Scottish Hill Race Predictions

Scottish Hill races are popular races where each individual race varies in both elevation and distance that challenge athletes. It is like cross country running, but the runner in a hill race many times does not have a pre-paved route to follow. Times are recorded and allow people to compete against one another and participate in professionally organized races. It also provides a great opportunity to create a model and estimate winning times in different hill races.

Getting Started

To start, pull data from http://www.statsci.org/data/general/hills. This tab-delimited dataset is the record time (Time) as of 1984 for 35 races and was collected by Atkinson for a popular regression diagnostic paper. There are some concerns about the integrity of the data, so part of our analysis will look for influential points and outliers.

Prerequisites

- A system running R or SAS (analysis will be done in both)
- General knowledge of regression models and the PROC REG statement in SAS

EDA

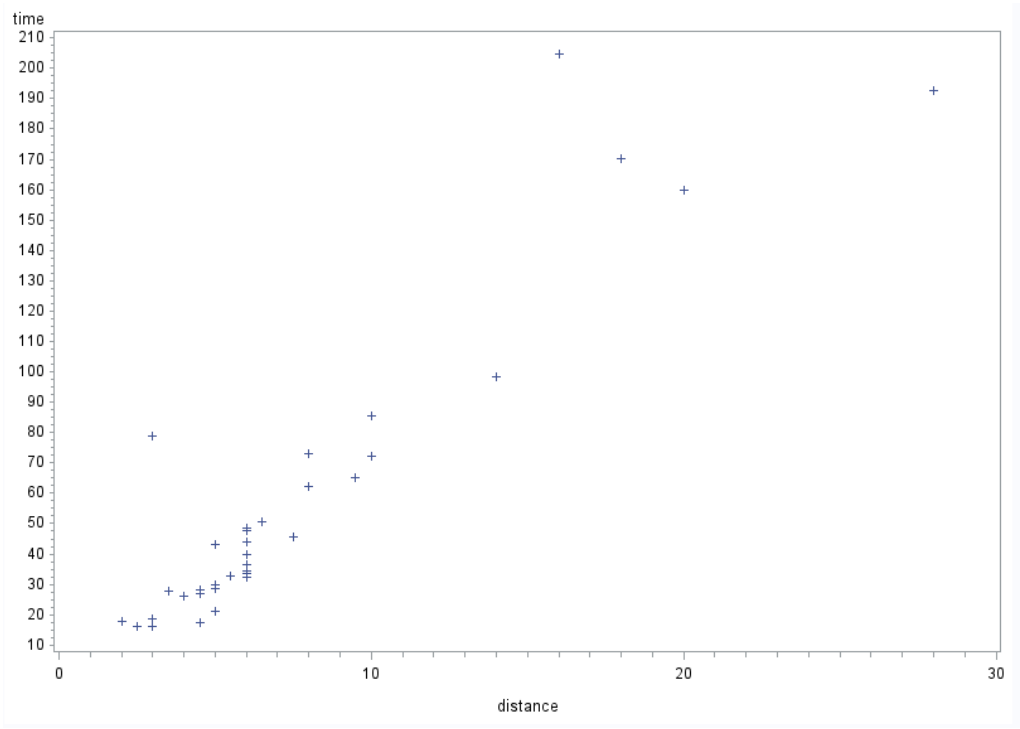
We want to first create several scatterplots, with *Time* on the x-axis and *Distance* or *Climb* on the y-axis. This provides us with an initial impression of the data, and allows us to visualize what we should expect and see any points we may need to take out. In SAS, we do so with two simple PROC GPLOT statements.

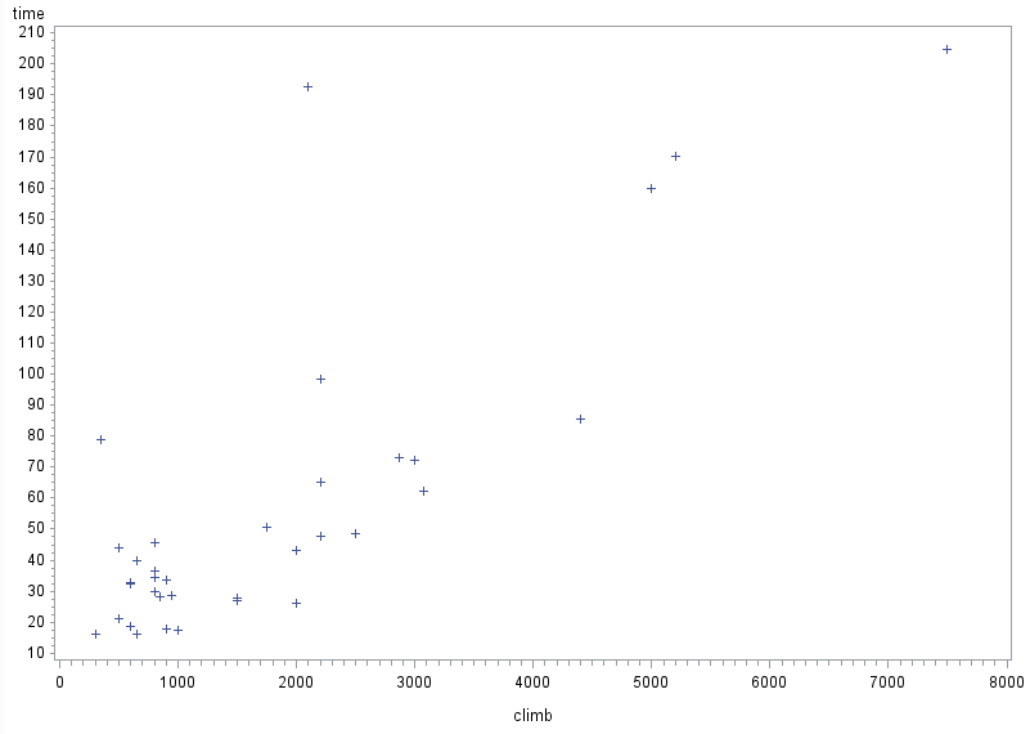
```
proc gplot data=hillraces;                /* Scatterplot for time and distance */
        plot time*distance;
run;

proc gplot data=hillraces;                /* Scatterplot for time and climb */
        plot time*climb;
run;
```

Here is the same process in R

```
plot(hills$Time~hills$Distance)           #Scatterplot for time and distance
plot(hills$Time~hills$Climb)              #Scatterplot for time and climb
```

The following graphs are produced:





We can see in both graphs that there is a general trend to the data, but a couple of points seem like outliers and we want to analyze those further.

Expanded EDA

We start by building our model. Our model is as follows:

$$\text{Time} = \beta_0 + \beta_1 * Distance + \beta_2 * Climb + \varepsilon,\ \varepsilon \sim N(0, \sigma^2)$$

In SAS, this would use a PROC REG statement

```
proc reg data=hillraces;
        model time = distance climb / r influence;
run;
```

and in R:

```
out.hill <- lm(Time~Distance + Climb, data=hills
```

These are the reported parameter estimates and standard error from the model.

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-8.99204	4.30273	-2.09	0.0447	0
distance	1	6.21796	0.60115	10.34	<.0001	1.74081
climb	1	0.01105	0.00205	5.39	<.0001	1.74081

This tells us that $\hat{\beta}_1$ = 6.21796, so for a one km increase in distance, time is predicted to increase by 6.21796 minutes, which from a hill running standpoint seems logical. Likewise, $\hat{\beta}_2$ = 0.01105, so for a one unit increase in elevation, time should increase 0.01105 minutes.

Calculating R-studentized residuals & Cook's Distance

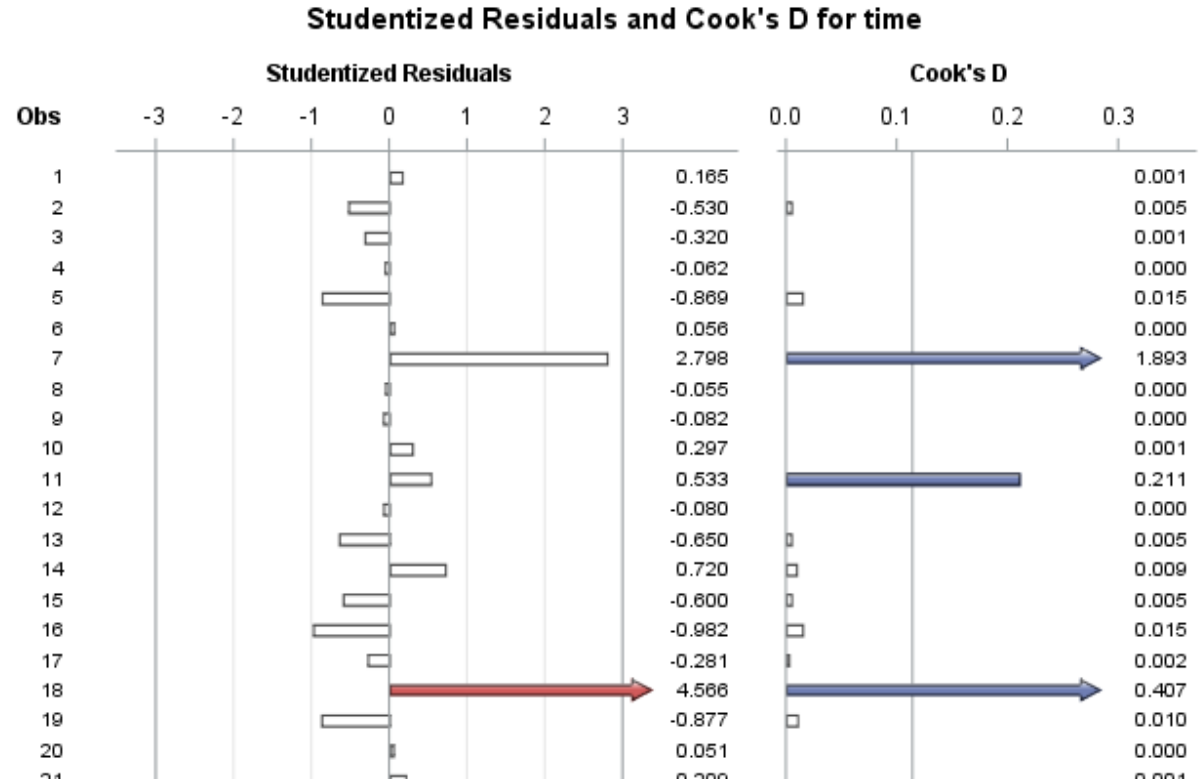
How we want to identify influential observations and outliers is to calculate the R-studentized residuals and Cook's Distance for each point. To do this in SAS, within the PROC REG statement, you need to incorporate after the model statement:

```
/ r influence;
```

Which tells SAS to calculate both R-Studentized residuals and Cook's Distance. In R, that looks like this:

```
cooks.distance(out.hills)        #Cooks distance from our model called out.hills
rstudent(out.hills)              #R-Studentized Residuals
```

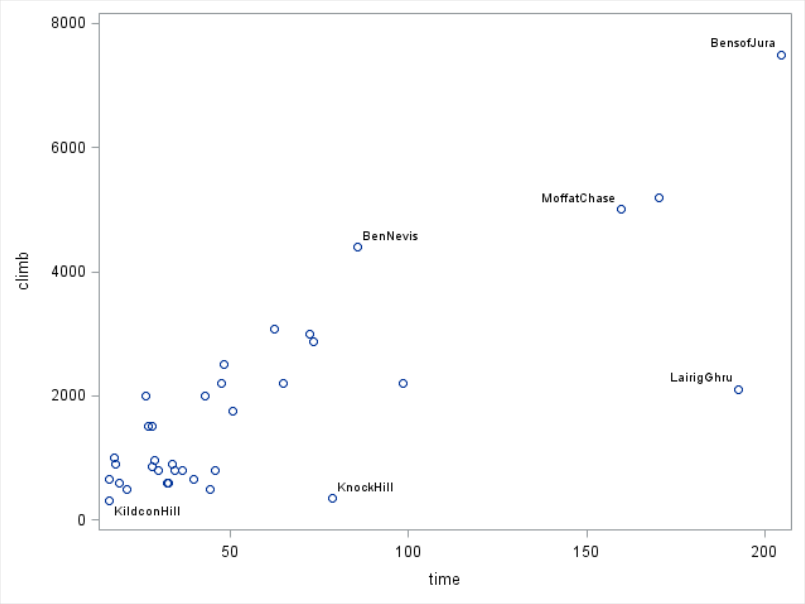
SAS provides a great graphic showing the R-Studentized residuals and Cook's distance, and flags influential observations



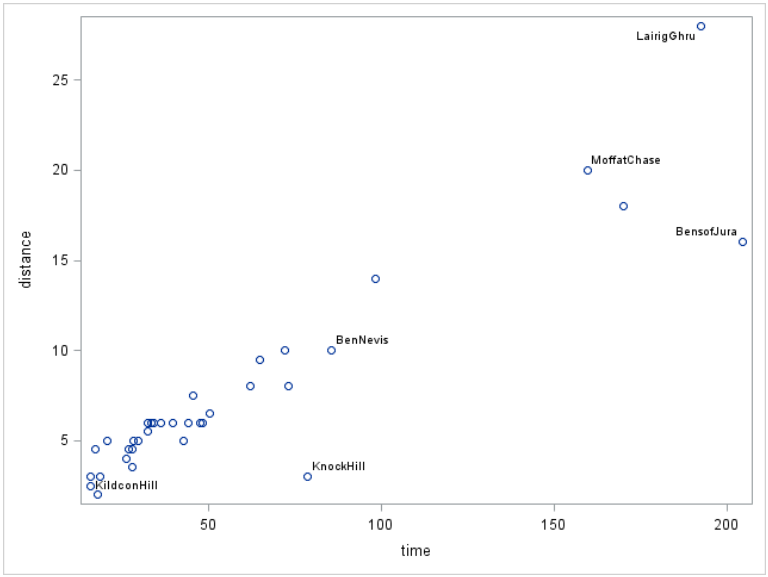
SAS flagged obs 18 in both R-Studentized residual and Cook's distance, while obs 7 and 11 were flagged by Cook's distance, so those we want to look at those.

Obs 7 is Bens of Jura, obs 11 is Lairig Ghru, and obs 18 is Knock Hill. To take a closer look, I plotted the same time vs. distance/climb charts as before, but labeled a few points to see what points those were and decide if they were points I would want to keep or remove from my data to improve my model.

Here is the time vs. distance chart:



In this chart, Bens of Jura is at the top-right in the chart and is extremely far from all the other points. It is influential because of that, but it is a good influence. Lairig Ghru doesn't follow the general data trend, and would pull my prediction down, thus being a bad, influential observation. Knock Hill is similar to Lairig Ghru, and I may consider taking it out.



In this time vs. distance chart, Bens of Jura does seem to drift from a line of best fit through the rest of the data, making it a potentially bad influence. Knock Hill is again a bad influence, making it a candidate for being removed. Lairig Ghru seems like a perfect, good influencer to the model in this case.

As a result, I would decide to take out Knock Hill from this data set.

Final Notes

We have seen that both distance and climb are statistically significant with p-values < 0.001, and that some of the data may positively or negatively affect our data. As a future project, we could continue our analysis so we can make predictions, or we could analyze more recent data and incorporate more hill races than just the 35 we analyzed.