The problem:

My goal for this project was to model a biathlon race and to use the resulting model to better understand the ways in which race conditions impact performance. I then wanted to use my model to predict likely outcomes for upcoming races. I begin by briefly discussing the sport itself.

The biathlon is a sport which combines two disciplines, cross-country skiing and target shooting. A race consists of either three or five skied laps which alternate with episodes of target shooting, each of five shots. Each missed shot incurs a penalty, in most cases, an additional 150 meter lap, which adds around 25 seconds to the total race time. Final placement is determined by total time to complete the race.

In this project, I have chosen to focus on the men's sprint race for two reasons. The first is the frequency with which it is held. The second is the relatively large number of athletes who compete each time. This race consists of three skied laps of around 3300 meters alternating with two rounds of five shots each, the first prone and the second standing. This race is held in a time trial format, which reduces the impact of direct competition between racers and in turn makes it easier to model.

One of the challenges that I faced in trying to model this race was the variability inherent in the structure of the biathlon. Because each missed shot increases total time for a racer by nearly a half minute, and because an athlete who averages 80% shooting accuracy (a not unreasonable estimate) has a roughly 33% chance of making all five shots, and a better than 25% chance of missing at least two. This alone introduces a difference of nearly a minute into the data. In addition, while I had available some information about some of the conditions under which the athletes were competing, the available information may be inadequate to explain the variability that is present. In particular, there may be attributes of individual racers which are both critical to obtaining a good model and impossible to know.

The data:

In this analysis, I used data from two different types of files, both found at the official race data site for the International Biathlon Union (IBU), https://biathlonresults.com. The first type contains data about the performances of individual racers in a given race, such as ski time, number of missed shots, and shooting time. The second contains data about the conditions of the race as a whole, such as a course description and information about weather and snow conditions during the competition. My goal was to use data about the race conditions to try to predict racer performance for a given race.

The attack:

In order to better model the race, I began by dividing it into separate pieces.

(1) The first piece of my model is the ski time. It is the most significant part of a racer's total time, but cannot be modeled directly due to variations in the length of a sprint race. As a result, I converted the ski times into speeds, and modeled those.

(2) The second piece of my model the number of missed shots. Both prone and standing shooting can be modeled (separately) as random variables with the likelihood of success of any single shot determined by the shooting accuracy of the racer.

(3) The third piece of my model is the range time, which is the most complex part and of necessity depends on the second piece. To predict range time, we can use the same process for both prone and standing range. We must estimate both the time spent actually on the shooting range and the time required to ski a penalty lap. We then combine this with the predicted number of missed shots to arrive at an estimate for range time.

I started by assuming that the components of the total race time are dependent on one or more of the course attributes or weather conditions. To investigate this, I first modeled each component as a linear function of each predictor variable in turn. I then used the correlation coefficients and the amount of change predicted in the response variable to select variables that had a strong impact on the given component. For each of these, I found that there were two or three variables that fared significantly better by these measures than the others.

In order to have a second means of evaluation for the predictor variables, I wanted to see how much predictive power each of them might have. (Note that while our first means of evaluation considered variable performance on races in their entirety, this one considers performance on racers individually.) To do this, I began to build the pieces of my model. My key assumption was that if race conditions impacted speed and shooting accuracy, then races run under similar conditions would have more similar outcomes than races run under very different conditions[1]. I implemented this in the following manner.

To investigate the predictive effects of altitude[2] on speed, I took all prior speed performance data for a racer. I then weighted the races, giving a lot of weight to previous races for which the the altitude was similar to that of the race being modeled, and very little weight to previous races for which the altitude was very dissimilar to that of the race being modeled. These weighted lists are then used to predict performance.

Here, I considered two things in selecting the best variables. The first was how frequently speed distributions based on that variable had the racer's actual speed in the middle 50% of his distribution of predicted speeds. The second was to what extent the racers who were faster than predicted were balanced by others who were slower than predicted. Once again, I found that there were two or three variables that were markedly better than the others.

Finally, for each piece of the model, I took those variables that performed well both here and above. They were

- For speed: wind speed and snow conditions
- For prone shooting accuracy: weather
- For standing shooting accuracy: altitude and wind speed
- For prone range time: weather, snow conditions, and event (within a season)
- For standing range time: weather and snow conditions

Now I was ready to assemble the parts of my model in order to predict the final outcomes of the races. Before I tested my model (which depends on the choice of variables used

[1] For simplicity, in the remainder of this paragraph and in the two which follow, we will assume that we are discussing speed. The processes for the other components of the model are analogous.

[2] Again for simplicity, we assume that the predictor variable of interest is altitude.The process for the other predictor variables is the same.

in weighting), I wanted to assess which combination of predictor variables was best at making predictions[3] by running my full model for each combination of variables produced by combining successful variables from above.

My full model consisted of two parts:

(1) The most significant contributor to total race time is the speed of the racer, which determines the time spent on each of the three ski loops. In order to model a racer's speed for a given race, we first fix a predictor variable, V. We take all of that racer's speed in prior races and weight them according to how similar the values for V for those races are to the value of V for the given race. We then randomly select a sample of ten speeds from this weighted list of prior race speeds and find their mean to produce a speed estimate for the given race. We repeat the process of drawing samples and averaging them until we have produced the desired number of predicted speeds.

(2) To produce a distribution of predictions for either prone or standing range times, we again begin with a weighted list of prior race results, this time consisting of pairs of range times and missed shots. We then draw bootstrap samples from this list and fit a line through the sample data in order to estimate both the portion of the range time devoted to shooting and the time required to ski a penalty lap. After simulating five shots[4], we compute an estimate of the total range time as

$$\text{range time} = \text{shooting time} + (\text{missed shots} \times \text{penalty lap times})$$

We again repeat this process until we have produced the number of predicted range times that we want.

The distribution of predicted total times is produced by converting each speed to its corresponding ski time and adding it to the predicted prone and standing range times.

Once we had run trials of the full model on an assortment of races for each variable combination, it was necessary to decide which weight combination produced the best results. Determining this required some measures of quality. I devised two measures, and combined their results in order to determine which of the variable combinations yielded the best results. The first of these measures how well the model predicts the racers' order of finish. The second of them measures how closely the racers' actual times correlate with the centers of the distributions of predictions. These measures allowed me to isolate a single combination of weights that seemed to outperform all of the other combinations on average. It was

- For speed : snow conditions
- For prone shooting accuracy : weather
- For standing shooting accuracy : altitude
- For prone range time : snow conditions
- For standing range time : weather

[3] We do not assume that the same predictor variable will be best for each part of the model. Rather, we recognize that snow conditions might best predict speed, while event might best predict prone range time.

[4] using the racer's shooting percentage

Evaluation of the model:

I finally had a single model, but in order to evaluate it, I needed some basis of comparison. Since my model produced a distribution of predictions for each racer rather than a single value, I wanted to compare my outcomes to the outcomes produced by another such model. Investigation of the career performances of those biathletes who had competed in a large number of races suggested that a reasonable model for the distribution of their race times would be an exponentially modified normal curve. I used this to create a naive model as follows:

(1) For each racer, fit an exponentially modified normal distribution curve to his prior race times.

(2) Draw a random sample of n times from the fitted model to produce a predictive distribution for the current race.

I then used both my model and this naive model to produce predicted time distributions for every racer in each race in the most recent four seasons.

At this point, I added two additional measures to the two measures of quality that I used above. The first was to compare the average of the difference between the means of the distributions and the actual times for the race for the two models. This value should be as small as possible. Te second was to find what percentage of the times on average that the racer's actual times fell in the middle 50% of their prediction distributions and then to compare the values for the two models. This time our desired value is as large as possible.

Discussion of model performance

In general, my model outperformed the distribution model. It correctly (or nearly correctly) placed 30% more racers than did the distribution model, (and fared 63% better than random change for this measure). On average, my model had 80% more race times within a given margin of error, though it only outperformed the naive model about two thirds of the time. The aspect of the comparison that is the most worrying is that the percentage of racers whose times fell in the middle 50% of their distributions, however this can be readily explained by the fact that the standard deviation of the predictions produced by the naive model were more than double to size of the standard deviations produced by my model.

Although my model did fairly well in comparison with the naive model, it has some significant shortcomings. The first is that the model does not permit us to make very precise time predictions, but rather provides us with a range of values with varying degrees of probability. However, due to the randomness inherent in the shooting portion of the race, it seems as if it would be impossible to find a useful model that gave a single prediction for each athlete in an event. In addition, because my model depends on snow and weather conditions at race time, which are unlikely to be known ahead of time, we are limited in our ability to use it to predict outcomes of races that have not yet occurred.