BEYOND THE NUMBERS 3.6

Outliers and Leverage Points

Name: Matthew Massey Section Number: ________

To be graded, all assignments must be completed and submitted on the original book page.

EXHIBIT 1

Heptathletes

Finish data for two 1992 Olympic Heptathlon events are shown below. A scatterplot of the data are shown just to the left of the table. Chouaa is the green data point and Barber is the red one.

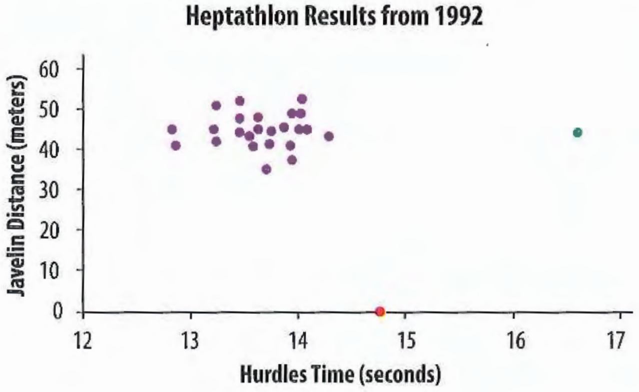

TABLE 3.6 Heptathletes

Name	Hurdles (seconds)	Javelin (meters)
Joyner-Kersee	12.85	44.98
Nastase	12.86	41.3
Dimitrova	13.23	44.48
Belova	13.25	41.9
Braun	13.25	51.12
Beer	13.48	48.1
Court	13.48	52.12
Kamrowska	13.48	44.12
Wlodarczyk	13.57	43.46
Greiner	13.59	40.78
Kaljurand	13.64	47.42
Zhu	13.64	45.12
Skjaeveland	13.73	35.42
Lesage	13.75	41.28
Nazaroviene	13.75	44.42
Aro	13.87	45.42
Marxer	13.94	41.08
Rattya	13.96	49.02
Carter	13.97	37.58
Atroshchenko	14.03	45.18
Vaidianu	14.04	49
Teppe	14.06	52.58
Clarius	14.1	45.14
Bond-Mills	14.31	43.3
Barber	14.79	0
Chouaa	16.62	44.4

Questions

1. What kind of association do you see in the scatterplot—positive, negative, neither? Support your answer.

The scatterplot shows NO LINEAR RELATIONSHIP. As X increases, Y generally stays the same, and the correlation r is close to 0.

2. Compute the correlation coefficient "*r*" for the entire data set. You should use a software package or an online applet as required by your instructor. Is this value of "*r*" consistent with what you answered in Question 1? Why or why not?

StatKey gives a correlation of -0.252

This is NOT consistent with my answer in question 1. I answered correlation ~= 0 because I visually estimated a non-linear relationship, however, StatKey correlation is affected by the outlier Barber (14.79, 0), which pulls the correlation down to a slightly negative linear relationship.

EXHIBIT 2

Language

In a scatterplot, outliers are data pairs that are not spatially close to the bulk of the data. Outliers are not necessarily a problem for the human inferences that arise from a correlation coefficient. However, if the removal of a single outlier causes a distinct change in the correlation, then that outlier is called an influence point. Influence points can disguise the essence of an association.

Questions

1. Looking at the scatterplot above, which athletes are outliers?

There appears to be two outliers Barber (14.79,0) and Chouaa (16.62, 44.4)

2. Compute the correlation coefficient "*r*" for the data set with Barber removed. Is Barber an influence point? Why?

With Barber removed, the StatKey correlation is 0.0036...close to my original visual estimation. Barber is definitely an influence point because it distinctly brings the correlation into a negative linear relationship of -0.252

3. Compare the value of "*r*" that you computed for the entire data set to the value of "*r*" that you computed with Barber removed. Which one best reflects the association seen in the scatterplot? Why?

This has basically been answered in the above questions already, but...

With Barber, correlation = -0.252. Without Barber, correlation = 0.0036. The correlation without Barber of 0.0036 best reflects the association in the scatterplot because the bulk of the dataset do NOT show any linear relationship