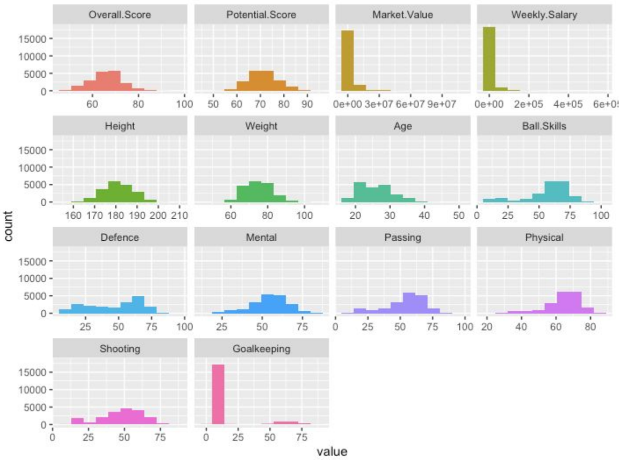
BAX 452 Assignment 3

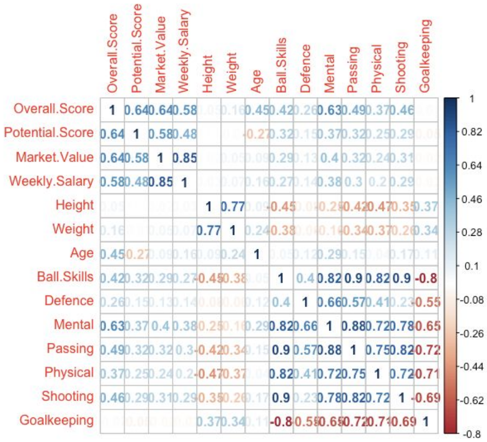
William Ma   Xiaochen Guo Jiaqi Lu

The purpose of this report is to build a model to predict the future market value of each player, with selected player features. This enables the UCD's soccer team manager to select the best player for a team with high future market value, to gain a commercial and strategic advantage over the other teams.

Section 1 Exploratory Data Analysis

The first step to our analysis is to conduct an exploratory data analysis on the dataset. We found that the data is relatively clean with no missing values in every column. Most variables are also found to be normally distributed, with no outliers. The histogram plots suggest that the Goalkeeping has a bimodal distribution, with two main data clusters (0~25, 45~80). This is because goalkeepers in a team tend to have higher goalkeeping points than the outfield players in the team. Meanwhile, we noticed that *Market. Value* and *Weekly. Salary* is left-skewed as their data range is too large. We will address these issues in the next section. We drew a heatmap to display the correlation between variables, and noticed that the *Market.Value* has a relatively higher correlation with *Overall.Score, Potential.Score, Weekly.Salary*. And there's a high possibility that multicollinearity exists among *Mental. Passing, Physical and Shooting* as they have strong positive correlations.

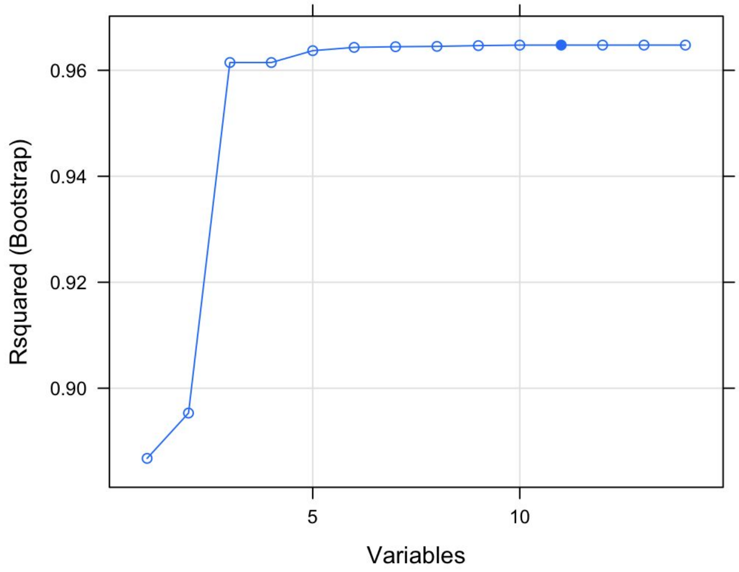




Section 2 Feature Selection

We conducted some data manipulation before diving into building the model. For the goalkeeping attribute, we created a dummy variable to indicate whether the player is a goalkeeper with 30 as the threshold. We also conducted log-transformation to *Weekly.Salary and Market.Value* to address the skewness. Moreover, we created one dummy variable for *foot_preference*, which is the only categorical variable with two categories: left foot and right foot.

After data manipulation, we checked the full model linear regression with the Log(Market.Value) as the dependent variable and included all features as independent variables. The model returned 0.96 for the Adjusted R^2, which is very high. Next, we used the recursive feature elimination (RFE), a.k.a. backward elimination, to apply initial feature selection. After recursive feature selection, we selected five features that contributed greatly to the increase of R^2 as shown in the plot.



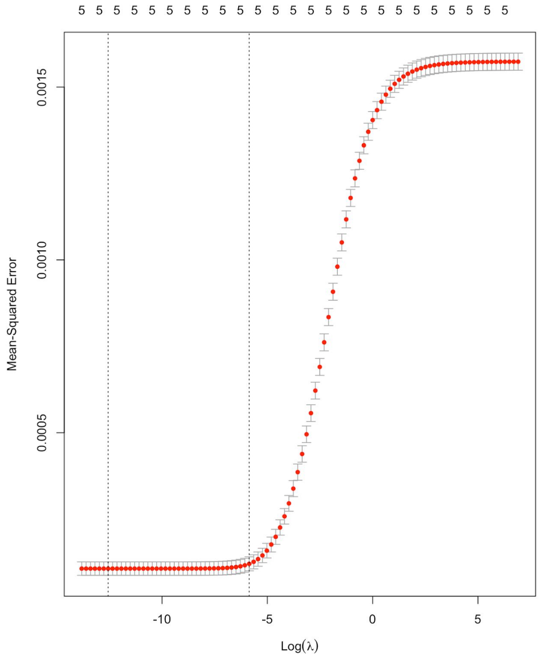
The following is our Reduced Linear Model:

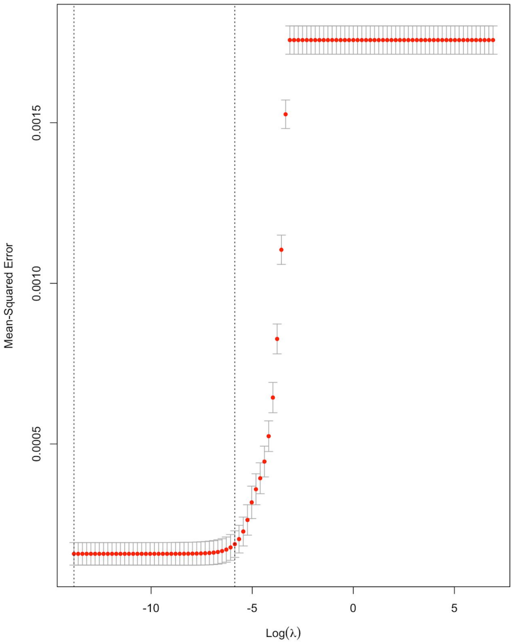
Market.Value ~ Weekly.Salary + Overall.Score + Age + foot_dummy + Goalkeeping.

The R^2 for the reduced linear model remained the same as 0.96.

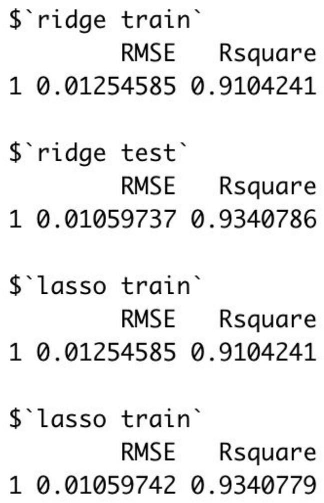
Section 3 Model Selection

We use ridge regression and lasso regression to perform regularization to enhance the prediction accuracy and interpretability of the resulting statistical model.
To see the prediction performance of the two models, we first partitioned the data into the training set and test set with a 70:30 ratio, respectively.
For ridge regression, we use cross-validation to determine the optimized lambda, which is 3.511192e-06. We use AIC to determine the best lambda for lasso regression, and the lambda is 1e-06.





For the training set, RMSE and R^2 for the ridge model and lasso model are equal, for the test part, RMSE of the ridge model is slightly lower and R^2 is slightly higher, which indicates that the performance of the ridge is a bit better.

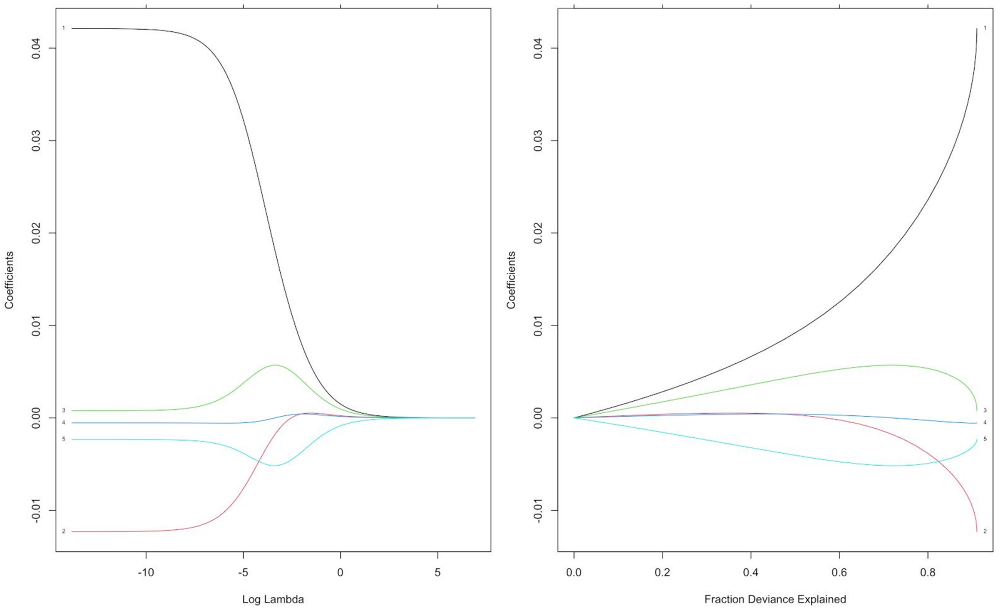


```
$`ridge train`
        RMSE   Rsquare
1 0.01254585 0.9104241

$`ridge test`
        RMSE   Rsquare
1 0.01059737 0.9340786

$`lasso train`
        RMSE   Rsquare
1 0.01254585 0.9104241

$`lasso train`
        RMSE   Rsquare
1 0.01059742 0.9340779
```

The lasso regression model is penalized for the sum of absolute values of the weights. It tends to give sparse weights (most zeros) because the L1 regularization cares equally about driving down big weights to small weights or driving small weights to zeros. Ridge regression penalizes the model for the sum of the squared value of the weights. Thus, the weights not only tend to have smaller absolute values and more evenly distributed but also tend to be close to zeros. Since we only have 5 variables to build models, Weekly.Salary, Overall. The score, Age, foot_dummy, Goalkeeping, and we think that all these five variables are important to the market value prediction.

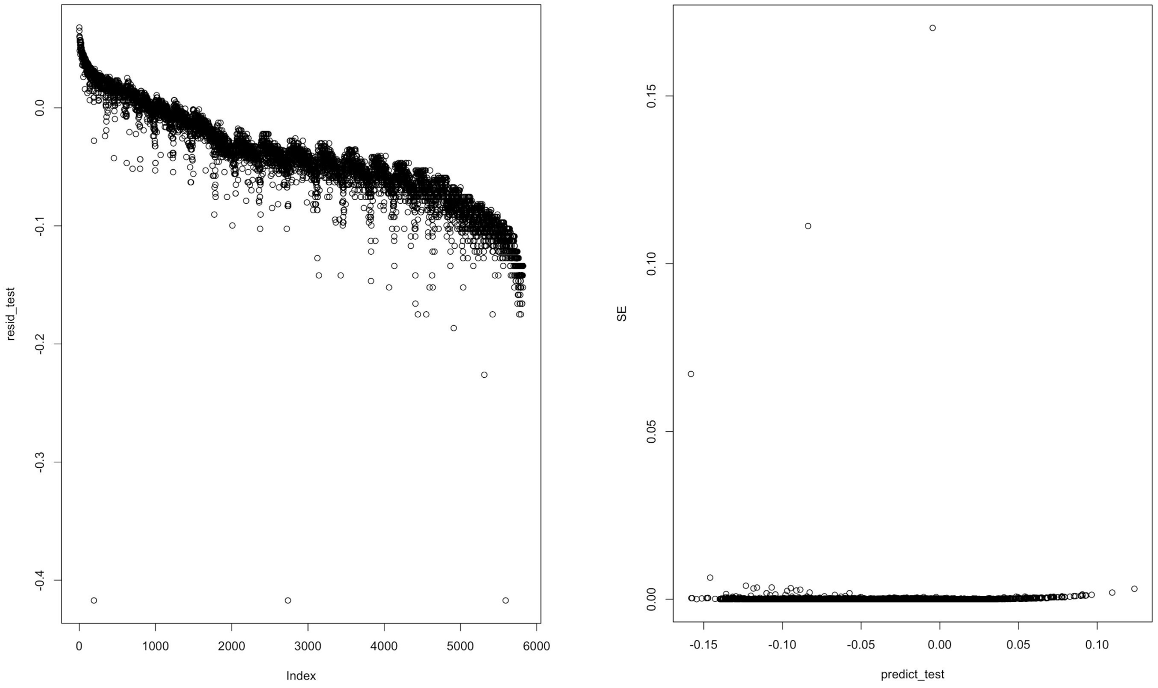
Combining these factors, we prefer the ridge regression model over the lasso regression model.

Section 4 Model Evaluation



As we can see from the fit plots of the Ridge Model above, the ridge model shrinks all the coefficients towards 0.

When it comes to the residual, without considering heteroscedasticity, the points in the residual plots are near zero. This shows the high fitting degrees.



As for the standard error of the two models, the standard error is nearly zero, which indicates that these two models have good predictions.

Section 5 Conclusion

Our prediction process is divided into data exploration, variables selection, model selection, and model evaluation. In the first section, we check the correlation between all variables and the distributions of each variable, to determine which variable should be chosen and if variable transformation is required. In the next section, we selected five variables to build the model using RFE. In the following section, we reviewed the probability of overfitting. Collecting too many data points in one population could cause overfitting, since there are 19402 soccer players in this dataset, we use regularization techniques to solve this problem. By using ridge and lasso regression, the model was modified, reducing model complexity and the risk of model overfitting. Although the model complexibility is reduced, the predictability with ridge and lasso regression remains very good. However, there are some limitations to our model, as the residual plots point towards heteroskedasticity as an issue that needs to be addressed. There are many reasons that could cause heteroskedasticity, such as the lack of variables. Since only 5 variables contributed greatly to enhancing the predictability of the model, this may be the case why heteroskedasticity exists. We can try to introduce more variables that are outside of the dataset to the model, and check if the heteroskedasticity is eliminated.