Final Project — MATH 185 (Winter 2019)

Salaries of Baseball Players

The data in the file baseball.txt are a subset of data collected on professional American baseball players. Focus on the following variables:

- salary — The yearly salary of each player (there are 59 NA's in salary — remove these rows)
- homeruns_career — The total number of homeruns a player has made over the course of his career
- division — The division a player is in (either W or E)

Part I: Rank sum test

(a) Are the salaries similar in the two divisions? Examine whether the distributions of player salaries are the same in both leagues. Use the Wilcoxon test with the normal approximation and calculate the value of the test statistic, its expected value and variance under the null hypothesis of no difference, and give the test result.

(b) The data contain ties. Explain how to simulate the distribution of the test statistic under the null. Implement this procedure and check that the significance levels you computed in the previous question are robust to the effect of ties.

(c) Test your result by implementing the Wilcoxon test in R and give a point estimate (using the Lehmann-Hodges estimator) and a confidence interval for the difference in median salary between the two leagues.

Part II: Rank sign test

(d) Consider the possibility that walks and runs have the same median. Find a confidence interval for the difference in medians and use it to test for equal medians at 95%. What do you conclude?

Part III: Smoothing

(e) Examine the relationship between the log of outcome variable salary and explanatory variables homeruns_career and division using exploratory plots.

(f) Can we predict salaries with the predictor variable homeruns_career? Perform a regression of response variable log_salary on the variable homeruns_career. Plot the regression lines onto the data and comment on the fit. Give an interpretation of the model, keeping in mind that the salaries are reported on a log-scale. Comment on the effect of extreme observations in homeruns_career.

(g) Try smoothing procedure, for example a Kernel smoother ksmooth() or local polynomial regression locpoly(). Plot the fit for various smoothing parameters and discuss what you find.

(h) Use LOO-CV to estimate the bandwidth in a local-polynomial fit of degree 1 (see lecture code for something nearly identical). Calculate the smoothing matrix S in this case (the expression for S was given in lecture 9 for local linear).