Prediction Analysis of Team Attendance Data Using Multiple Regression Prediction Models

By Peter B Larsen
For the Tampa Bay Rays Position of Analyst; Strategy and Analytics

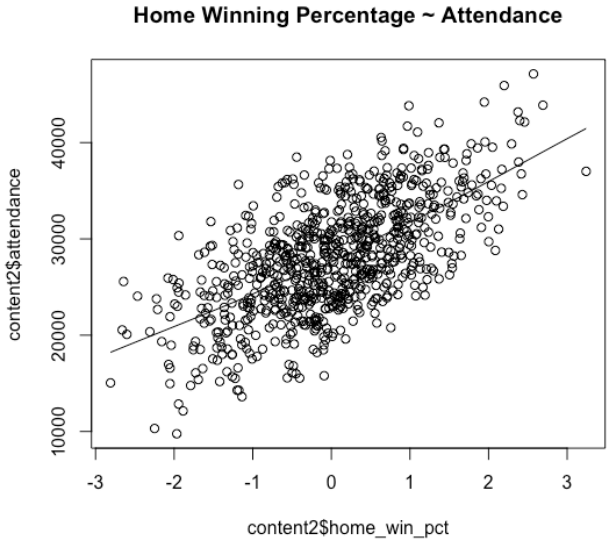
Using RStudio, the dataset 'data_for_exercise.csv', and the 'glmnet' library found at https://cran.r-project.org/web/packages/glmnet/index.html, three types of regression prediction models can be created: a basic linear regression model, a ridge regression model, and a LASSO regression model. Below, each models type is described and details of the modelling process and results are listed. At the end, a comparison of the models and selection of which one performs the best was chosen.

Linear Regression:

Linear regression in one or more variables works off the basis that there is 1 or more controlling variables and then 1 response variable, in the form of an equation like so:

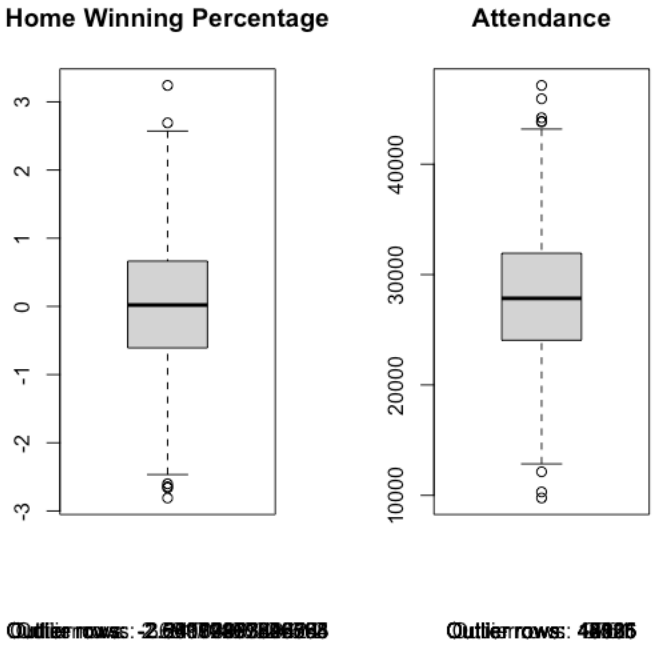
$$Y = \beta_0 + \beta_1 X + \epsilon$$

Here, the Y is the response variable, in our case attendance. The first term on the right side represents the intercept, or the baseline of the problem, being what one would expect if all predictors were constantly zero. The second term in the summation represents the regression coefficient multiplied by its respective predictor variable. Finally, the epsilon represents the error terms inherent in predictions. The goal of accurate modelling is to reduce the error terms to as close to zero for as many predictors as possible. The image below is a scatter plot of 'attendance' and 'home_win_pct'.



Using a Correlation matrix allows us to determine which values were affecting the attendance the most, with a score near 0 representing almost no effect, and a score nearer to 1 or -1 representing a positive or negative relationship between the variables. The matrix easily shows that 'home_win_pct' had a correlation of 0.63764692 with 'attendance', indicating that there may exist some connection. No other variable had above a |0.3| correlation with attendance, indicating that for linear regression, the only variables affecting 'attendance' would be just 'home_win_pct'. Interestingly, the correlation matrix is also a way to find multicollinearity, and the matrix does not contain any values above a |0.3| correlation, indicating that aside from 'home_win_pct' being correlated with 'attendance', the response variable, no other variables share a high correlation with each other. This will be important when it comes to the ridge regression model. The below model is the result of running the 'lm()' function in RStudio on our model.

Model 1: attendance = 27974 + 3810*(home_win_pct)



When modelling, there are often outliers when it comes to linear regression. In order to determine values that were considered outliers, boxplots can be used by the function 'boxplot.stats' as can be seen in the above image. A model could then be created using the dataset without the outliers, in an attempt to create a model that accurately explains the regular expected attendance rather than be thrown off by extremely poor or extremely excellent attendance nights. Of note: it is very improper to arbitrarily remove values, but for the purposes of modelling and examining the difference between a raw dataset and a clean dataset, the opportunity to see if the model could improve by removing 13 entries of the 810 original, or roughly 1.6% of the original set, was taken). A new correlation matrix was created, producing a new correlation value between 'home_win_pct' and 'attendance' of 0.61192028, which is lower than the original model. All models kept 'home_win_pct' in it's Z-score

Model 2 (outliers removed): attendance = 27947 + 3654*(home_win_pct)

Four forms of model fit analysis can be used to determine which model performs the best. R^2 and adjusted R^2 are used to explain the amount of variation in the data that the model can explain, on a scale from 0 being none of the variation is explained to 1 being all of the variation is explained. After performing a model summary, both models are not ideal. Model 1 has an R^2 of 0.4066 and an adjusted R^2 of 0.4059, while Model 2 has an R^2 of 0.3744 and an adjusted R^2 of 0.3737. The other two forms of model fit analysis are Akaike's information criterion (AIC) and the Bayesian information criterion (BIC), which work using the likelihood function and the sample size and number of parameters. Here, the lower the AIC and BIC are, the better the model. The AIC and BIC of both models were both incredibly high (above 15000), though the Model 2 performed slightly better on both, a change from the R^2 values.

Linear regression models seem to be struggling to explain all the variation inherent in the dataset, so an experiment was conducted to see how accurate a LR model could be in this instance by using 80% of the values as a training set for a new model and the remaining 20% to see how accurate said model could be as the test set (to see if the original models possibly were problematic due to overfitting).

Model 3: attendance = 28031 + 3949*(home_win_pct)

This model has a better R^2 and adjusted R^2 than either of the two previous models, at 0.4293 and 0.4284, and a much lower AIC and BIC (12000s). The thing to note here is that one cannot necessarily rely on these accuracy metrics as they are based on a smaller subset of the data, so they should not be compared. The accuracy of this model can be tested by seeing how it predicts the remaining 20% of the data kept as the test set.

Using the correlation accuracy, we can see that all three models actually produce the same correlation accuracy (55.58% correlation accuracy implies that the actuals and predicted values have similar directional movement in ~56% of situations). Therefore, it cannot be determined by this if the models are different.

Ridge Regression:

Least-squares regression seeks to minimize the sum of squared residuals (RSS):

$$RSS = \sum(y_i - y_i^*)^2$$

Here, y_i is the actual response value for the ith observation while y_i^* is the predicted response. Ridge regression seeks to do a similar approach, but instead minimize the equation:

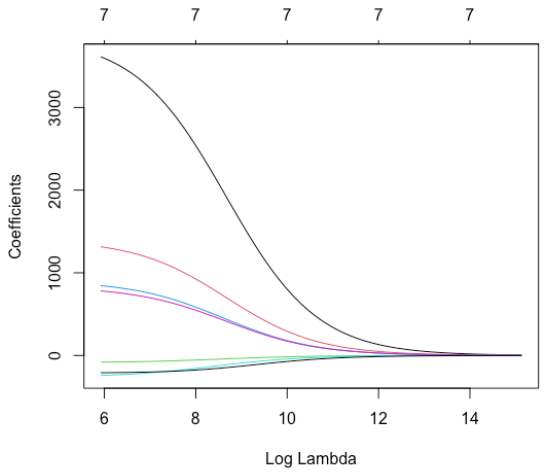
$$RSS + \lambda\sum\beta_j^2$$

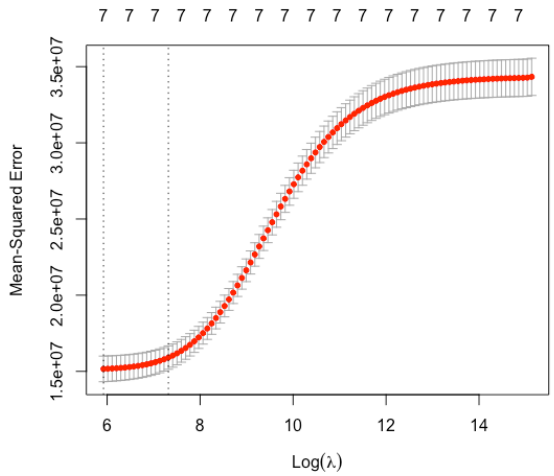
Here, the second term is the shrinkage penalty, which means the model's goal is to find a value for lambda which produces the lowest possible test MSE (mean squared error). Using the 'glmnet' library from the link at the beginning of this analysis, and the 'glmnet' function from said library, a ridge regression model can be created. Of note here: the differences between a ridge regression and a LASSO regression is the setting of the alpha value in the 'glmnet' function. An alpha of 0 produces a ridge regression model, an alpha of 1 produces a LASSO regression model, and an alpha between 0 and 1 produces an elastic net model. All have different uses. The ridge regression model is excellent for working around high multicollinearity, something that, as determined in the Linear Regression section, is not present to a high degree here in this problem. This should not affect the model too much, but it does not make use of the strengths of this specific model type.

If all seven predictor variables are used ('home_win_pct', 'away_win_pct', 'opposing_team', 'day_of_week_effect', 'temp','pct_season_completed','is_bobblehead'), the model produced looks like so:

Model 3: attendance = 25485.63347 + 3613.21804 (home_win_pct) + 1314.42453(away_win_pct) - 81.16213(opposing_team) + 844.43528(day_of_week_effect) - 241.74838(temp) + 780.36778(pct_season_completed) - 208.82847(is_bobblehead)

This model produces an R^2 of 0.5653589, the best value yet for any model. In order to examine the effectiveness of this model, one or more parameters can be removed to examine whether the variation explanation increases or decreases. A model which removed 'opposing_team' produced an R^2 value of 0.54932150, a reduction in effective variation modelling. A model which removed both 'opposing_team' and 'day_of_week_effect' produced an R^2 of 0.4609004, much lower than the full modelwith all predictors. Thus, the model with all predictors present appears to model 'attendance' the most accurately compared to the models with predictors removed. The images below show how the best lambda is found for a ridge regression model.





LASSO Regression:

LASSO regression is almost identical to ridge regression, but instead of seeking to minimize this:

$$RSS + \lambda \sum \beta_j^2$$

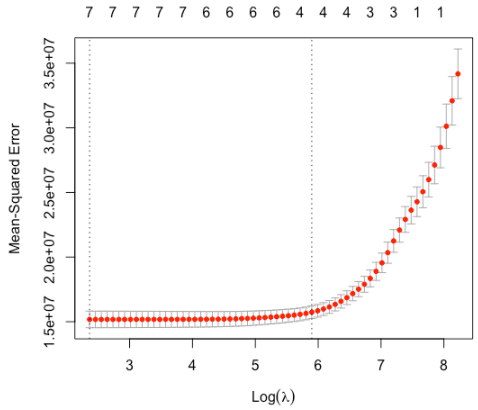
LASSO regression seeks to minimize this:

$$RSS + \lambda \sum |\beta_j|$$

The goal is to find lambda using k-fold cross validation, and using the 'glmnet' function from the library with the same name with an alpha value of 1 (see the previous section for an explanation of what changes to the alpha value does), a LASSO regression model is created. This

Model 4: attendance = 25334.59059 + 3835.04361(home_win_pct) + 1388.35820(away_win_pct) - 85.54363(opposing_team) + 895.63247(day_of_week_effect) - 250.09088(temp) + 795.53385(pct_season_completed) - 175.76394(is_bobblehead)

This model produces an R^2 value of 0.567456. The image below shows a plot of the model and how the best lambda is found for a LASSO regression model.



Overall Analysis:

As a whole, the LASSO regression model performed the best from an R^2 standpoint, with roughly 57% of all variation in the dataset being explained by Model 4. Linear regression really did not create an accurate model, even when attempting to train or controlling for errors. Ridge regression was a very close second to LASSO regression in terms of model accuracy. Unfortunately, the AIC and BIC metrics from linear regression don't have straight 1-1 equivalents for the models created using 'glmnet', so only R^2 valeus could be compared. As for the variables that seem to control 'attendance' the most, 'home_win_pct' seems to have the highest effect, which makes sense, as people want to go to a game where they can reasonably expect their home team to win. Other variables have an effect, but none nearly as large as that.