American League Dilemma

AN INSIDE LOOK ON THE EFFECT OF THE DESIGNATED HITTER RULE

By: Mitch Sutrick, Brett Zahn, Brent Nicolet,
Saisharan Chimbili, and Tyler Minushkin

Some of the feedback we received from our presentation and how we handled it can be seen below.

- **Put NL in the same model (Dummy variable for league)**
  - **We go more in depth with the NL starting on page 1 where we show if there was a significant change in Percent Change of Attendance between AL and NL.**
  - **We changed our first regression to include an AL/NL dummy variable (Page 4).**
- **Use change in attendance for AL and NL as figure 1, other figures are diagnostic plots, add residual plots (maybe)**
  - **We moved this to our first figure to show more clearly why the DH rule is an important addition to the American League.**
  - **We moved this to our Figure 1 to show that attendance spiked from 1972 to 1973 for American League as a whole.**
- **We included two different models. One to test for the % change in attendance in relation to the league. Our second test involved % change in attendance in the AL in relation to designated hitter home runs**

Introduction

One of America's favorite traditions that dates back to the 19th century is baseball. From having backyard games to stadium-filled crowds, America's favorite pastime has grown in popularity. Because of this, teams were forced to have bigger stadiums or relocate to keep with the demand of the game. One factor that creates demand for a franchise is wins. The more wins a team has, the higher the franchise's total home attendance will be.

However, fans of the game do not just want to see their team win. Fans want to be entertained for nine innings and see their team come out with a victory at the end. So what draws people in and excites them the most? Hits, which range from singles, doubles, triples to home runs which are one of the greatest spectacles of the sport due to their difficulty. After the 1972 season, the designated hitter rule was implemented into the MLB to replace pitchers for hitters in the American League (AL). Because of this, hits were increased once pitchers stopped batting. Thus, AL teams with high number of hits had a higher total attendance once the rule was implemented.

Overall, our analysis does not show that the percentage change in attendance between 1972 and 1973 was significantly higher for AL teams than NL teams. Additionally, our analysis does not show that AL teams with more DH home runs had a higher percent change in attendance in 1973.

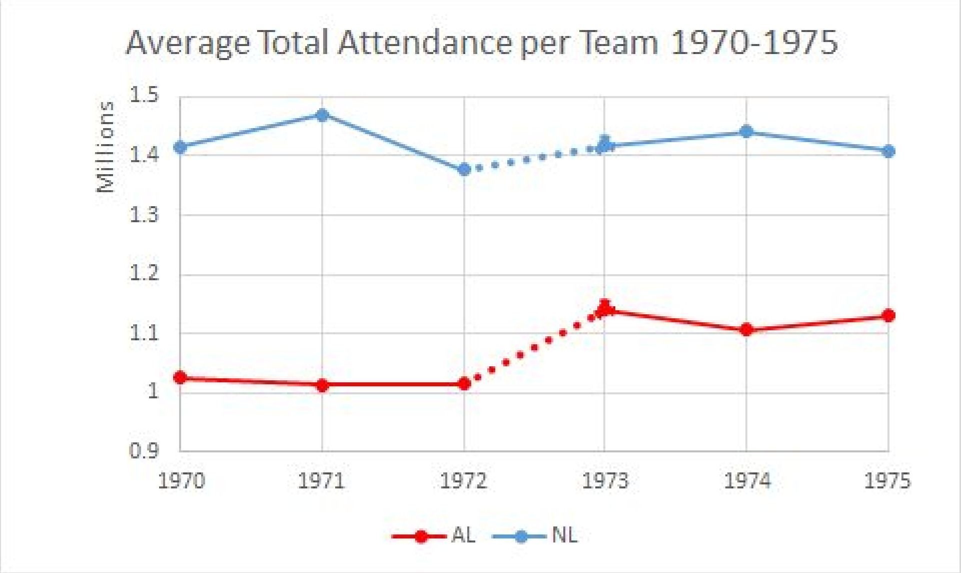


Figure 1a: The average total season attendance for teams in NL (blue) and AL (red) from years 1970-1975.

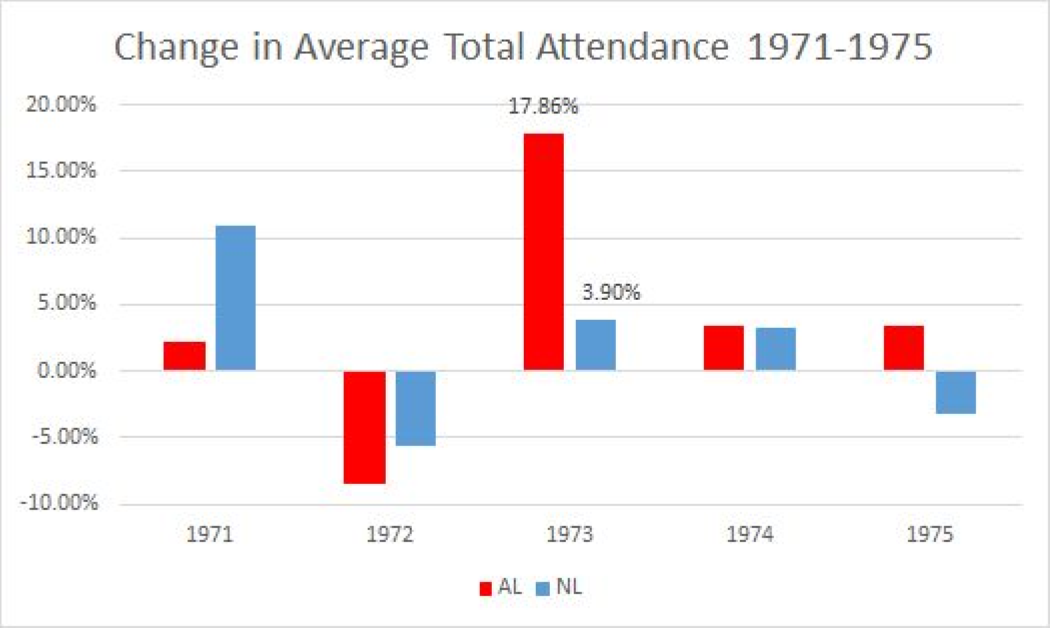


Figure 1b: The percentage change in average total season attendance between seasons for NL (blue) and AL (red) for years 1971-1975

Methods

1. **Data Sources**
   a. **"BASEBALL DATA"**
   http://seanlahman.com/baseball-archive/statistics/
   The data set used was created and updated yearly by Sean Lahman and Ted Toracy. The database is maintained through the use of Chadwick Baseball Bureau's Retrosheet, a tool that updates play-by-play throughout the MLB season. The data is then groomed by another of their softwares, Chadwick, that is used in baseball scorekeeping and keeping track of, and calculating, stats. Git is then used to track any changes in the data obtained. Finally, cross reference register is used to keep track of players, managers, and umpires from professional, college, and international leagues. We looked at NL and AL data for home runs from 1972-1973.
2. **Data Cleaning**
   a. We take the Excel file from the data source. We then obtained the number of home runs hit by the designated hitter for each AL team for the 1973 season and added it to the excel file. Additionally, we calculate the percentage change in total

attendance for each team for the 1973 season from the 1972 season and add this as a column to the excel file. We then import the file to R Studio and cut down our data to only contain the information for the year 1973 and cut the table to contain only the variables we want: percent change in attendance, league, and Designated Hitter home runs. We then make a dummy variable for league called AL and have a value of 1 set for AL and a value of 0 for NL.

3. Linear Model

a. *Variable Definitions:*
   i. *Percent_Change:* % change in attendance from 1972 to 1973
   ii. *Percent_Change_AL:* % change in attendance for AL from 1972 to 1973
   iii. *cat_league_AL:* dummy variable that is equal to 1 if American League (AL) and 0 if National League (NL)
   iv. *DH_Home_Runs:* home runs hits by a designated hitter

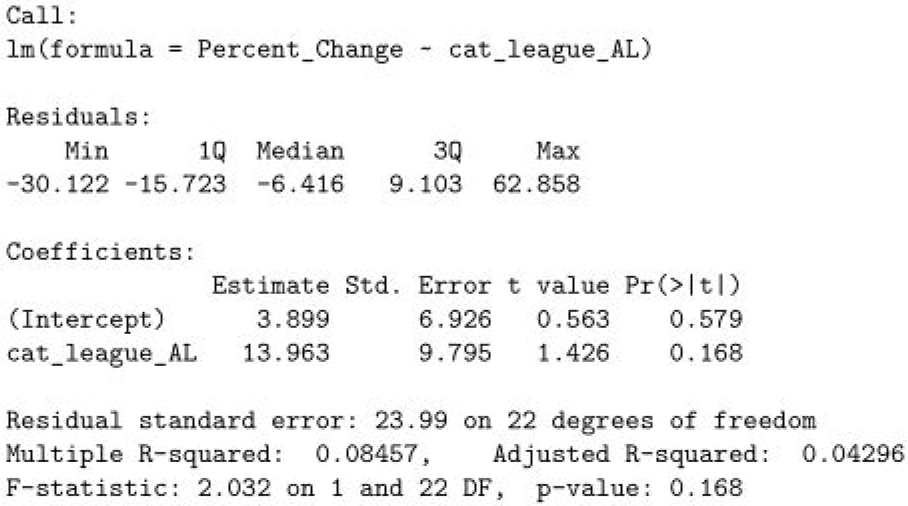
b. *Percent_Change* $= \beta_0 + \beta_1$ **cat_league_AL*
   i. This model estimates the percentage change in attendance between 1972 and 1973 based on whether the team was in the NL or AL.
   ii. We use this model in order to determine the significance of the difference of attendance growth between NL and AL. This shows whether AL attendance growth was significantly higher than NL attendance growth after the implementation of the DH rule.

c. *Percent_Change_AL* $= \beta_0 + \beta_1$ **DH_Home_Runs*
   i. This model estimates the percentage change in attendance for AL teams based on the number of home runs by the team's designated hitter.
   ii. We use this model in order to determine whether the attendance growth in the AL was significantly correlated with the number of home runs by designated hitters. This shows whether the treatment (DH rule, measured as home runs by DH's) had a significant relationship with attendance growth.

d. We chose to do a simple linear model because our project is focused around a natural experiment and we were curious about the causal relationship.

Results

1. Summary Table For Percent Change in Attendance Between Leagues



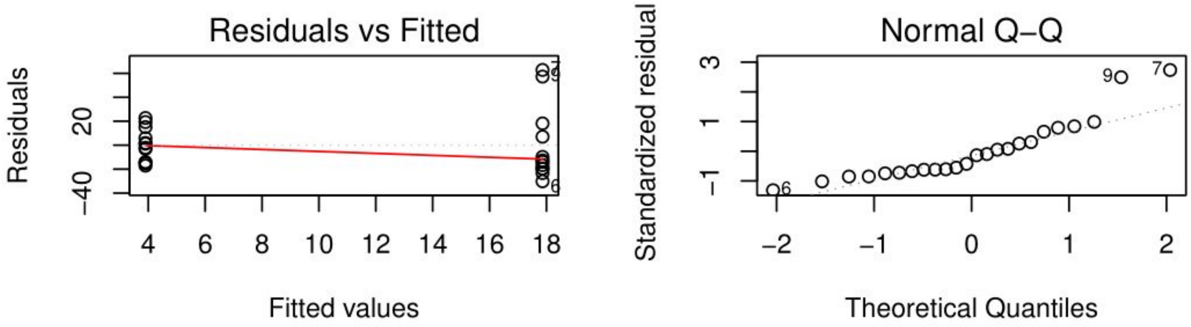
```
Call:
lm(formula = Percent_Change ~ cat_league_AL)

Residuals:
    Min      1Q  Median      3Q     Max 
-30.122 -15.723  -6.416   9.103  62.858 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.899      6.926   0.563    0.579
cat_league_AL   13.963      9.795   1.426    0.168

Residual standard error: 23.99 on 22 degrees of freedom
Multiple R-squared:  0.08457,    Adjusted R-squared:  0.04296 
F-statistic: 2.032 on 1 and 22 DF,  p-value: 0.168
```

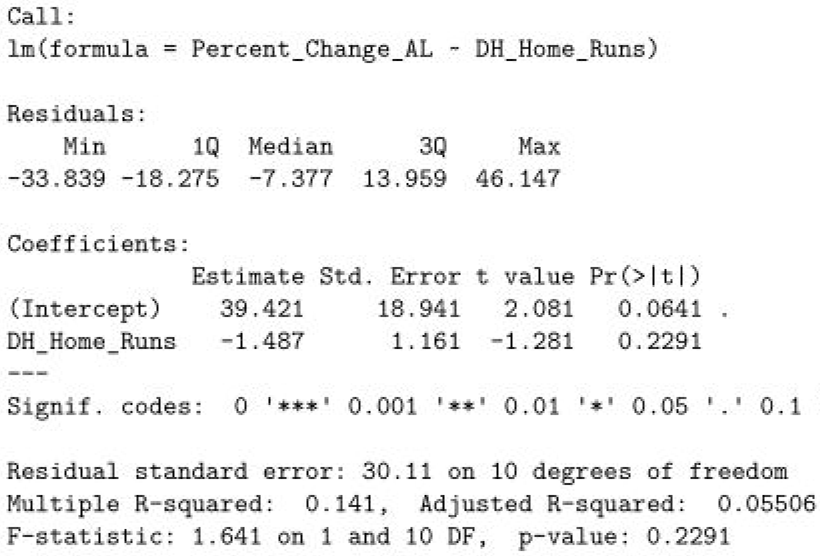
In the summary table, we analyze the *cat_league_AL* on *Percent_Change*. If the team is in the American League, there will be a 17.859% point increase in the percent change in attendance. If the team is in the National League, there will be a 3.899% point increase in the percent change in attendance. However, our p-value is insignificant for *cat_league_AL* at 0.168; thus, *cat_league_AL* does not show a significant correlation with *Percent_Change*.



In the Residuals vs Fitted plot we see the residuals for the individual league from their mean value. On the left we see the residuals for the NL centered around the β_0 value of 3.899

and the AL centered around β_0+ β_1 value of 17.859 but with a much larger variation and spread of the residuals.

2. Summary Table For Percent Change in Attendance for AL and Homeruns Hit by DH



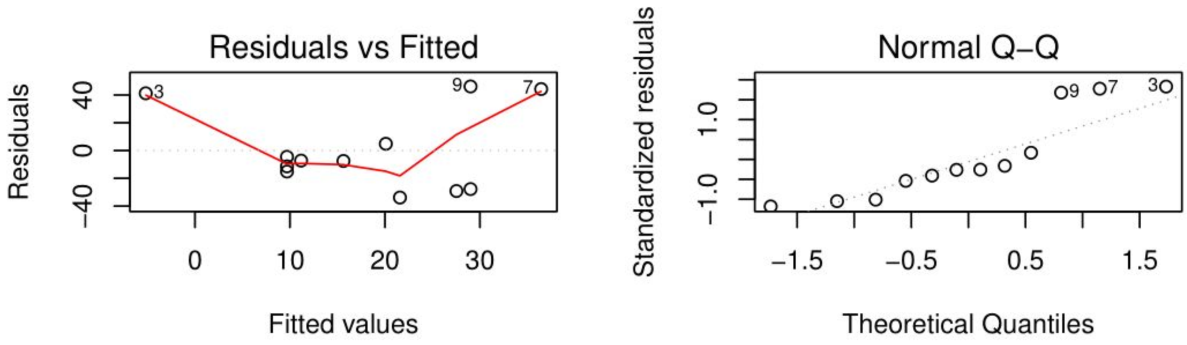
```
Call:
lm(formula = Percent_Change_AL ~ DH_Home_Runs)

Residuals:
    Min      1Q  Median      3Q     Max 
-33.839 -18.275  -7.377  13.959  46.147 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)  
(Intercept)    39.421     18.941   2.081   0.0641 .
DH_Home_Runs   -1.487      1.161  -1.281   0.2291  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 30.11 on 10 degrees of freedom
Multiple R-squared:  0.141,  Adjusted R-squared:  0.05506 
F-statistic: 1.641 on 1 and 10 DF,  p-value: 0.2291
```

In the summary table, we analyze the *DH_Home_Runs* on *Percent_Change_AL*. If the AL team's designated hitter hits one more home run, there will be a -1.487% point decrease in the percent change in attendance for AL teams. However, our p-value is insignificant for *cat_league_AL* at 0.2291; thus, *DH_Home_Runs* does not show a significant correlation with *Percent_Change_AL*.



In our Residuals vs Fitted plot you can see that the relationship is not shown to be linear but that is most likely due to the small sample size of our data. Most of the data is fairly linear except for three values pulling the data up which is due to our three cities having the largest percent change in attendance. The three cities and their percentages are: Kansas city (80.72%), Milwaukee (75.16%), and California (36.05%). Our Normal Q-Q plot shows a fairly normal distribution of our residuals.

Conclusion

Despite the average percent change in attendance being much higher for AL teams than NL teams, the fact that the league variable was not significant in the first regression prevents the conclusion from this analysis that the difference between AL and NL was statistically significant. This may be partly due to the very high standard deviations of percentage change distributions within the AL and NL. In fact, the percentage changes range from -12.26% to 80.72% for the AL, and from -13.18% to 26.64% for the NL. Overall, we cannot conclude that the implementation of the DH rule resulted in greater attendance growth for the AL than NL.

This analysis does not show that the number of home runs by designated hitters was correlated with attendance growth in the AL. So, we cannot conclude that more effective designated hitters corresponded to greater attendance growth in the AL.

One major limitation of this analysis is that there are numerous factors that affect attendance growth that cannot be accurately modeled or studied with the available data. These include but are not limited to day of week distribution for games, promotions for games, weather, and stadium size changes. Perhaps a model with more predictor variables would yield different results. Additionally, there are potentially many more changes between the 1972 and 1973 seasons besides the DH rule implementation that may have affected attendance growth. There was also a limited data set to use, since each team can only account for one observation per season in the model. Perhaps the results would show clearer trends if there were more teams to observe. Another potential issue is that home runs are not the only metric for offense, runs or hits by the designated hitter could be significant predictors for attendance growth. Although wage can be a confounding variable for home runs and the attendance growth of the franchise's games, there is no wage data during the period of the analysis. This is due to the antitrust exemption from the MLB. The antitrust exemption allows the MLB teams to have monopsony power over their players; thus, they are able to control how much they pay their players, where their players are traded to, and more. This gives zero power to the players or other franchises on getting better players (reserve clause). Players and teams were finally able to trade or sign better players once free agency opened up in 1975 after the Kurt Flood case was taken to Supreme Court. Overall, there are many potential confounding variables for attendance growth that if accounted for may create a more significant analysis.