1 Instructions

1a, 1d, 2a, and 2c will require you to perform a simple regression analysis. A regression analysis should include the following information. Other parts will not require all of the following information — answer the question as written.

- Non-technical summary of analysis
- Hypothesized Model: State your hypothesized model as $\hat{y} = \beta_0 + \beta_1 x$ with numerical values for β_0 and β_1.
- A scatterplot showing your hypothesized model overlaying the original data.
- Parameter estimates and confidence intervals where requested (unless otherwise noted, use the t-distribution for tests/intervals)
- Coefficient of determination
- A test for validity in your model (either slope or coefficient of correlation). Remember all parts of a hypothesis test.
  - Hypotheses
  - Test Statistic
  - Either p-value or rejection region
  - Technical conclusion
- Predictions and estimates (where requested) for your model.
- For now, you do not need to validate assumptions. We'll do more of that next week.

2 Questions

Problem 1: 50 points
Using the `hofbatting.csv` file that you've come to know and love, conduct a regression analysis to determine if On-Base Percentage (OBP) can be used to predict Slugging Percentage (SLG).

(a)
Test whether this model is providing useful information. In other words, is the slope non-zero? (If not, the model $y = \bar{x}$ could not be dismissed.)

(b)
What is the expected slugging percentage for players with an OBP of 0.32? Give a confidence interval for $\alpha = 0.05$.

(c)
What would you expect the slugging percentage of a new inductee with an OBP of 0.32 to be? Give a confidence interval for $\alpha = 0.05$.

(d)
We previously identified Willard Brown as an outlier. Redo the analysis from the previous problems excluding his data. Did it make a difference?

Problem 2: 50 points
In baseball, it is hypothesized that we can use the run differential to predict the number of wins a team will have by the end of the season. Use the file `Teamdata.csv` to test this concept.

(a)
Create a column of data for Run Differential (runs - runs allowed or $R - RA$) and a column for Win Percentage ($W/(W + L)$). Use these values to determine if the Run Differential can be used to predict the percentage of wins a team will end up with.

(b)
Create a plot showing the confidence intervals for mean estimation and prediction overlaid on the original scatterplot. Reduce the dot size so that you can see the lines/zones.

(c)
Bill James, the godfather of sabermetrics, empirically derived a non-linear formula to estimate winning percentage called the Pythagorean Expectation.

$$W_{pct} = \frac{R^2}{R^2 + RA^2}$$

Create a new variable representing the Pythagorean Expectation. Now use this new column to replace the Run Differential and re-run your analysis.

(d)
The 2001 Seattle Mariners had 116 wins and 46 losses with a +300 Run Differential in the data. Find this row in your data and pretend it's a new team with identical stats.

Use both the Run Differential and Pythagorean Expectration models to create confidence intervals for the Win Percentage expected for another team with identical performance to the Mariners.

(e)
What are the pros and cons of each of these models (PE and RD)? Which would you rather use? Why? Think about practical usage as well as statistical accuracy.