Studying Nationalistic Bias in Figure Skating Judging

Kyle Wang

Formalizing the Question:

We are interested in whether judges exhibit nationalistic bias—specifically, whether a judge assigns systematically higher scores to competitors from the same country. We expect this nationalistic bias to manifest in the scores that a judge gives. Because of this assumption, we must be very careful how we formulate the hypothesis we wish to test so as to avoid any sort of post-treatment conditioning.

In the example of the Olympic Winter Games 2022, the nine judges for the short program are given and seemingly, each judge comes from a different country. We make the assumption that across scores, a judge's relative scoring compared the other judges stays relatively constant. That is, conditioned on all other features, the ratio of two judges' scores stays the same. Put another way, we might believe that a judge's score is linear in the scores the other judges make. This seems a reasonable assumption that captures an intuition that one judge might consistently give higher scores than another judge. The question then is whether given this model, the addition of nation features is also informative.

Formally, suppose we can identify individual judges and associate them with specific score columns. Let i index a performance and j index judges. We model a given judge’s score as a function of the scores assigned by the other judges, along with an indicator for shared nationality:

$$\text{JudgeScore}_{ij} = \beta_0 + \beta_1 \text{SameCountry}_{ij} + \sum_{k \neq j} \beta_k \text{JudgeScore}_{ik} + \epsilon_{ij},$$

where $\text{SameCountry}_{ij} = 1$ if judge j and competitor i share a nationality. The coefficient β_1 captures whether judge j scores same-country competitors higher relative to how other judges score the same performance. The hypothesis of interest is

$$H_0 : \beta_1 = 0.$$

Data Quality and Concerns

The above hypothesis test is usually accomplished with an F-test/t-test but these tests make some broad assumptions regarding the underlying data generation process, specifically, the normality and independence of the error terms. First, we immediately know that the errors of scores from the same performance will be correlated since they are generated by the same competitor in the same "experiment". While we maybe able to eliminate this dependence by aggregating across performances, this significantly reduces our effective sample size and thus, the power of our hypothesis tests.

More broadly, we know the Olympics bounds the number of competitors from a single country and since we have only this one competition to consider, we might only have one or two athletes from a specific country competiting. Thus, there could be individual-specific errors that we simply don't have the data to distinguish between. A similar issue to this is that for F-tests, the pivot statistic can arbitrarily be inflated or deflated based on the variance of the covariates. Since we are only using a single competition to test our hypothesis, the power of our test seems like it will lower than expected for this sample size.

Finally, we seemingly have a class imbalance in the SameCountry feature. There are exactly 9 judges but either 25 or 29 competitiors depending on the short or free programs. Necessarily, this means that most scores will have $\text{SameCountry} = 0$. Thus, the predicted β_1 will have inflated errors which will in turn, deflate the true F-test statistic.

Implications and Next Steps

This analysis could serve as a useful pilot, but conclusions drawn from a single competition should be interpreted cautiously. The most obvious solution is to acquire data across more competitions, though this may require more feature engineering on our end in order reduce the errors on our coefficients. Without such data, failure to reject H_0 should not be interpreted as strong evidence of fairness, but rather as a reflection of limited power.

For now, we will proceed as follows.

0. Get the judge nationalities. It is unclear if isu.org has their nationalities but it definitely has the names of the judges so we maybe can manually do this step.
1. Merge judge nationalities into dataset or maintain a dictionary that we can use to lookup judge nationality.
2. Do an EDA on our dataset to learn if there are any feature engineering things we need to do. Check if it is even possible to distinguish judge scores or if the judge scores are randomly permuted.
3. Perform analysis described above.