DATA 621 - HW1 - Regression Model

Samantha Deokinanan

September 3, 2020

Task

Using the training data set, build at least three different multiple linear regression models, using different variables(or the same variables with different transformations). Since we have not yet covered automated variable selection methods, you should select the variables manually (unless you previously learned Forward or Stepwise selection, etc.). Since you manually selected a variable for inclusion into the model or exclusion into the model, indicate why this was done.

Discuss the coefficients in the models, do they make sense? For example, if a team hits a lot of Home Runs, it would be reasonably expected that such a team would win more games. However, if the coefficient is negative (suggesting that the team would lose more games), then that needs to be discussed. Are you keeping the model even though it is counter intuitive? Why? The boss needs to know.

The Dataset

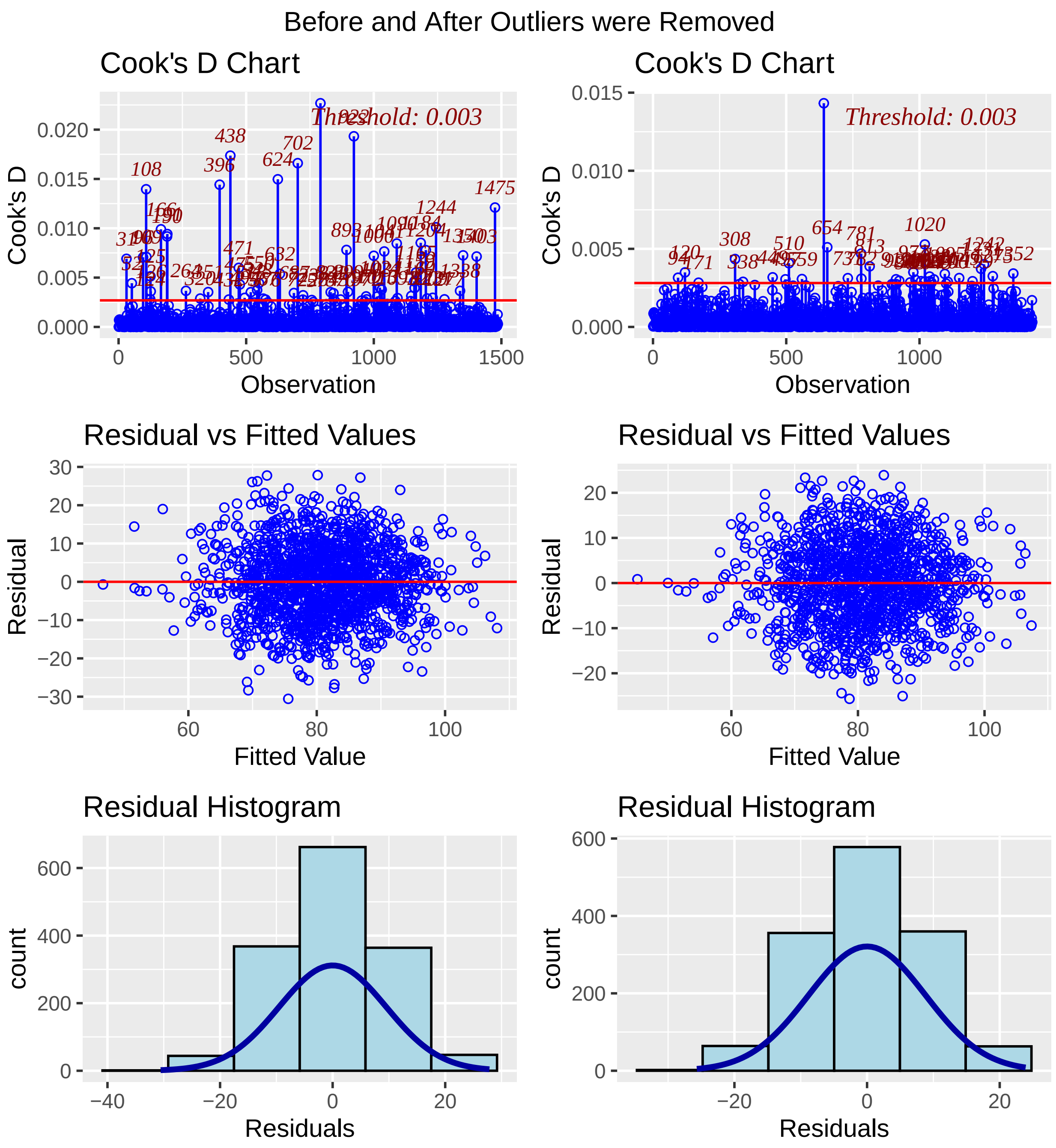
The data set contains approximately 2200 records. Each record represents a professional baseball team from the years 1871 to 2006 inclusive. Each record has the performance of the team for the given year, with all of the statistics adjusted to match the performance of a 162 game season. Aim is to predict the number of wins for the team.

Data Transformation, Outliers and Missing Data

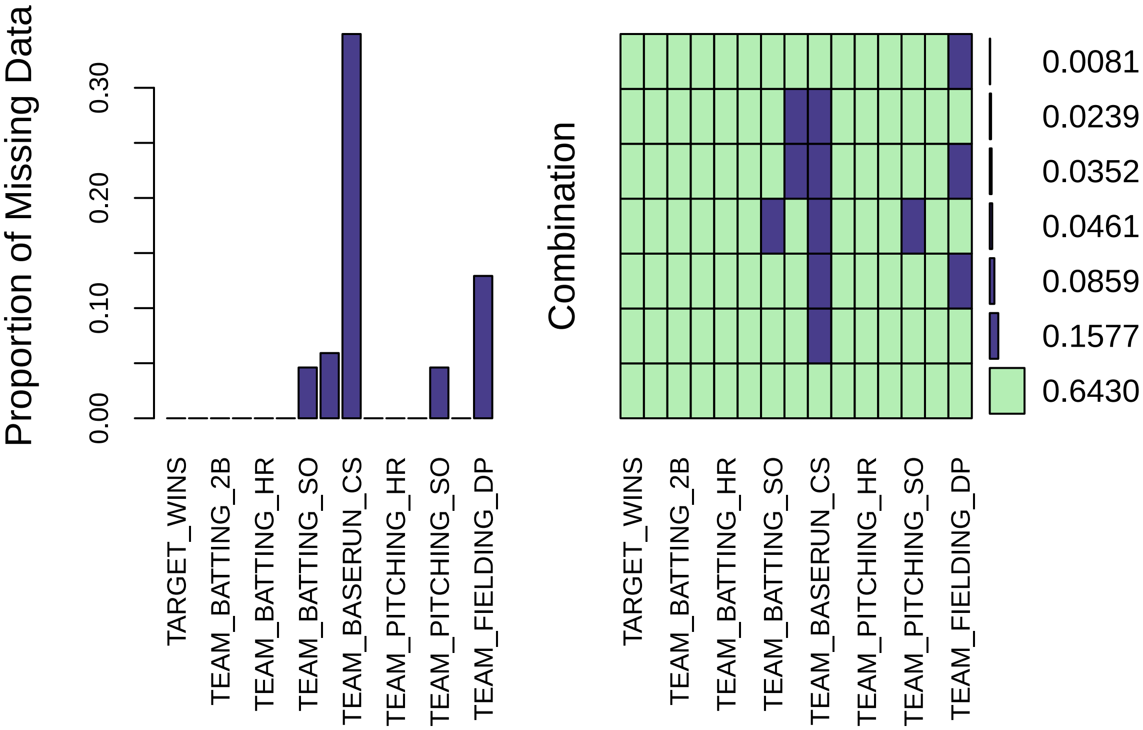
A quick exploration of the data, and it is already apparent that the `TEAM_BATTING_HBP` variable has nearly 92% of the its data missing. Since it would be very difficult to accurately impute such a large proportion, this variable will be excluded from further analysis.

Next, the outlier plots revealed that there are a few extreme values that can influence the analysis. Because the objective is so create a multivariate regression model, declaring an observation as an outlier based on a just one feature could lead to unrealistic inferences. Therefore Cook's distance is use to decide if an individual entity is an extreme value or not.

Cook's distance is a measure computed with respect to a given regression model and therefore is impacted only by the predictors included in the model. It computes the influence exerted by each data point on the response variable. Fitting the full model during this exploratory stage, the process remove the most influential points. Note from the plots that all influential observations are not necessarily outliers. Less than 3% of the data was removed.

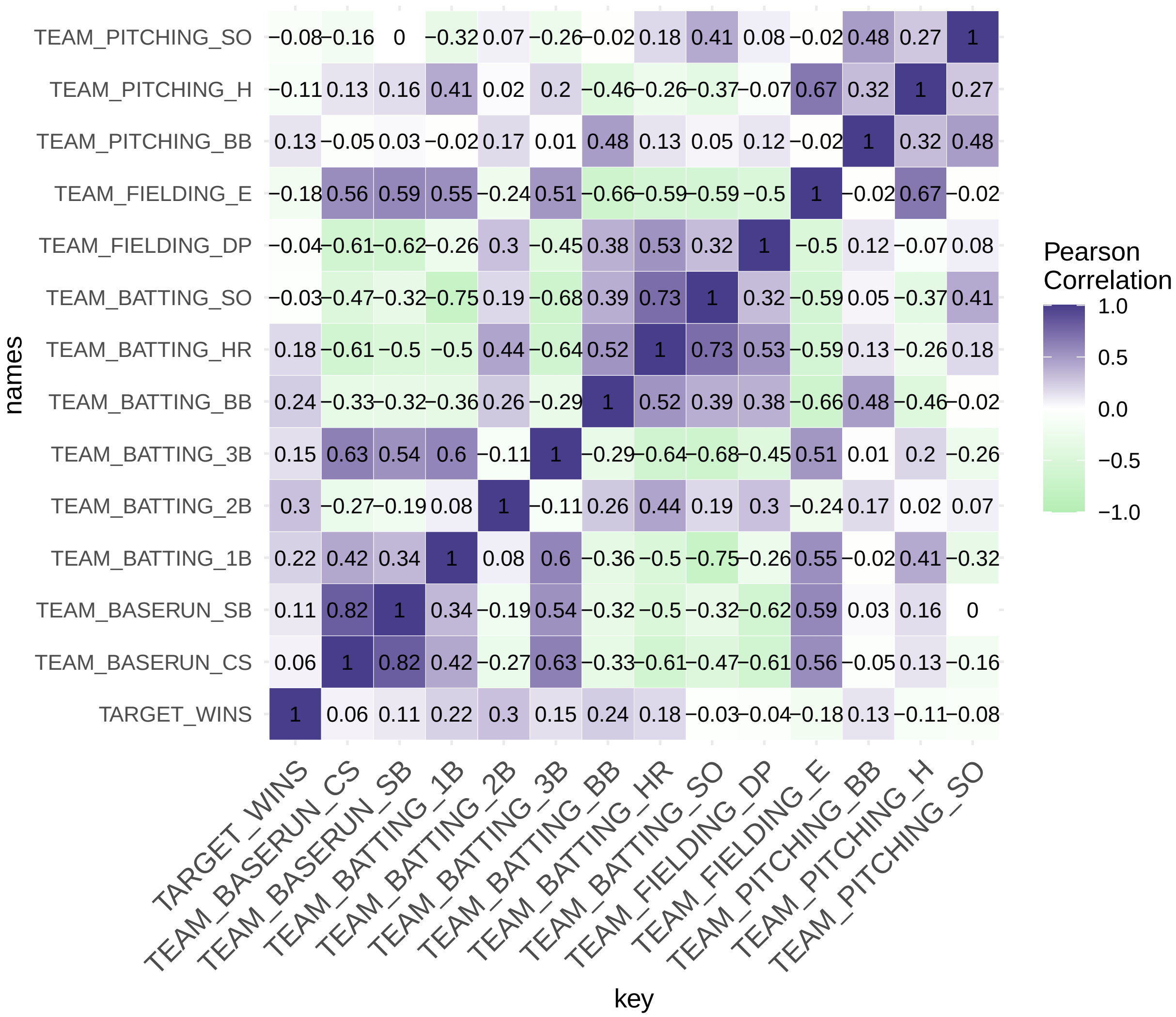


Now, an investigation is perform to calculate the amount of missing/imputed values in each variable. The plot helps to understand that almost 85% of the samples are not missing any information, whereas below states the percentage of missing values per variables. The method of choice to handle these missing values is imputation using predictive mean matching, to replace missing data with a randomly chosen value from several similar cases.



Initial Tests

Because the data consist of variable called `TEAM_BATTING_H`, base hits by batters, i.e. it is linear combination of H = 1B + 2B + 3B + HR, there are concerns of possible multicollinearity. Therefore, the single hits by batter found and base hits were removed. Next, a collinearity diagnostic test is done to examining the diagnostic output for variance inflation factor, tolerance, and Farrar-Glauber F-test. The F-statistic for the variable `TEAM_BATTING_HR` is quite high (42.1158) followed by the variable `TEAM_PITCHING_HR` (F-value of 33.8885). So, the test shows that there are multiple variables that will be the root cause of multicollinearity. Moreover, as expected, there are high partial correlations found to be statistically significant. As a solution to deal with multicollinearity, there are several remedial measures will be used as a result of this diagnostic test. Some included removal of highly correlated variables and stepwise regression analysis were done.



Lastly, because a robust regression process will be performed, higher order polynomials variables were introduced into the full model.

Regression Model: Stepwise Regression with Repeated k-fold Cross-Validation

A stepwise variable selection model is conducted to determine what are the variables that can help predict the number of wins for the team. The stepwise variable selection allows variables to be added one at a time to the model, as long as the F-statistic is below the specified α, in this case $\alpha = 0.05$. However, variables already in the model do not necessarily stay in. The steps evaluate all of the variables already included in the model and remove any variable that has an insignificant F-statistic. Only after this test ends, is the best model found, that is when none of the variables can be excluded and every variable included in the model is significant.

Here the dependent variable is the continuous variable, `TARGET_WINS`, and the independent variables are the full model to identify the most contributing predictors. In addition, a robust method for estimating the accuracy of a model, the k-fold cross-validation method, was performed evaluate the model performance on different subset of the training data and then calculate the average prediction error rate.

After the steps, the final model resulted below, with $adj.R^2 = 0.43$, suggesting that this model accounts for nearly 43% of the variation in the dependent variable with the independent variables, which is acceptable as a good model. With this method of stepwise regression, AIC (Akaike Information Criteria) quantifies the amount of information loss due to simplification. That is, based on the AIC, the final model outputted is the simplest model without impacting much on the performance.

r.squared	adj.r.squared	rsme
0.43	0.43	11.76

Prediction

Studying the coefficients of the model suggest that winning is in favor if the team batting hits more doubles, triples and home runs. Moreover, increase in the number of stolen bases, and a decrease in caught steals, double plays, error, and walks allowed would all lead to a win for the batting team. It is noteworthy that the model suggests that a decrease in single hits by batter and an increase in strikeouts by batters which seems counter intuitive. But these variables were kept because when a batter steps to the plate, the player is more likely to strike out than to get a hit. Trying to hit the ball out of the park will come with strikeouts but it will also increase the chances of hitting home runs (even 1B, 2B, 3B), and that is pretty good exchange that most teams are willing carry out.

	x
(Intercept)	129.9363340
TEAM_BATTING_2B	0.1020279
TEAM_BATTING_3B	0.2272429
TEAM_BATTING_HR	0.1127316
TEAM_BATTING_BB	-0.1797268
TEAM_BATTING_SO	0.0363787
TEAM_BASERUN_SB	0.0875630
TEAM_BASERUN_CS	-0.0672105
TEAM_FIELDING_E	-0.0804263
TEAM_FIELDING_DP	-0.3730934
TEAM_BATTING_1B	-0.0557781
TEAM_BATTING_2B_1	-0.0001514
TEAM_BATTING_3B_1	-0.0006398
TEAM_BATTING_BB_1	0.0001826
TEAM_BATTING_SO_1	-0.0000316
TEAM_BASERUN_SB_1	-0.0000584
TEAM_BASERUN_CS_1	0.0003098
TEAM_PITCHING_BB_1	0.0000031
TEAM_FIELDING_E_1	0.0000168
TEAM_FIELDING_DP_1	0.0008994
TEAM_BATTING_1B_1	0.0000435

Using the test data and the final step model, a comparison in the prediction statistic was conducted. It is apparent that the model's prediction are not too off from the training data. However, the predictions resulted in a higher kurtosis, thus there are expectation of greater extremity of the deviations, and not centered near the mean.

dataset	n	mean	sd	median	trimmed	min	max	skew	kurtosis	se
Training Data	2213	80.77	15.63	82.00	81.28	0.00	146.00	-0.41	1.13	0.33
Test Prediction	259	79.87	11.90	80.46	80.23	25.52	121.74	-0.67	4.04	0.74