Question 1: Salaries of Baseball Players

The data in the file http://www.stats.ox.ac.uk/~nicholls/CompStats/baseball.txt are a subset of data collected on professional American baseball players. Focus on the following variables:

1. **salary** - The yearly salary of each player (there are 59 NA’s in salary - remove these rows)
2. **homeruns_career** - The total number of homeruns a player made over their career
3. **division** - The division a player is in (either W or E)

(a) EDA and Smoothing

a.1 Examine the relationship between the log of outcome variable **salary** and explanatory variables **homeruns_career** and **division** using exploratory plots.

a.2 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**.

a.3 Try a 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. Investigate also `lowess()` and `dpill()`.

a.4 Could anything be done to improve the fit?

a.5 Optional Extensions: (i) use LOO-CV to estimate the bandwidth in a local-polynomial fit of degree 1 (see lecture code L6.R for something nearly identical); (ii) Calculate the smoothing matrix $S$ in this case (the expression for $S$ is given below and in PS2) and verify the fit by hand. Calculate the DOF for GCV, $\text{trace}(S)$, and the DOF for variance estimation, $2\text{trace}(S) - \text{trace}(S^T S)$.

(b) Rank sum test

b.1 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 by hand the value of the test statistic, its expected value and variance under the null hypothesis of no difference and give the test result.
b.2 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.

b.3 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. State any assumptions.

(c) Rank sign test

c.1 Consider the possibility that walks and runs have the same median. Why should we treat these data as paired?

c.2 Find a confidence interval for the difference in medians and use it to test for equal medians at 95%. What do you conclude? State any assumptions.

Question 2. What we test depends on what we can assume.

(a) Suppose we have two samples \(X_1, \ldots, X_n\) and \(Y_1, \ldots, Y_m\) and we suspect they have different variances but the same mean and median. For example

\[
n=100; \ x=\text{rnorm}(n,0,1); \\
m=100; \ y=\text{rnorm}(m,0,1.5); \\
\text{wilcox.test}(x,y) \ #accepts \text{H0}
\]

The Wilcoxon Rank Sum test isn’t designed to detect this sort of difference. Invent your own test statistic and use a permutation test to test for a difference, \(F_X \neq F_Y\). I used

\[
z=c(x,y) \\
T=\text{abs}(\text{sd}(z[1:n])-\text{sd}(z[(n+1):(m+n)])) \ #my \text{ test stat is sum of } |\text{sd}(x)-\text{sd}(y)|
\]

Explain why my choice is not robust to outliers. Can you make it robust?

(b) Suppose we have two distributions with the same median but completely different shapes (for example they might be skewed in different directions…hint). Can the Wilcoxon Rank Sum test detect that the two distributions are different? The answer is yes, for at least some cases. Construct an example.

Formulae for question 1(a.5) For data \((x_i, y_i), i = 1, 2, \ldots, n\) and, for \(x \in \mathbb{R}\), if \(\hat{m}(x)\) is a local linear smoother with kernel \(K(x)\) and weight function \(w_i(x) = K((x - x_i)/h)\), then

\[
\hat{m}(x) = \sum_{j=1}^{n} \frac{\hat{w}_j(x)}{\sum_{k=1}^{n} \hat{w}_k(x)} y_j
\]

where

\[
\hat{w}_j(x) = w_j(x)(T_{n,2}(x) - (x_j - x)T_{n,1}(x))
\]

and

\[
T_{n,1} = \sum_{k=1}^{n} w_k(x)(x_k - x), \quad T_{n,2} = \sum_{k=1}^{n} w_k(x)(x_k - x)^2
\]