Regression Analysis of Wins per Season

Michael Cambaliza
Brendan Morrison

May 5, 2020
Final Project
PSTAT 126

Introduction

Our project will be focused on estimating the amount of wins a baseball team achieves, given the variables in the data set “Lahman” provided by the Lahman Baseball Database. We will study whether wins can be predicted by our predictors. We would like to figure out which of our predictors have the greatest impact on wins. The dataset “Teams” from the package “Lahman” contains 48 variables and 2895 observations. We decided to exclude variables that we deemed irrelevant for our model which included teams before 1961 because of shortened seasons and exclude the 1972, 1981, 1985, 1994, and 1995 seasons because of shortened seasons as well due to strikes. Including this data would have skewed our results and models. All of the variables shown are based on a given season by a given team, not totals around the league that year. The variables we feel that would help us in our research are *Wins* which represents the total amount of wins a team achieved in a season; *Hits* are the total amount of hits a team reached; *ERA*(Earned Run Average) which is the average amount of runs a team allowed per game to their opponent; *Hits Allowed* are the total amount of hits the opposing team had against a team; *Home Runs* represents the total amount of home runs a team had; *Runs* means the total amount of runs a team scored ; *Runs Allowed* represents the total amount of runs the team allowed the opposing team to score; and *Attendance* which represents the total amount of fans at their home games. We set the variable *Wins* as the response of this project.

Question of Interest

I. Can a team's amount of wins be predicted by hits, ERA, hits allowed, home runs, runs, runs allowed, and attendance?

Regression Method

The first thing we need to do to create a sufficient model is to refine our base model (if needed) to only include predictors that are relevant and make our model better. We will do this using stepwise regression and confirm the strength of our model with best subsets regression.

Once we have the relevant predictors in our model, we need to make sure it follows the LINE conditions of linear regression before we can start to answer our question of interest. We will use residual analysis to make sure our model follows all of the line conditions. If we see any issues with our predictors we will attempt the correct transformation if we deem it necessary.

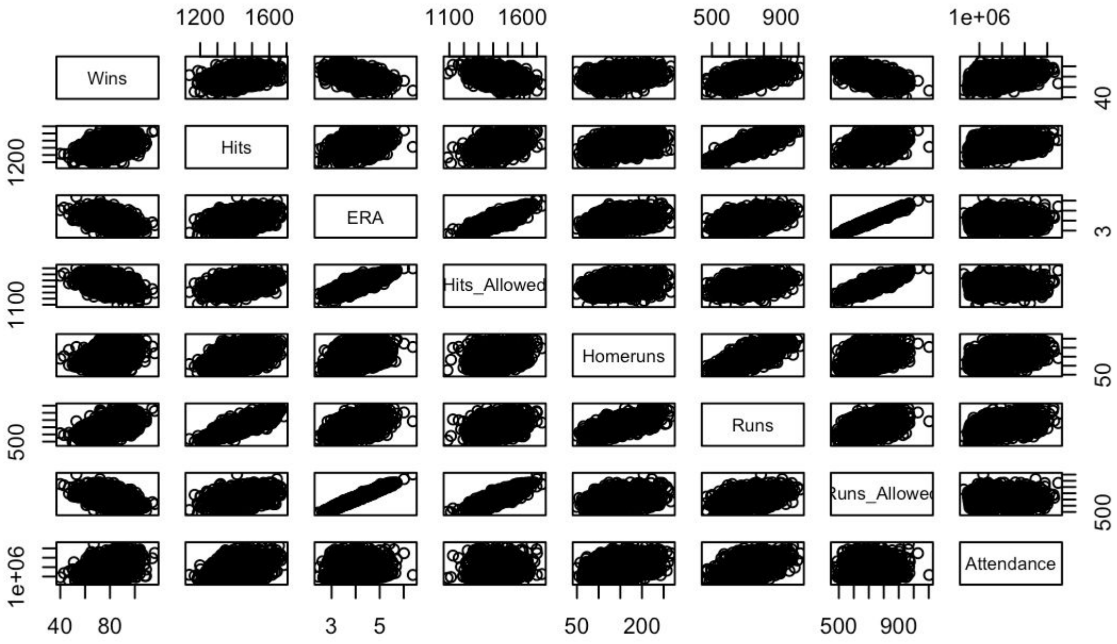
After clearing our four LINE conditions we can then proceed to answering our question of interest by reviewing our coefficient of determination in our model and check if the predictors we chose for our model help us indicate the amount of wins a team achieves per season with the p-value.

Regression Analysis, Results, and Interpretation

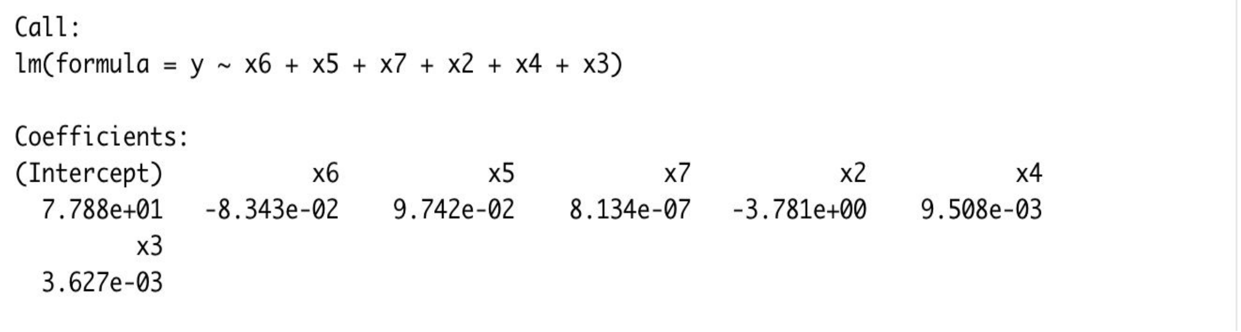
To start off our analysis we create the following model and our variables:

- y = Wins
- x1 = Hits
- x2 = ERA
- x3 = Hits Allowed
- x4 = Home runs
- x5 = Runs
- x6 = Runs Allowed
- x7 = Attendance

The Y (Wins) variable is what we are trying to predict with x1,x2,...,x7. We first plotted all of our predictors against our response variable *Wins* with the pairs() function in R. This will show us the relations of the predictors on response as well as predictors on eachother. The results we got back are shown below:



We use a stepwise function to perform a series of hypothesis tests comparing reduced and full models, showing if any predictors are irrelevant in predicting *Wins.*We will use the step() function in R to help us with determining appropriate variables.



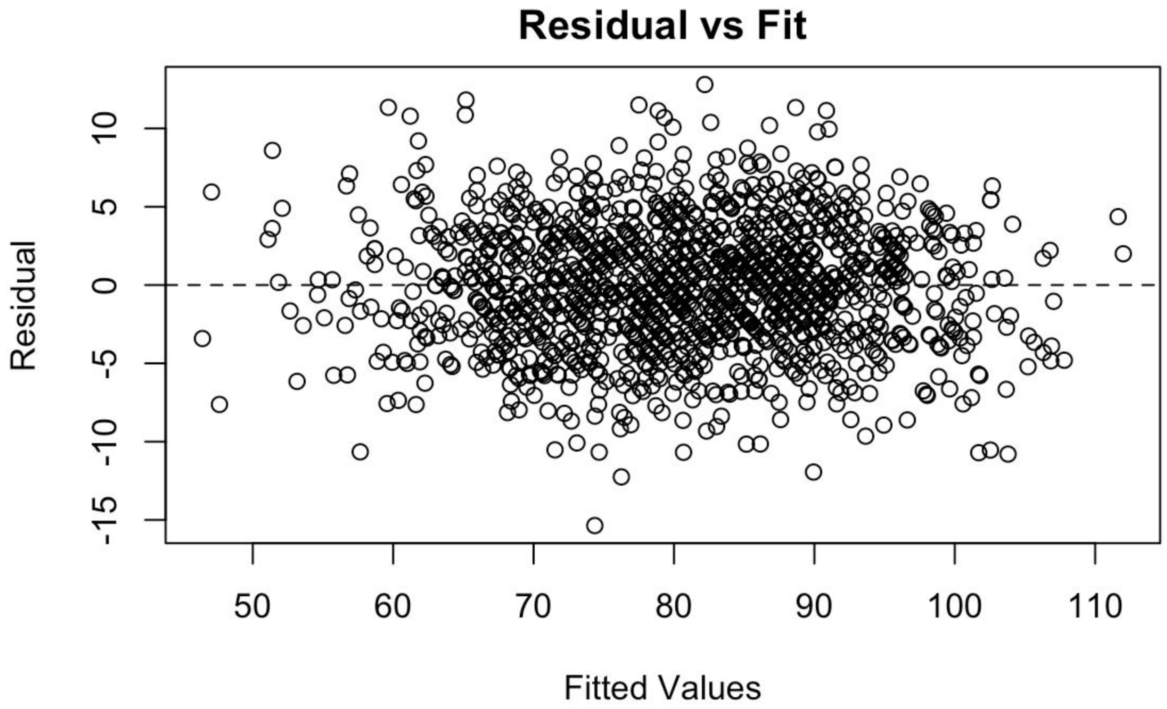
```
Call:
lm(formula = y ~ x6 + x5 + x7 + x2 + x4 + x3)

Coefficients:
(Intercept)           x6           x5           x7           x2           x4  
  7.788e+01   -8.343e-02    9.742e-02    8.134e-07   -3.781e+00    9.508e-03  
         x3  
  3.627e-03  
```

When running the step() function on our model, it returns the model (y ~ x2 + x3 + x4 + x5 + x6 + x7) as the best model.

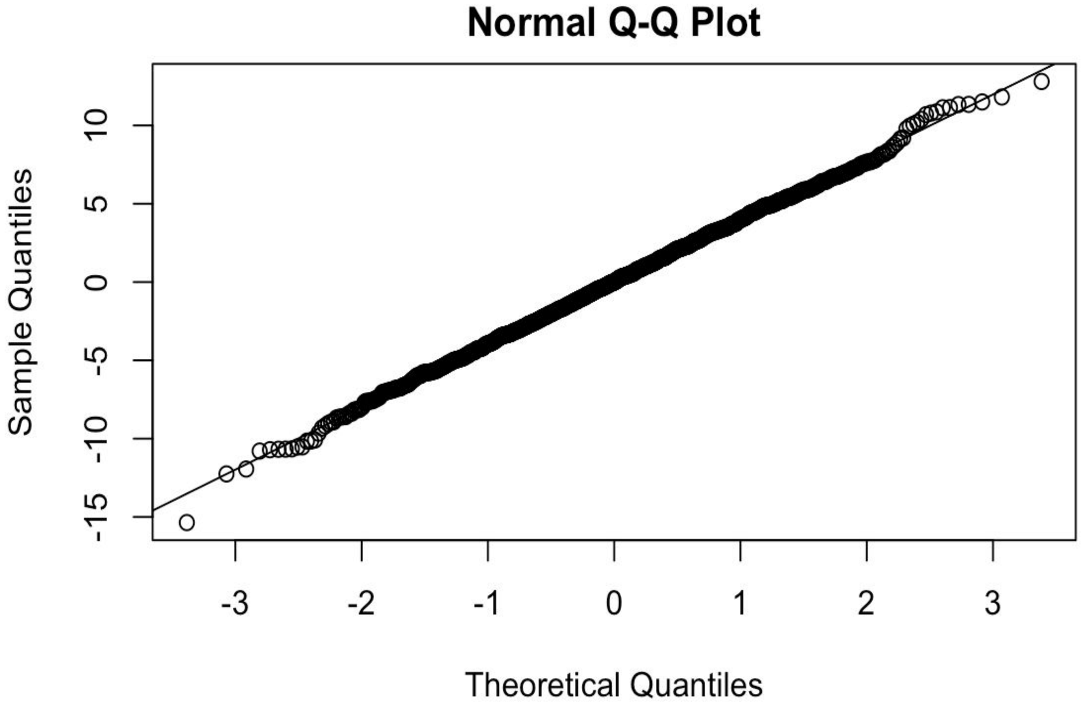
The leaps function gives us a list of the best models for each number of predictors (model with one predictor, two predictors...etc) and their respective adjusted R^2 values. The model with the greatest adjusted R^2 value is the best model. When running the leaps function on our data, we can see that the model including all of our predictors except x1(*Hits*) is the best model for predicting the number of *Wins*, much like what we saw when running the step function which further solidifies our model as appropriate for our question of interest.

Now that we have a good model for predicting *Wins*, we will perform a residual analysis in order to make sure our model follows the four LINE conditions required for linear regression. In order to test our model for linearity and equal variance we have to create a residual vs fit plot.



Our residual vs fit plot does not have any sort of non linear pattern or fanning effect. It is relatively uniform and randomly scattered about 0. This plot tells us that our model's predictors have linear relationships with the response as well as equal variances.

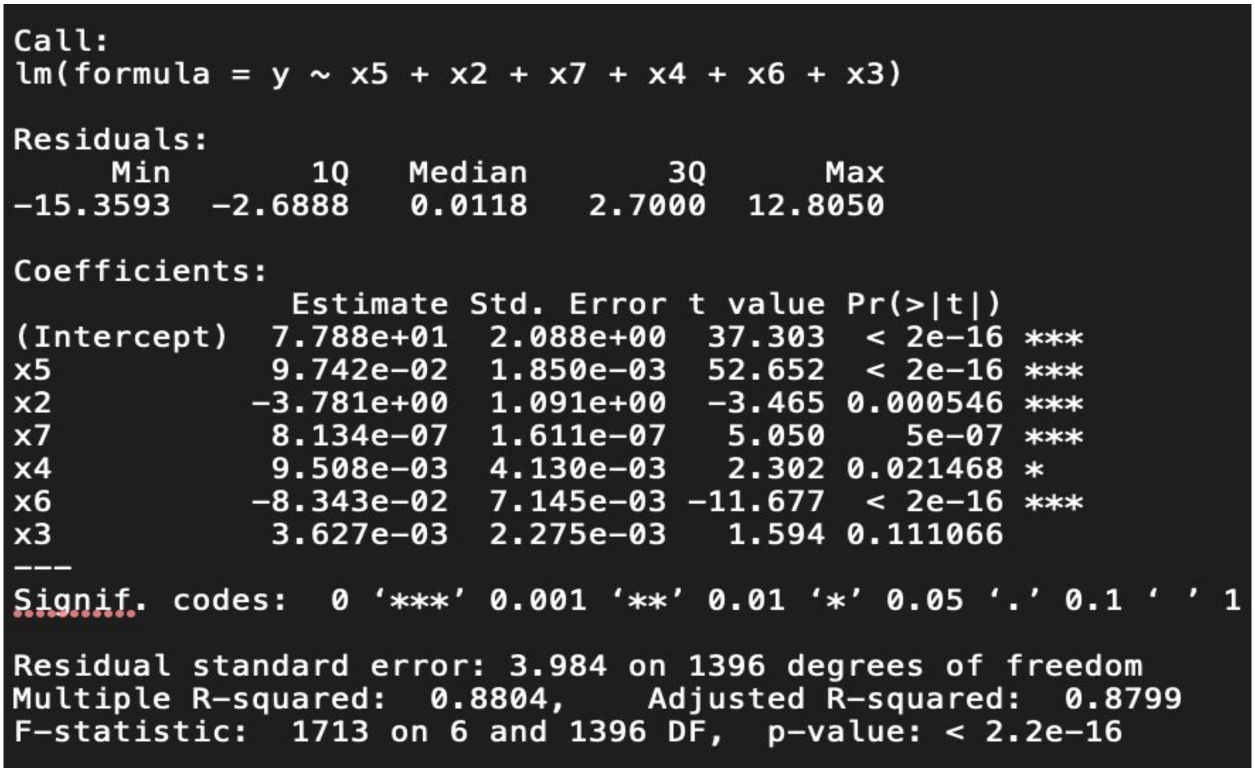
In order to test for normality we create a Q-Q plot to see if anything is abnormal and prevents us from having a normally distributed model.



Since the plot falls for the most part on the line, our model is normally distributed. There does seem to be a few outliers, but since it is so few, the effect is negligible.

Since our model follows the line conditions, there is no need to transform the values to achieve linearity.

Now that we have a well fitted linear model for predicting wins we need to see if we can accurately predict the number of Wins based on ERA, hits allowed, home runs, runs, runs allowed, and attendance.



```
Call:
lm(formula = y ~ x5 + x2 + x7 + x4 + x6 + x3)

Residuals:
     Min       1Q   Median       3Q      Max 
-15.3593  -2.6888   0.0118   2.7000  12.8050 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.788e+01  2.088e+00  37.303  < 2e-16 ***
x5           9.742e-02  1.850e-03  52.652  < 2e-16 ***
x2          -3.781e+00  1.091e+00  -3.465 0.000546 ***
x7           8.134e-07  1.611e-07   5.050    5e-07 ***
x4           9.508e-03  4.130e-03   2.302 0.021468 *  
x6          -8.343e-02  7.145e-03 -11.677  < 2e-16 ***
x3           3.627e-03  2.275e-03   1.594 0.111066    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.984 on 1396 degrees of freedom
Multiple R-squared:  0.8804,    Adjusted R-squared:  0.8799 
F-statistic:  1713 on 6 and 1396 DF,  p-value: < 2.2e-16
```

When we run the summary function on our model we can see that we have an adjusted R^2 value of 0.8799, meaning that about 88% of the variability of the number of wins is explained by our predictors. The very small p values in the table indicate that all the predictors in our model are relevant in the prediction of the number of wins. Most of our beta values are what we would expect in that we have positive beta values for runs, attendance and home runs implying a positive correlation and negative beta values for ERA and runs allowed implying a negative correlation. The beta value for hits allowed was not what we expected in that it implies a positive correlation between hits allowed and number of wins, but this is not surprising because the corresponding p-value is relatively large.

Conclusion

In conclusion, it is clear that the number of wins by a given baseball team can be accurately predicted by the ERA, hits allowed, homeruns, runs scored, runs allowed, and attendance. Additionally, we could say that the number of hits allowed has less of an impact on the total number of wins than the rest of the variables in observation of their respective p-values. The data shows that if a team wants to maximize their win potential they should invest in players who produce the most runs and players who prevent the most runs from scoring, not players who just hit home runs and gain a ton of hits. The variable that is the most surprising in our data is the attendance variable because the amount of people in a stadium should not have any influence on how well a team plays because they have no direct impact on each play, but logically if a team is winning more people show up to the ballpark so it would make sense that a higher attendance would reflect the amount of winning a team is doing.

Appendix

```
# Import dataset and packages
library(Lahman)
library(leaps)
data(Teams)
# Reducing dataset to the variables we want to use. We also excluded seasons in
1972, 1981,1985, 1994, 1995 , and pre-1961 due to an unequal number of games
per season.

teamdata <- Teams[c(1360:1589,1614:1813,1840:1917,1944:2152, 2210:2895),
c('W','H', 'ERA', 'HA', 'HR', 'R', 'RA', 'attendance')]
colnames(teamdata) <- c('Wins', 'Hits', 'ERA', 'Hits_Allowed', 'Homeruns', 'Runs',
'Runs_Allowed', 'Attendance')

head(teamdata)

# Creating a matrix plot of the variables to see relationships of predictors on
response
# as well as relationships between predictors.
pairs(teamdata)
```

```
# Performing stepwise regression to reveal any unnecessary predictors and further
refine our model.
y <- teamdata$Wins
x1 <- teamdata$Hits
x2 <- teamdata$ERA
x3 <- teamdata$Hits_Allowed
x4 <- teamdata$Homeruns
x5 <- teamdata$Runs
x6 <- teamdata$Runs_Allowed
x7 <- teamdata$Attendance
mod_reduced = lm(y ~ 1)
mod_full = lm(y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7)
step(mod_reduced, scope = list(lower = mod_reduced, upper = mod_full))

# Running the leaps function to make sure that the model we chose from the step
function is
# the best possible model from the set of predictors.
leaps(cbind(x1, x2, x3, x4, x5, x6, x7), y, method = 'adjr2', nbest = 1)

# Creating a residual vs. fit plot
fit <- lm(y ~ x5 + x2 + x7 + x4+ x6 + x3)
yhat = fitted(fit)
e = y - yhat
plot(yhat, e, xlab = 'Fitted Values', ylab = 'Residual', main = 'Residual vs Fit')
abline(h = 0, lty = 2)

# Creating a Q-Q plot on the model
qqnorm(e)
qqline(e)

#Running the summary to see correlation of coefficients,R^2 , and p-values
summary(fit)
```