Lab 12: Baseball

Multiple Regression

"Baseball is ninety percent mental. The other half is physical." - Yogi Berra

Part 1

1. What are some questions we can answer about baseball (or other sports) using statistics?
2. How do you think publicly available baseball data is collected? What sources of error could be associated with different collection methods?

There are a number of publicly available baseball datasets. One main source of data that we will be using this week is the Lahman database which contains a number of data sets with different units of observation. Below is the first few rows and some of the columns for two of these data sets: Teams and Batting.

yearID	teamID	franchID	G	W	L	R	RA	name
2020	ARI	ARI	60	25	35	269	295	Arizona Diamondbacks
2020	ATL	ATL	60	35	25	348	288	Atlanta Braves
2020	BAL	BAL	60	25	35	274	294	Baltimore Orioles
2020	BOS	BOS	60	24	36	292	351	Boston Red Sox
2020	CHA	CHW	60	35	25	306	246	Chicago White Sox
2020	CHN	CHC	60	34	26	265	240	Chicago Cubs

playerID	yearID	teamID	G	AB	R	H	BB	SO
abreual01	2020	NYA	2	0	0	0	0	0
abreubr01	2020	HOU	4	0	0	0	0	0
abreujo02	2020	CHA	60	240	43	76	18	59
acunaro01	2020	ATL	46	160	46	40	38	60
adamewi01	2020	TBA	54	185	29	48	20	74
adamja01	2020	CHN	13	0	0	0	0	0

3. What is the unit of observation for the Teams data set? What about for the Batting data set?
4. What is a question you could answer using the Teams dataset but not the Batting data set, and vice versa?
5. What is a question that we would need more granular (measured on a finer/more specific part of the game) data than the Teams and Batting dataset provide to answer?
6. Roughly since 1962 MLB teams have played 162 games in a season. What do you think the distribution of wins looks like? Sketch a plot and describe.
7. What do you think the relationship between wins and runs looks like? Sketch a plot and describe.
8. Some people believe analytics is ruining baseball because teams are more cautious which makes the games less entertaining. Do you agree or disagree? Why?

Part 2

Use the following code to load in the Teams dataset from the Lahman database.

```
library(Lahman)
data(Teams)
```

9. Subset the Teams dataset to only include years from 2000 to present day. What are the dimensions of this filtered dataset?
10. Plot the distribution of wins. Describe the relationship.
11. Plot the relationship between runs and wins. Describe the relationship (form, direction, strength, presence of outliers).
12. Plot the relationship between runs allowed and wins. Describe the relationship. How does it compare to the relationship between runs and wins?
13. Fit a simple linear model to predict wins by runs. Write out the equation for the linear model and interpret the slope. What are the R^2 and adjusted-R^2 values?
14. What is the average number of season runs and wins? Based on the previous model, how many games would you predict a team that scored the average number of runs would win? What about a team that scored 600 runs? What about 850 runs?
15. Fit a simple linear regression model to predict wins by runs and runs allowed. Write out the equation for the linear model and interpret the slopes. What are the R^2 and adjusted-R^2 values? Compare to the simple linear regression from the previous question.
16. Fit a multiple regression model to predict wins using at least two other variables in this data set. How does the R^2 and adjusted-R^2 values change? Do you think the new model you created predicts wins better?
17. Does this seem like a reasonable setting to conduct statistical inference / make generalizations? To which population might a sabermetrician wish to generalize?
18. Is it reasonable to draw a causal conclusion from these models? Why or why not or under which circumstances?