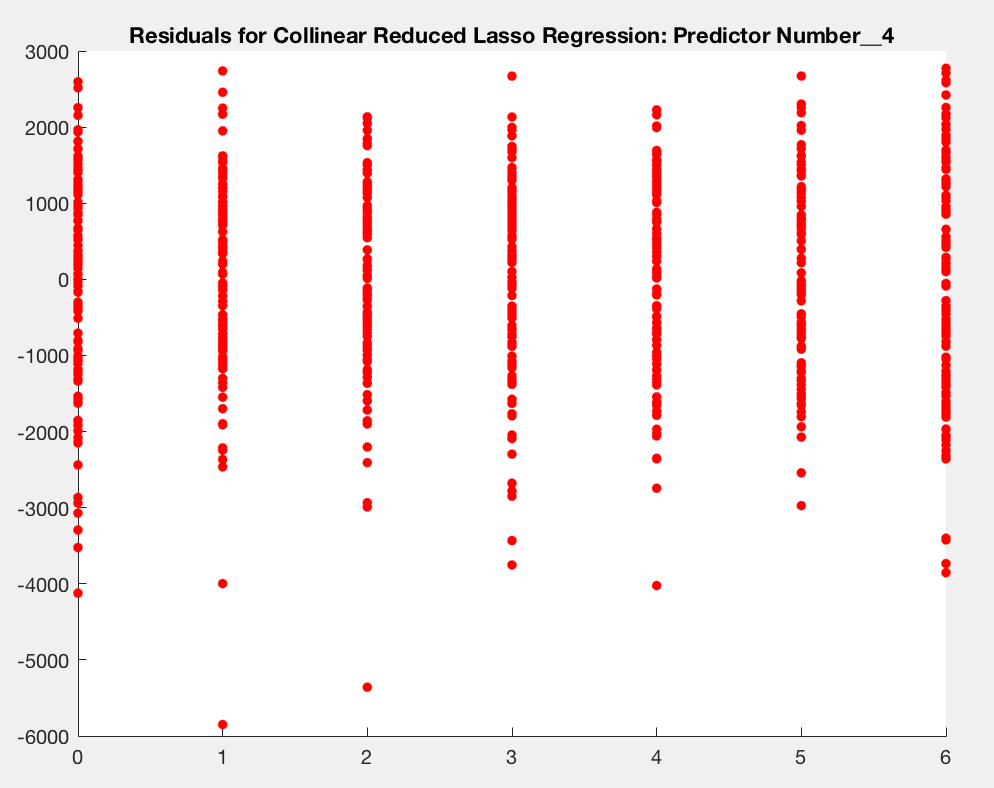
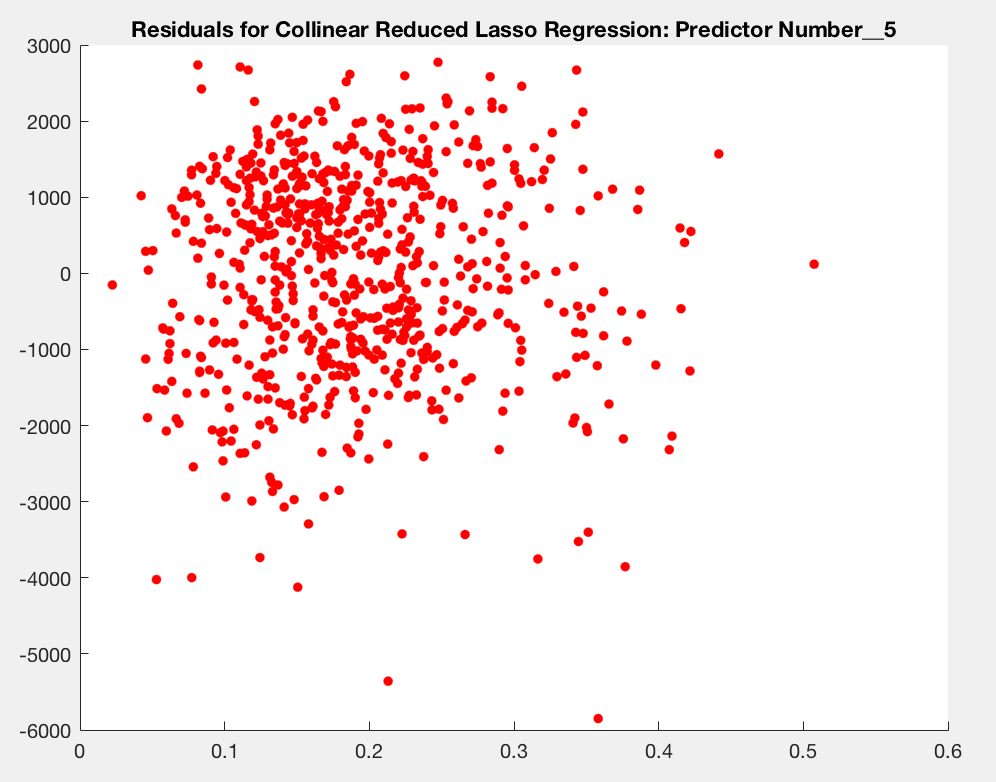
**FIGURES 23-28**



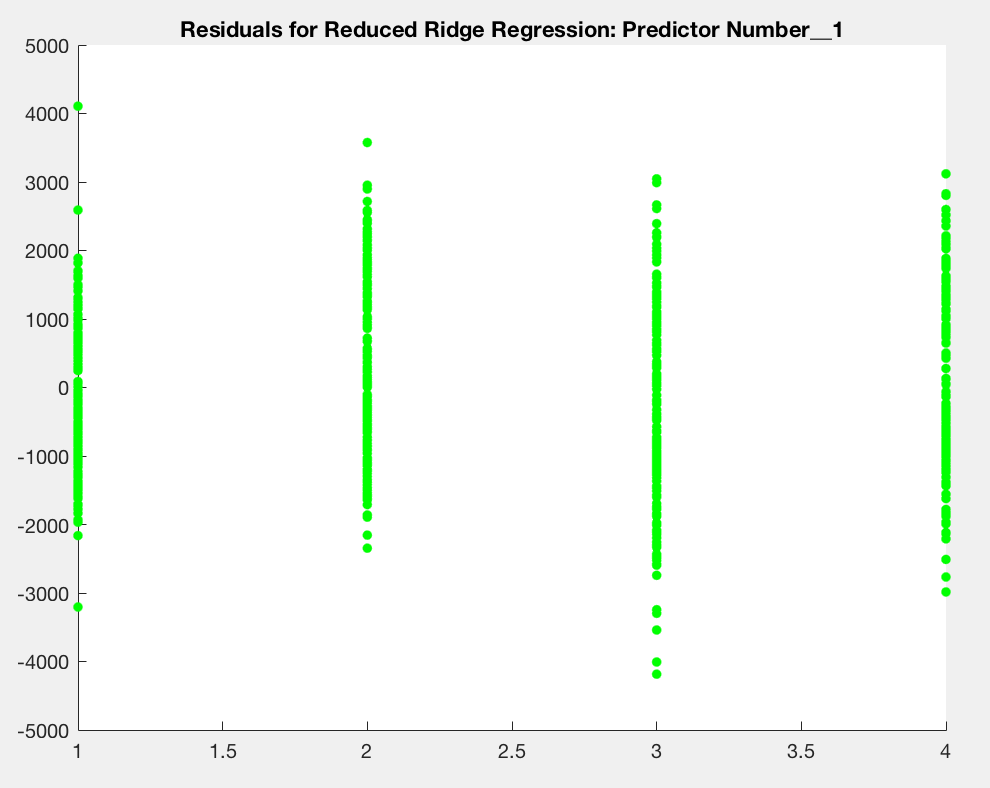
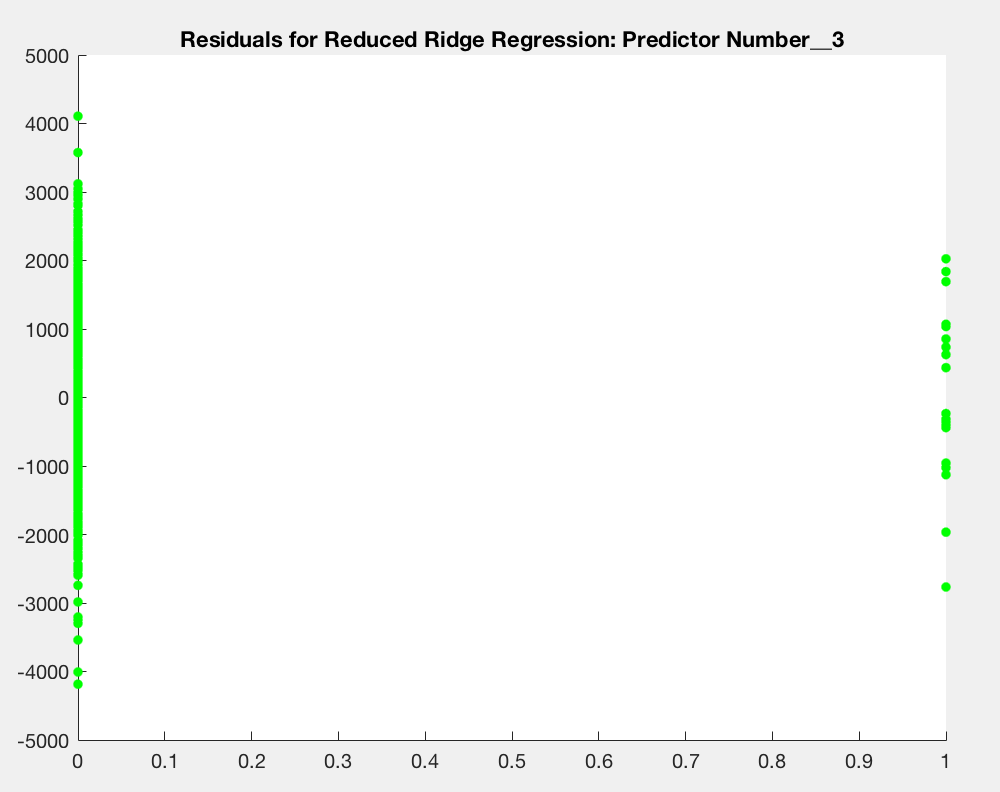
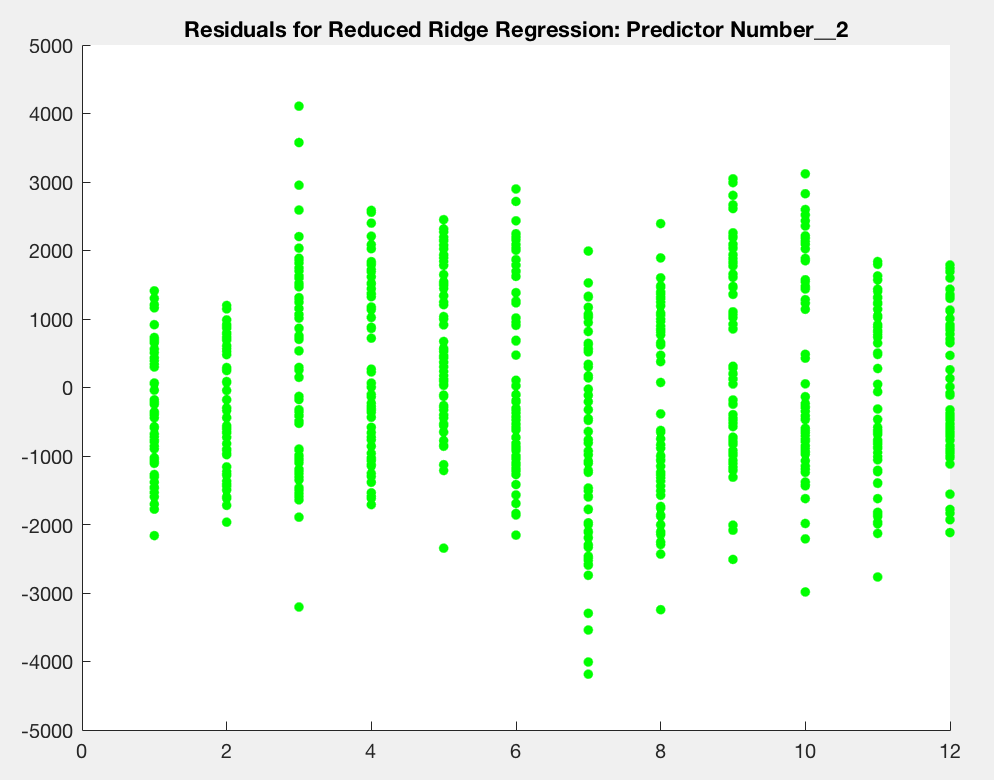


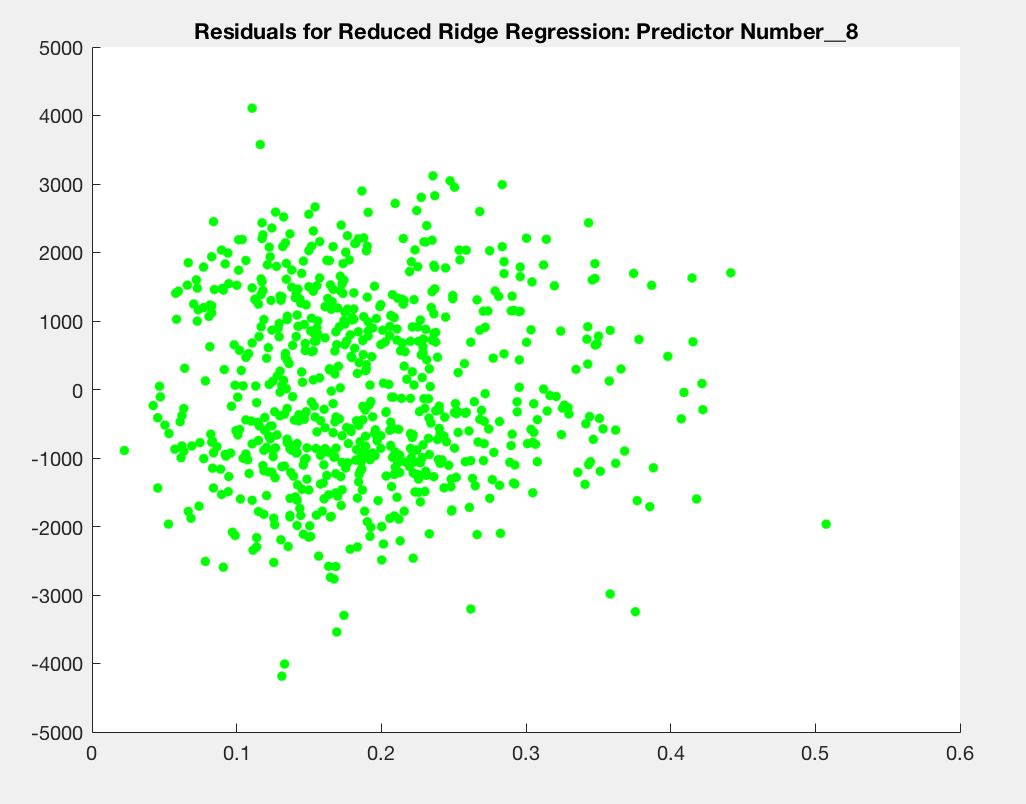
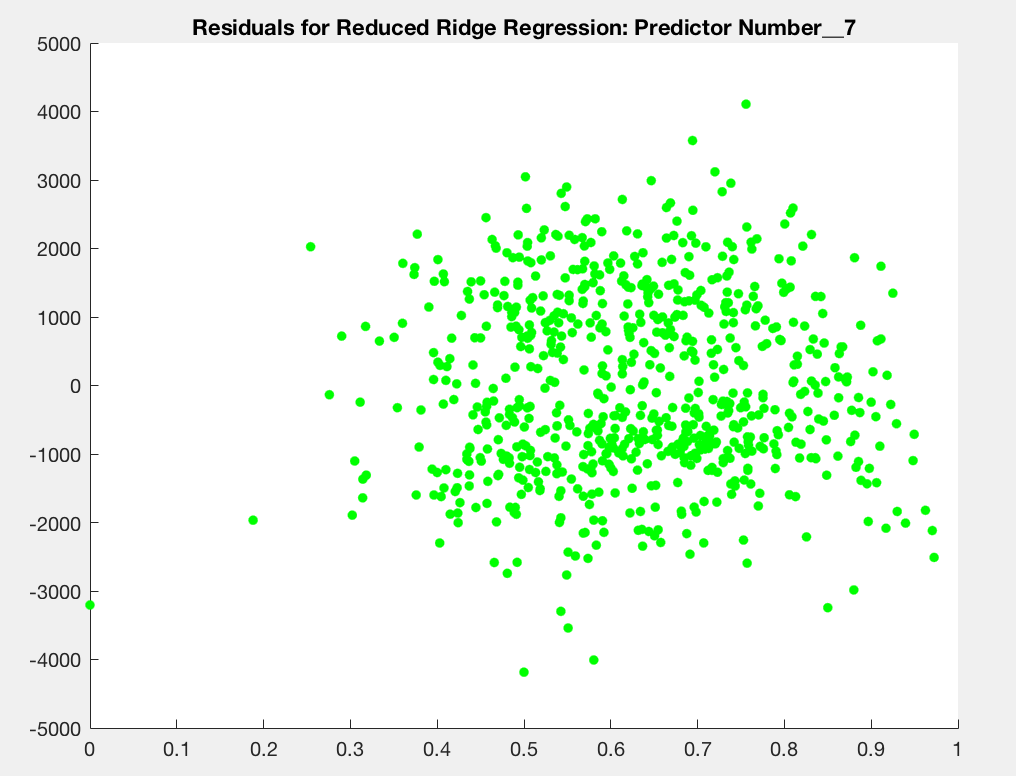
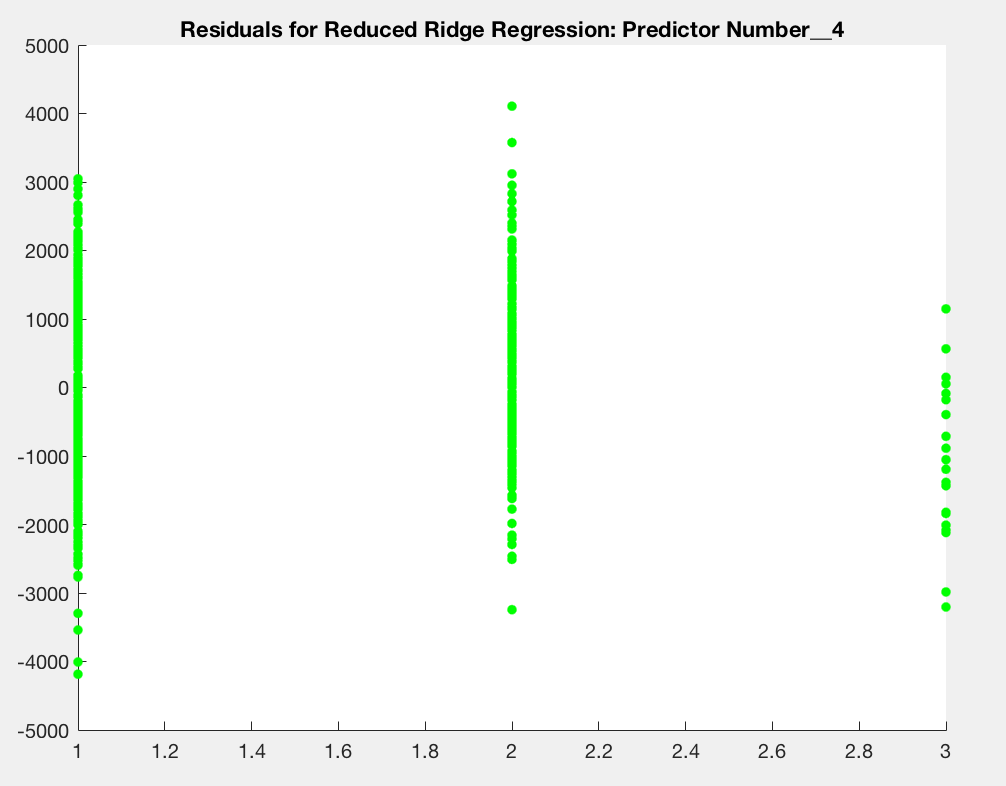
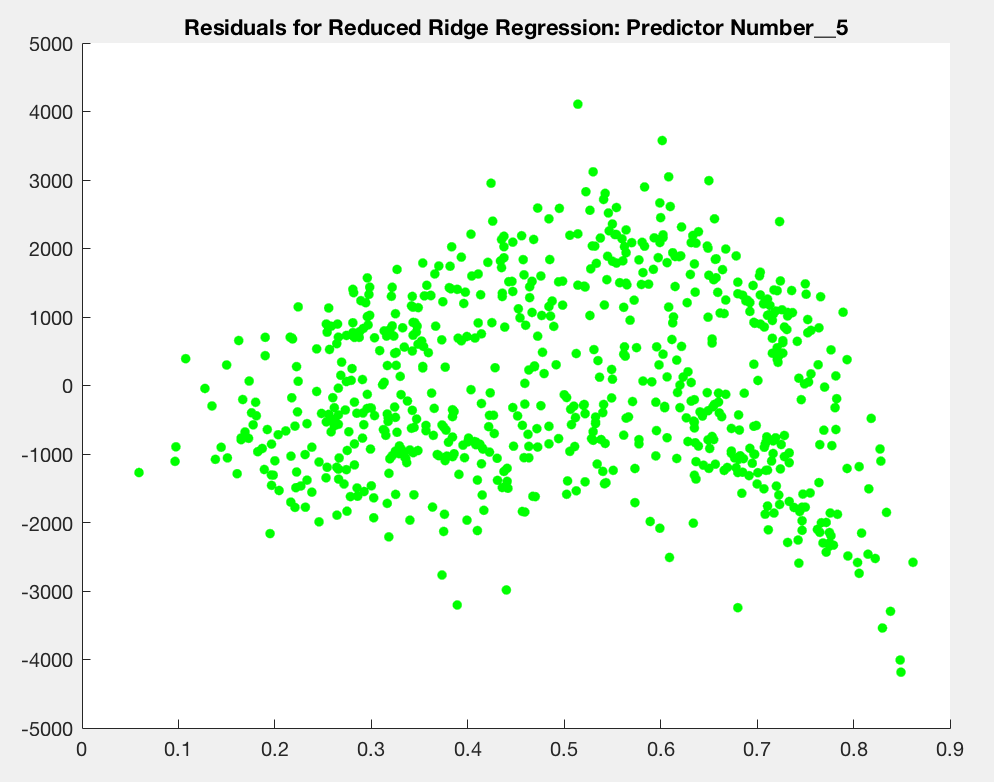
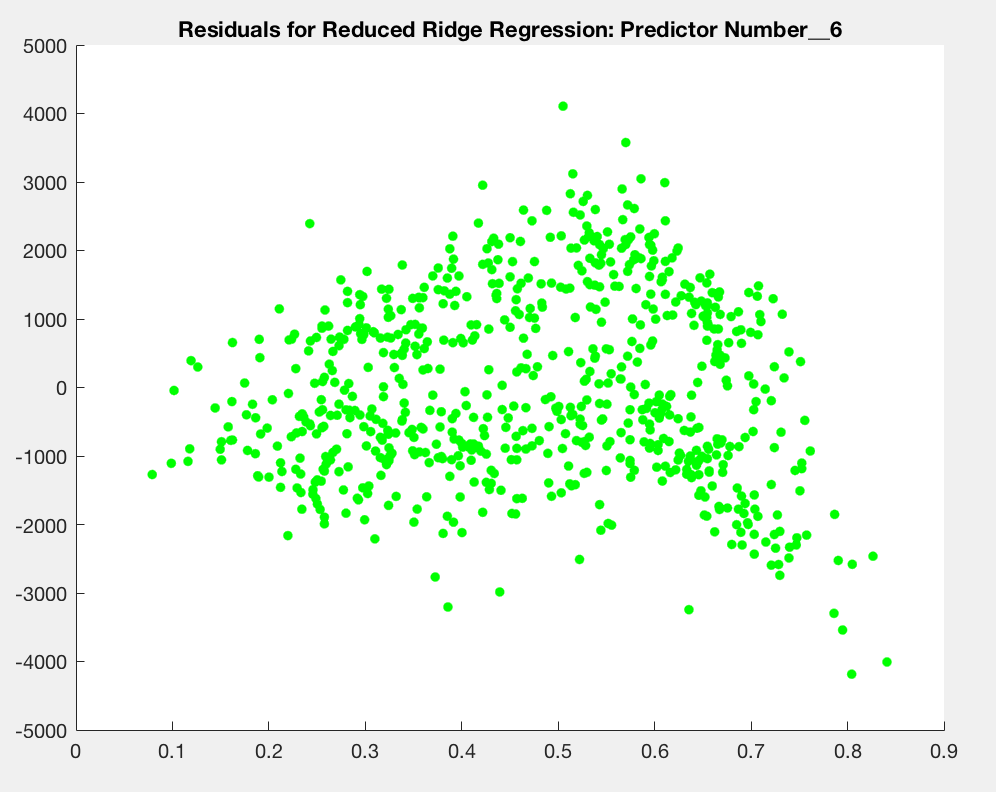
**FIGURES 29-33**





**FIGURES 33-41**

****

****

Bibliography:

[1] Fanaee-T, Hadi, and Gama, Joao, "Event labeling combining ensemble detectors and background knowledge", Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelberg, doi:10.1007/s13748-013-0040-3.