Scottish Hills

Mitchell Pudil

February 2, 2019

1 Introduction

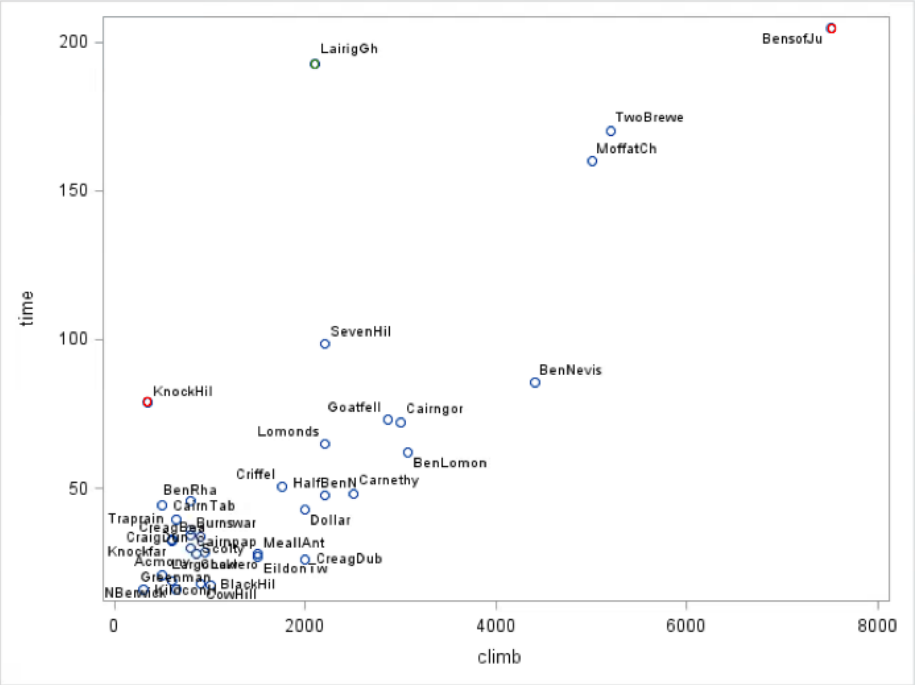
Scottish hill racing, also known as fell running, is a sport of running and racing off road, over upland country where the gradient climbed is a significant component of the difficulty. The name arises from the origins of the English sport on the fells of northern Britain, especially those in the Lake District. It has elements of trail running, cross country and mountain running, but is also distinct from those disciplines. I was asked to perform diagnostic tests to assess the validity and characteristics of a dataset containing information of various races.

2 Data

The hill racing data being used was gathered by Atkinson for a popular regression diagnostics paper and is available at http://www.statsci.org/data/general/hills.txt. It contains the record time for 35 hill races, as well as the corresponding distance and hill climb.

3 Exploratory Data Analysis

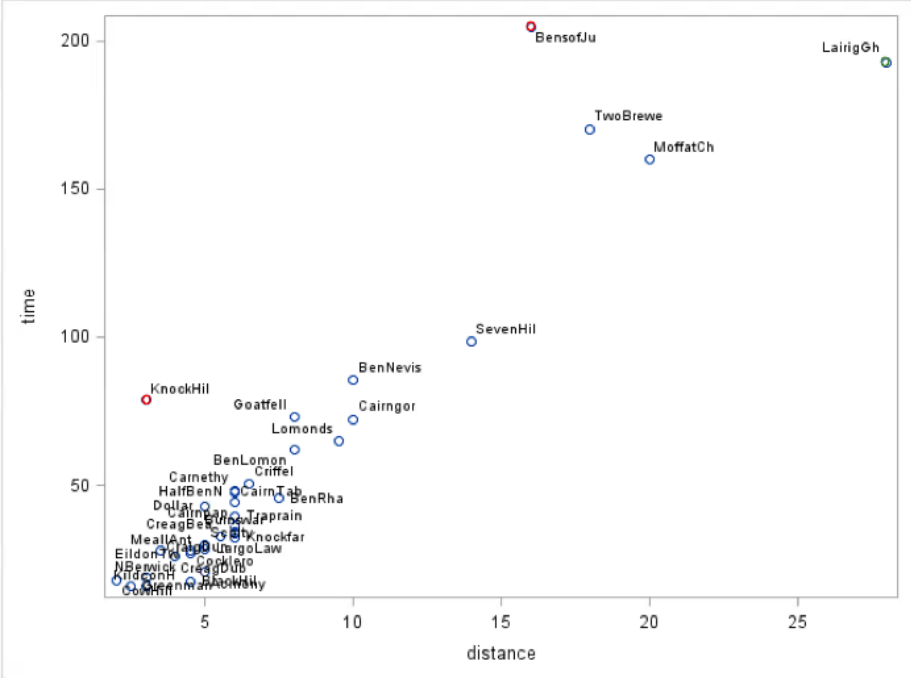
To begin, let's consider the correlations between climb and time, as well as the between distance and time. A scatterplot for climb and time is shown below:



The red circles indicate a race that, for the regression of time on distance and climb, was marked as a bad influential observations. That means that those observations carry a lot more weight in determining our parameter estimates than just one observation would, and that they do not follow the general trend we would expect. Thus, it is an outlier as well as an influential observation. However, not all influential observations are bad. The race circled in green is a good influential observation, which means that while it did carry more weight in determining our parameter estimates, it is not an outlier. This is generally the case when we have a data point that is far away from others but that follows

the same trend.

The scatterplot of distance vs. time is shown next:



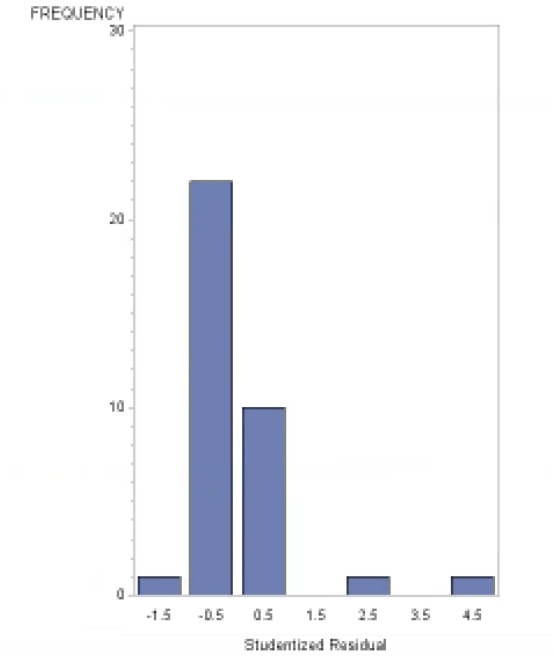
We can see from the plots that the bad influential observations come from the Bens of Jura race and from the Knock Hill race. However, the good influential observation in the dataset is the Lairig Ghru race.

4 Analysis

For the analysis, I was asked to determine various statistics to further analyze the influence of certain observations and the characteristics and validity of the dataset as a whole.

First, I was asked to compute the R-studentized residual for the Cairn Table race. I found the value of the residual to be 0.72. Next, I was asked to create

the histogram of the R-studentized residuals and determine if there is evidence of violating the normality assumption through both the Shapiro-Wilk normality test and the histogram itself. The histogram of the R-studentized residuals are shown below:



With a p-value <0.0001 from the Shapiro-Wilk normality test, it appears that the residuals are not normally distributed. This is important to note since if one were to do a regression analysis as is, the standard errors would be wrong, and so too would be any inference.

I was informed that Kildcon Hill has a record time of 15.95 minutes, but the regression model predicts 12.9762 minutes, and then asked if this observation would be an outlier. I determined that it wasn't an outlier since the R-studentized residual for this race is 0.209, which is much less than the cutoff for outliers.

I was also told that there is some concern that the record time for the Knock Hill race was recorded correctly and asked to use the appropriate diagnostic statistic(s) to investigate and conclude whether or not this concern is based in rumor or data. I found that the studentized residual for Knock Hill is 4.56, so it is an outlier. Because of this, there is evidence to suggest that the record time for the Knock Hill race could have been incorrectly recorded.

Next, I was asked to compute the leverage for the Ben Nevis race. I determined the leverage to be approximately 0.1216, which suggests that the Ben Nevis race is not a very influential observation.

Additionally, I was asked to compute the Cook's Distance for the Moffat Chase race. I found that distance to be only 0.052. so the Moffat Chase race is not an influential observation.

I was instructed to see if the races at Moffat Chase hill and Lairig Ghru hill are an influential observation. I discovered that the Moffat Chase hill race to not be influential since its Cook's Distance is very small. However, the race at Lairig Ghru hill is influential because its Cook's Distance is since its Cook's Distance is 0.211. The Lairig Ghru hill is a good influential observations since there are few observations similar to it, but it follows the general trend we would expect.

5 Conclusion

We now know that there are a few influential observations in this dataset that we need to be careful about. Particularly, there are two bad influential observations that act as outliers that should probably be tossed out if further analysis on the data is to be done. However, there is a good influential observation that allows us to have better predictions in areas where we do not have much data.