Figure Skating Scores: Prediction and Assessing Bias

Acknowledgments

I would like to thank my advisors, Jim Waldo and Joe Blitzstein, for encouraging me to pursue a topic that I love. I would not have had as much fun writing about something that was not figure skating.

Thank you to my friends who provided food, edits, and valuable moral support through this process, and to my family for forgiving me for ignoring their phone calls.

Contents

Chapter 1

Introduction

Figure skating comes into the public eye once every four years during the Winter Olympics. Since the time of Nancy Kerrigan, Tonya Harding, and Michelle Kwan, figure skating has lost much of its fan base in America [1].

The country's most popular sports — basketball, football, soccer, baseball — have huge followings, markets, and statistics dedicated to them. Many teams offer analysis jobs [2]. The National Basketball Association has an entire website dedicated to player and team stats [3]. Literature on basketball has investigated player effectiveness, the hot-hand effect, and much more [4, 5].

Figure skating no longer holds enough public attention to warrant a similar amount of attention by the sports statistics community. No one is aggregating counts of jumps landed or producing win probabilities. Tickets for major international competitions rarely sell out, and elite skaters who are not at the top of the sport often struggle to make ends meet [6].

The sport is also distinct from games like basketball, football, and even chess in that success is not measured by pairwise matches. Thus, existing methods of sports statistics do not translate well to the context of figure skating, and there is a serious lack of any sort of analysis of even the highest levels of competitive figure skating.

This thesis seeks to begin filling this void of analysis with three major contributions:

- The collection and parsing of detailed scores from every major international skating competition since 2005 into a format amenable to analysis.
- The creation and evaluation of predictive models using linear regression and multilevel modeling.
- The assessment of judging bias under the current scoring system, with the conclusion

that there is strong statistical evidence for a small amount of nationalistic bias that does not significantly affect standings.

Chapter 2 aims to familiarize the reader enough with figure skating to understand the challenges of prediction and score analysis. Chapter 3 summarizes the process of data collection and parsing and gives basic visualization of the data. Chapter 4 builds a linear regression and two hierarchical predictive models. Chapter 5 assesses the question of judging bias through nonparametric and regression methods.

Chapter 2

Background

Figure skating is a judged sport, making it very different from the most popularly analyzed sports in the U.S. where success is measured by victories in matches against a single opponent or team. It also differs from situations like wine judging because there are fewer predictors available for measurement [7].

Judged sports are unique because they open up worries about bias and subjectivity. Even in basketball people have raised concerns of referee bias [8]. When judges are explicitly assigning point values to someone's performance, the stakes get even higher. In wine competitions, judges can be "blind" to what wines they are tasting [9]. In sports where judges assess based on what they see, it is impossible to avoid all human biases.

This chapter will explain details of the current scoring system that inspire our modeling approaches in prediction (Chapter 4) and assessing bias (Chapter 5). We also provide a brief background of the international competitive scene to motivate the implications of our results for the sport.

2.1 Competition

There are four disciplines of figure skating. Each discipline is a separate event at each competition, and thus skaters in different disciplines are awarded separate medals.

- Men's singles
- Ladies' singles
- Pairs
- Ice dance

Though the disciplines share the basic structure of the scoring system, they require different technical elements, and as a result the ranges of scores are different.

Competitors perform technical "elements" such as jumps, spins, step sequences, and choreographic sequences. They are also evaluated on artistry and skating skills ("program components" or "components"). The final score is broken into a technical score for the elements and a program components score for the artistry and skating skills.

International figure skating is governed by the International Skating Union (ISU). At every competition, each skater or couple skates two performances or "programs": the short program and the free skate. The short program and the free skate are scored independently of each other. These programs are known as the "segments" of a competition. We will use "segment" and "program" interchangeably. Within a segment, every competitor will perform sequentially in a predetermined start order and judges will score the performances in real time. The total score is determined by the sum of the two segment scores, and the final ranking is based on the total scores.

2.2 Scoring

Up until 2004, the ISU used the 6.0 system to score skating competitions internationally. A judging scandal at the 2002 Winter Olympics in Salt Lake City prompted the development of a new judging system [10]. The original 6.0 system aggregated individual judges' rankings of skaters [11]; the new International Judging System (IJS) ranks skaters on a much more complicated points system [12].

Each segment has a panel consisting of ISU officials who score each of the performances in the segment. Typically 9 judges sit on the panel for each segment. In addition to the judges, there is a technical panel that determines the validity of the technical elements performed.

IJS breaks a skater's segment score into two parts: the technical score and the program components score.

The *technical score* rewards skaters for the technical elements they perform. It is composed of the elements that the skater performs, so a skater receives a certain number of points per element. Each element has a base value, an indication of how difficult the element is. For example, jumps with more rotations have higher base values. For each element, every judge awards a grade of execution (GOE) for the quality of the element. GOEs are integers between -3 and 3 inclusive. These GOEs are aggregated and scaled according to the base value. The element score is then the sum of the base value and the aggregated GOE. The technical score is the sum of the points awarded for every element in the program. Technical scores tend to vary more for a skater across competitions, because skaters can lose many points for making

mistakes such as falling. However, skaters will typically receive a similar number of points for executing a clean triple axel, so technical scores can be more consistent across skaters.

The *program components* score rewards skaters for their artistry and skating skills. There are five program components simplified as follows:

- Skating Skills
- Transitions
- Performance
- Choreography
- Interpretation of the Music

Each judge assigns a point value divisible by 0.25 from 0.00 to 10.00 inclusive for each of these component types. These scores are aggregated using the trimmed mean (the mean after throwing out the highest and lowest scores). Each of these aggregated scores (there will be one for each component type) are summed and multiplied by a factor depending on the segment. The reason for this scaling is to approximately balance the technical and program components score. Program components scores tend to vary less for an individual skater between competitions, because there is less variation in a skater's artistry and skating skills between performances.

Finally the technical panel will determine any deductions. These include points taken off for falls, music and costume violations, and other rule infringements.

The segment score for a skater's program is the sum of the technical score, the program components score, and any deductions.

2.3 The Figure Skating Season

The main international skating competitions of the season run from October to March.

- The Grand Prix circuit consists of 6 international competitions in October and November. Skaters compete at a maximum of 2 of these competitions.
- The Grand Prix Final (GPF) in December invites the 6 skaters from each discipline with the highest combined placements at their 2 Grand Prix events.
- The European Championships in January permits European countries to send delegations of skaters.

- The Four Continents Championships in January or February are the equivalent of the European Championships, but for all other continents excepting Antarctica (the event's name refers to the Americas, Asia, Oceania, and Africa). These continents are combined because of the historical dominance of European figure skaters, but North American and Asian skaters now dominate many disciplines [13].
- The Olympic Winter Games occur once every four years in February. Usually skaters view qualifying for or medaling at the Olympics to be their ultimate goal.
- The World Championships are the final event of the season in March.

2.4 Existing Literature

We will discuss relevant literature in the appropriate chapters.

2.5 Implications

We now discuss how this background motivates our methods in the subsequent chapters.

For Chapter 4 (prediction), the distinctions between the four disciplines of figure skating suggest separate predictive models for each. Because skaters rarely move between disciplines, we gain no new information from combining information across disciplines in prediction.

The similarity across skaters in scoring technical elements suggests that we gain information from observing skaters in a discipline collectively. The close relation between skaters and their program components scores suggest that we should model some sort of inherent, per-skater quality, or at least some sort of reputation factor. These properties motivate us to pursue hierarchical models in prediction: we can learn something about a particular technical element from all of the skaters, but some skaters are better at certain types of elements than other skaters.

Because the four disciplines use the same scoring system, when answering questions about judging bias in Chapter 5 we can more safely aggregate the data. In applying methods to determine the existence of bias, we should control for skater effect when looking at program components scores because of the strong per-skater quality evident in components.

The main competition types demonstrate what kind of data we have, and explain which competitions are more significant for the skating community.

Chapter 3

Data

The results for every international competition listed in the background chapter since the introduction of IJS are available online; see Appendix A. However, the detailed scores containing GOEs, base values, and every judge's individual marks are available only in PDF format. We asked the ISU for data in a more easily parsed format, which they did not provide.

Therefore the first major task of this thesis was to find a way to assemble all of this data into a form amenable to statistical analysis. This chapter outlines the basic steps, with more details in Appendix B. Many of the techniques were inspired by Mitchell (2015) [14]. We then visualize the distributional properties of the data to preface applying statistical techniques in later chapters.

3.1 Overview

All analyses in this thesis use Python 2.7 [15]. All plots were generated using `matplotlib` [16].

The first step of data collection was to retrieve all of the published competition information available at the links listed in Appendix A. This was relatively straightforward and was accomplished using the `requests` [17] and `BeautifulSoup` [18] libraries.

The second step was to parse the non-scorecard information, which was stored in HTML files: results summaries, panels, and entries. This was also straightforward using `BeautifulSoup` [18].

The third and most difficult step was the parsing of scorecards. Though the distribution of scorecards in PDF format is likely convenient, the IJS is inconsistent in the visual design of

PDFs, even in the same season. Fortunately the general structure has remained similar. More details are available in Appendix B.

The last step was data cleaning for situations where scoring was incomplete.

Once the data was in a consistent format, it was straightforward to create and manipulate datasets using `pandas` [19].

3.2 Data Cleaning

We detail the main assumptions made for incomplete information.

For competitions with many entrants, only the top 16-24 skaters after the short advance to the free skate. Thus the reported "total scores" for some skaters at certain competitions may only reflect one program, because those skaters who do not qualify for the free skate will only perform the one short program.

We include total scores for competitions where free skate qualification is necessary, so some total score data points will only reflect one program. However, free skate qualification is a very telling metric as there is a class of skaters who are almost never in danger of not qualifying. We therefore include these results.

We do remove withdrawals, because they do not reflect the complete performance of a skater. One competition, the Grand Prix of France in 2015, was canceled after the short program due to the November 2015 Paris attacks. The short program scores were saved as the total scores, so skaters with otherwise high total scores have this outlier on their resume. Therefore we remove these data points when using total score as a predictor.

We also had to resolve inconsistent spelling of skater and judge names across competitions, which is discussed in Appendix B.

3.3 Data Structure

Figure 3.1 shows an example of a scorecard, which is the detailed breakdown of a skater's segment score. The top row of information is a summary of the scorecard, with skater information, the technical score, the program components score, the deductions, and the segment score. The top half of the scorecard lists out the elements that the skater performed. The bottom half lists the program components. Each column underneath "The Judges Panel" represents the scores of a single judge for every element and component.*

*Note in that the judging panel here is presented "in random order." This randomness was introduced with the premiere of IJS to protect judges from being pressured to vote with a particular bloc. Though anonymous judging has since ended, at the time this meant that scores could not be deterministically matched to particular judges. Columns of scores within the same scorecard represented the same judge. However, it is not clear

Rank	Name	Nation	Starting Number	Total Segment Score	Total Element Score	Total Program Component Score (factored)	Total Deductions
1	Yuzuru HANYU	JPN	5	110.95	61.81	49.14	0.00

#	Executed Elements	Info	Base Value	GOE	The Judges Panel (in random order)									Ref	Scores of Panel
1	4S		10.50	3.00	3	3	3	3	2	3	3	3	3		13.50
2	4T+3T		14.60	3.00	3	3	2	3	3	3	3	3	3		17.60
3	FCSp4		3.20	1.29	3	3	2	3	2	2	2	3	3		4.49
4	3A		9.35 x	2.71	3	3	3	2	2	3	3	3	2		12.06
5	CSSp4		3.00	1.43	3	3	3	3	3	3	2	2	3		4.43
6	StSq3		3.30	1.50	3	3	3	3	2	3	3	3	3		4.80
7	CCoSp3p4		3.50	1.43	3	3	2	3	3	3	2	3	3		4.93
			47.45												**61.81**
	Program Components			**Factor**											
	Skating Skills			1.00	9.75	9.75	9.50	9.75	9.00	10.00	9.50	10.00	9.75		9.71
	Transition / Linking Footwork			1.00	9.75	9.75	9.50	9.75	9.00	9.50	9.50	9.50	9.75		9.61
	Performance / Execution			1.00	10.00	10.00	10.00	10.00	9.75	10.00	10.00	10.00	10.00		10.00
	Choreography / Composition			1.00	10.00	10.00	10.00	10.00	9.25	9.75	9.75	10.00	10.00		9.93
	Interpretation			1.00	10.00	10.00	10.00	10.00	9.50	10.00	9.75	9.50	10.00		9.89
	Judges Total Program Component Score (factored)														**49.14**
	Deductions:														**0.00**

x Credit for highlight distribution, base value multiplied by 1.1

Figure 3.1: The scorecard for Yuzuru Hanyu's world record short program at the 2015-2016 Grand Prix Final [20]. The top row contains summary information. The rest of the top half details the technical score. The bottom half breaks down the program components score. Each column under "The Judges Panel" represents all the scores of a single judge.

We store all of the data in the above scorecard in the data structures detailed in Appendix B and maintain all of the structure shown. Once the data was parsed and stored in this structured format, it was easy to pull out the desired information for analysis.

3.4 Exploratory Data Analysis

There are 13 seasons of data, from 2005-2006 to 2017-2018, for a total of 134 competitions. We give the number of individual skater results in Table 3.1. One data point is one skater's overall result in a competition, so at a particular competition, a discipline with 6 skaters would give 6 data points. We also report the number of unique competitors per discipline. In pairs and ice dance, sometimes an individual will have multiple partners throughout the years; we count each of those couples as a unique skater. The men and ladies disciplines are the most popular, as they lack the barrier to entry of finding a partner, which is necessary for pairs and ice dance.

whether the order of the judges would be the same for all skaters' scorecards. Emerson and Arnold (2011) assert that the columns were permuted between scorecards at the 2010 Olympics [21]. ISU regulations, particularly outdated ones, are very hard to find online and it is not clear what the exact rules were surrounding this randomness. However, anonymous judging ended for the 2016-2017 season, so data from fall 2016 onward enables the matching of judges to scores and resolves the ambiguity of judge shuffling.

	men	ladies	pairs	ice dance
Total Number of Results	2218	2311	1358	1716
Number of Unique Skaters	318	404	213	264

Table 3.1: Summary of the total amount of data available per discipline, 2005-2018. One data point or "result" is one skater's result at a competition.

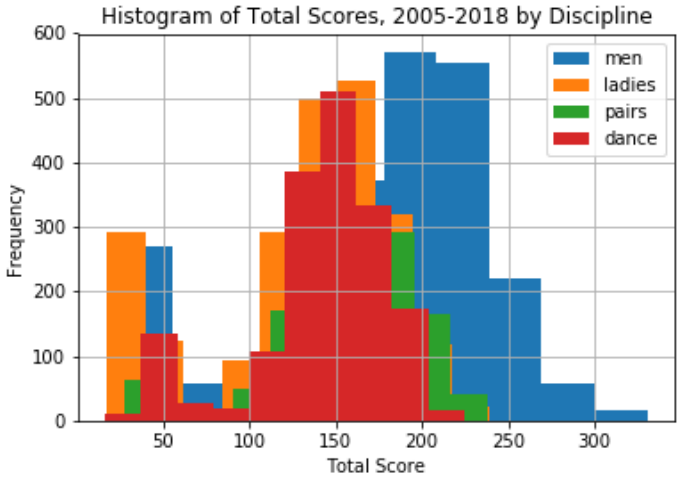

Figure 3.2: Total score distributions for the four disciplines, aggregated over all the data. Men's scores are much higher on average. The bimodal distributions reflect our choice to leave in failure to qualify to the free skate as a "total score" data point.

3.4.1 Total Scores and Competitors

Figure 3.2 shows histograms for the total scores, separated by disciplines using different colors. Men's total scores are clearly much higher on average than in the other three disciplines: because they perform more difficult elements, they receive higher technical scores. To balance out this technical score, their program components factor is also higher than in the other disciplines. Ladies, pairs, and ice dance scores are distributed more similarly, though we can see that the highest scores for ladies and pairs are higher than the highest scores for ice dance.

The total score distributions are bimodal due to the decision to include data points where skaters failed to qualify to the free skate. These data points have much smaller total scores because the total score does not include the free skate.

In Figure 3.3 we plot some statistics about competitors and score trends over time. The first plot shows the number of unique competitors appearing in each season. There has been a drop in the number of men's and ladies' competitors since pre-2010, though the numbers have climbed back for the 2017-2018 season. There are clear peaks for the Olympic years (2010, 2014, 2018), except in the pairs discipline for the 2018 season.

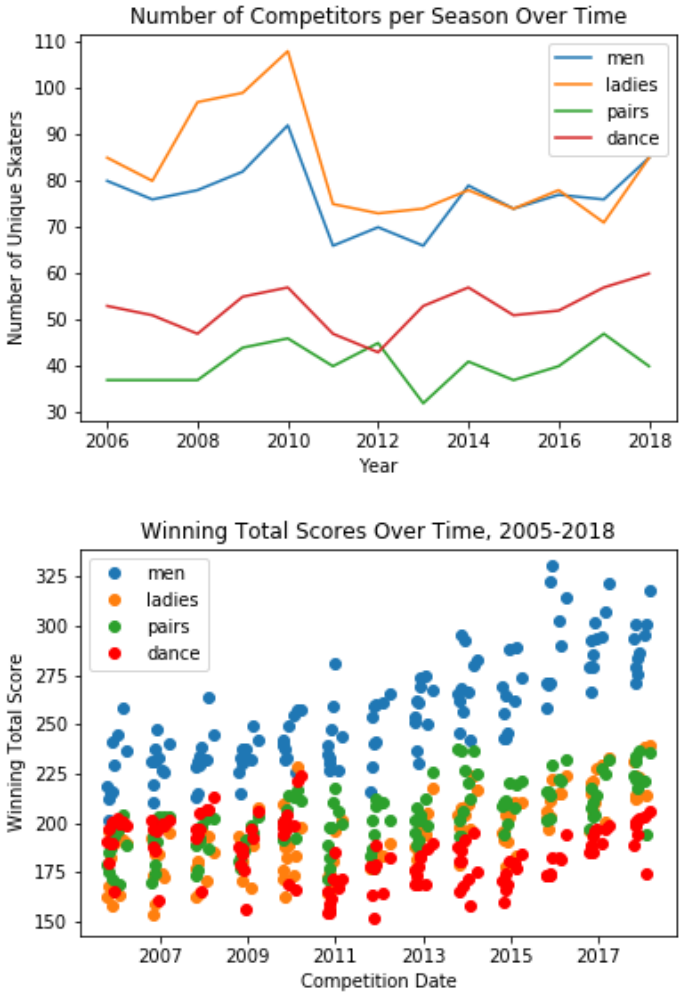

Figure 3.3: Trends in the number of competitors and total scores over time by discipline. **Top**: The number of unique competitors in a season over time. Note the local maxima in Olympic years (2010, 2014, 2018). **Bottom**: The total score that won each competition plotted against the date of the competition. Clearly scores have been increasing over time.

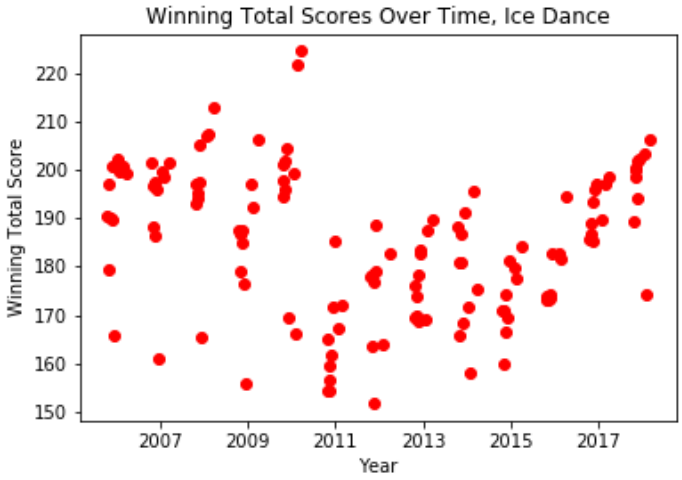

Figure 3.4: Winning total scores over time in the ice dance discipline. The data is confusing prior to 2011, when ice dancers would perform either 2 or 3 programs per competition. Starting in the 2010-2011 season, we see the more consistent trend of score increase over time.

The second plot of Figure 3.3 shows the winning score plotted against competition date. These scores have increased steadily over time, with the most significant change in the men's discipline where technical difficulty has increased the most dramatically.

In ice dance, there appears to be a drop in 2010. We plot only the ice dance data points in Figure 3.4. The data points show no clear trend before 2010, but then total scores grow steadily between 2011 and 2018. The reason for this is a change in the structure of ice dance. Ice dancers used to compete up to 3 total programs per competition. The third was removed after the 2009-2010 season. However, not every competition in these earlier seasons required the ice dancers to perform all 3 programs, which is why we see some very low winning total scores in the earlier data. For instance, at the 2009-2010 Grand Prix Final the ice dance competitors only performed 2 programs [22], while at the 2010 Olympics just three months later they performed 3 [23].

Thus, in fitting any ice dance model we will only use the data starting in the 2010-2011 season to avoid confusion with the number of programs.

Finally, in Figure 3.5 we show histograms of ice dance scores from 2005-2018 for teams from three countries, USA, Italy, and China, as an example of "powerhouse countries" in figure skating. The total scores of all US ice dance teams are shown in the red histogram: the US has more results and higher total scores than the other two countries. China has the fewest results, and they are all lower scores. Italy is somewhere in the middle. This reflects these countries' statuses in ice dance: the US has been very successful since 2006, Italy has a developing

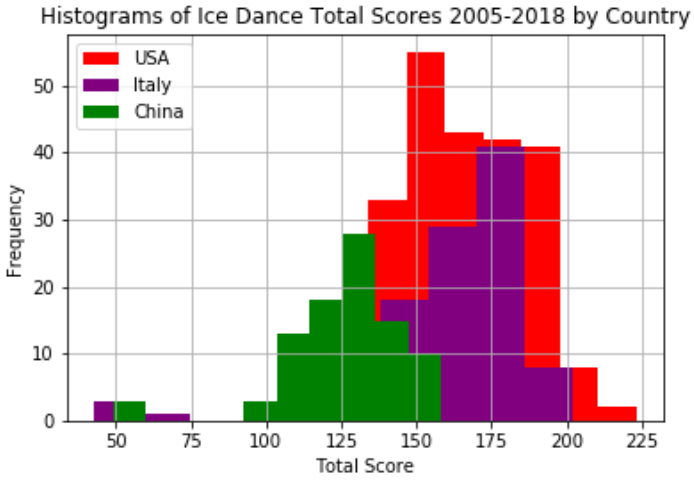

Figure 3.5: Histograms of total ice dance scores for teams from USA, Italy, and China in 2005-2018, illustrating the relative strength of certain countries in this discipline. The US has more teams competing and outperforming the other countries. China has a much smaller, less high-scoring contingent, while Italy is somewhere in the middle.

program that has produced a world champion team, and China does not place as much of an emphasis on ice dance (their strength is pairs).

3.4.2 Personal Bests and Start Order

In Chapter 4, one predictive model uses skating order and personal best scores in each segment as predictors.

As an example, in Figure 3.6 we show the histograms of personal best segment scores (the highest seen in our dataset) for individual skaters in the men's discipline. We separate by the short and the free. These have approximately Normal distributions. In Chapter 4 we will refer to these predictors as "reputation."

In general skaters who perform near the end of the skate order receive higher scores, for reasons we will detail later [24, 25]. Figure 3.7 shows this trend for competitions in the 2016-2017 season. We plot "normalized start order," which is the start order divided by the total length of the skating order, against free skate score. There is a clear upward trend in every discipline, reinforcing our belief that program scores will increase later in the skating order.

3.4.3 GOEs and Components

In Chapter 2 we discussed grades of execution (GOEs) scored for technical elements and program components scores (components) awarded for skating skills and artistry. In this section

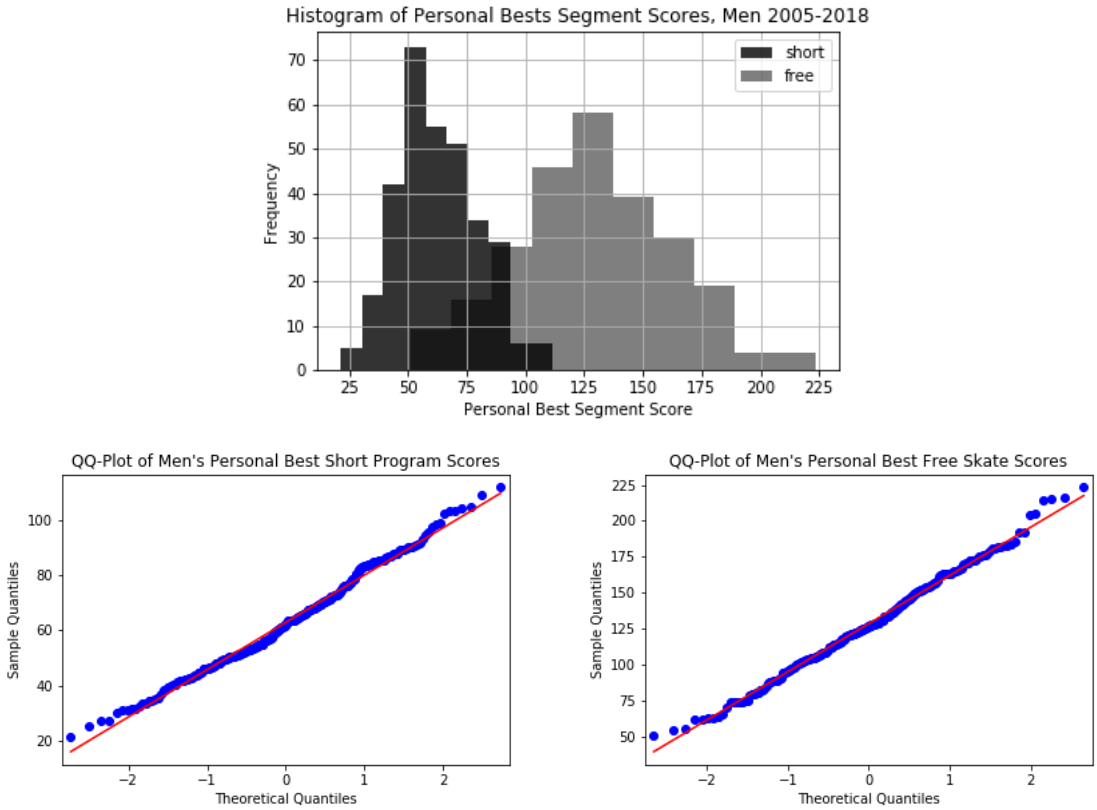

Figure 3.6: Personal best segment scores in the men's discipline. Both the short program and free skate personal bests are approximately Normal. **Top**: Histogram of personal best segment scores separated by short and free. **Bottom-left**: QQ-plot of men's personal best short program scores. **Bottom-right**: QQ-plot of men's personal best free skate scores.

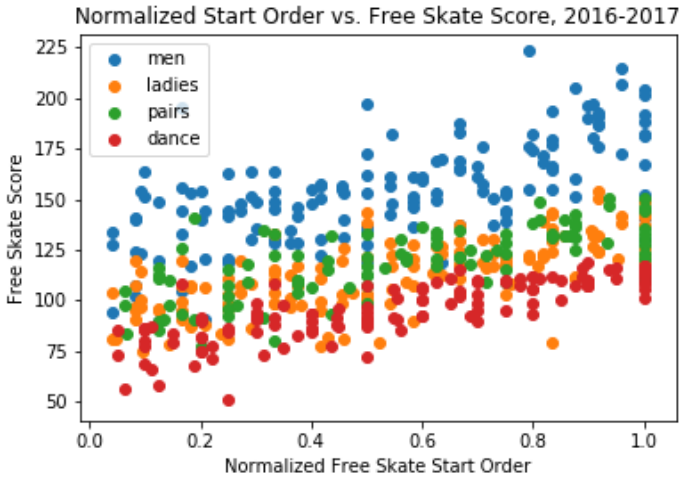

Figure 3.7: Normalized free skate start order vs. free skate score for the 2016-2017 season, by discipline. "Normalized start order" is the starting number of a skater divided by the total number of skaters in the skate order. There is a clear upward trend: later start orders tend to receive higher scores.

we examine their general distributions for the *individual* judges' scores, not the aggregated score.

Figure 3.8 shows histograms of GOEs in the data. Since the judges' individual GOEs are discrete and range from -3 to 3, they cannot be well approximated by continuous distributions. We observe all possible GOEs, and there is no extremely obvious distribution difference between the 2016-2017 GOEs and the aggregated GOEs, except perhaps a slight shift towards the right (higher GOEs) in the 2016-2017 data.

Recall components scores are between 0.00 and 10.00 in increments of 0.25, so although they are also discrete, the larger domain means they could be approximated by a continuous distribution such as a scaled Beta. Figure 3.9 shows histograms of individual judges' component scores in the data. We observe very few scores less than 4.00. While the aggregated scores look symmetric, the components in the 2016-2017 show a clear left skew. Therefore when modeling component scores, we want to be careful about our choice of distribution depending on what slice of data is under consideration.

We check to see if there are differences by discipline in Figure 3.10. There are no drastic differences in distributions of either GOEs or components between distributions: they all have most of the probability mass in the center of the distribution. Ice dance seems to have higher GOEs on average, most likely due to the less risky elements. They also seem to have higher components scores, which also makes sense because ice dance requires stronger foundational skating skills than the jumping disciplines (men, ladies, pairs).

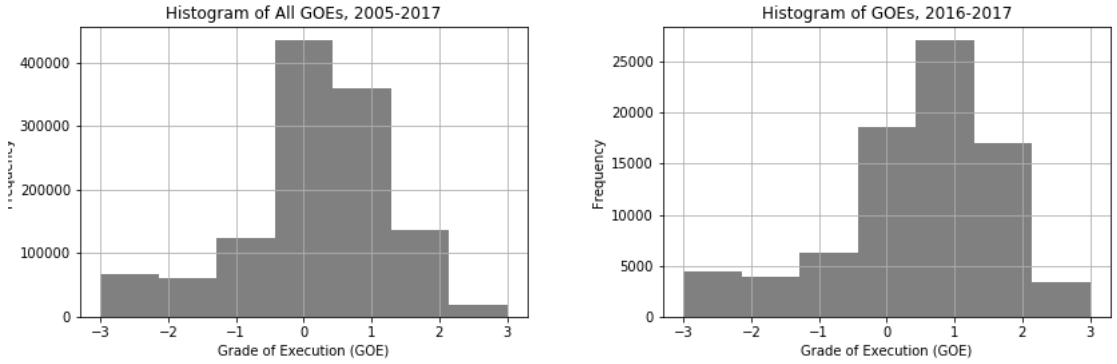

Figure 3.8: Histograms of individual judges' unaggregated grades of execution (GOEs). GOEs in the 2016-2017 season are slightly higher than GOEs across all years, 2005-2017. **Left**: All GOEs in the dataset. **Right**: GOEs in the 2016-2017 season.

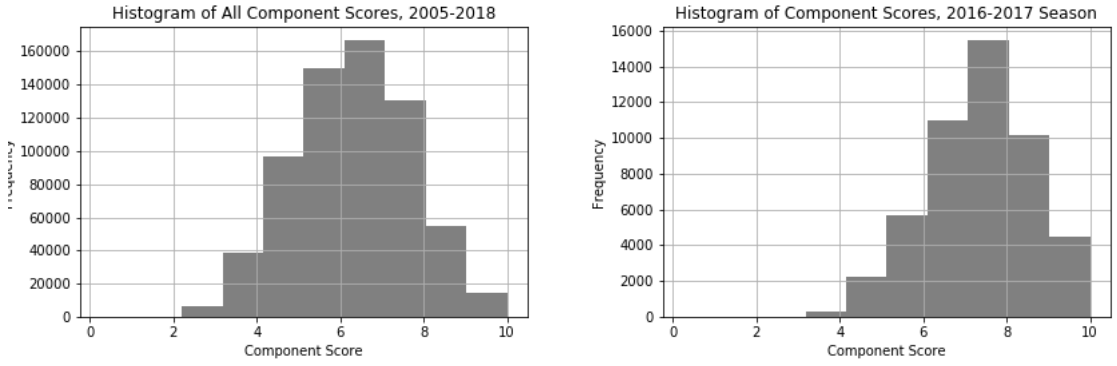

Figure 3.9: Histograms of individual judges' unaggregated component scores. **Left**: All components in the dataset. **Right**: Components in the 2016-2017 season. This distribution shows a left skew.

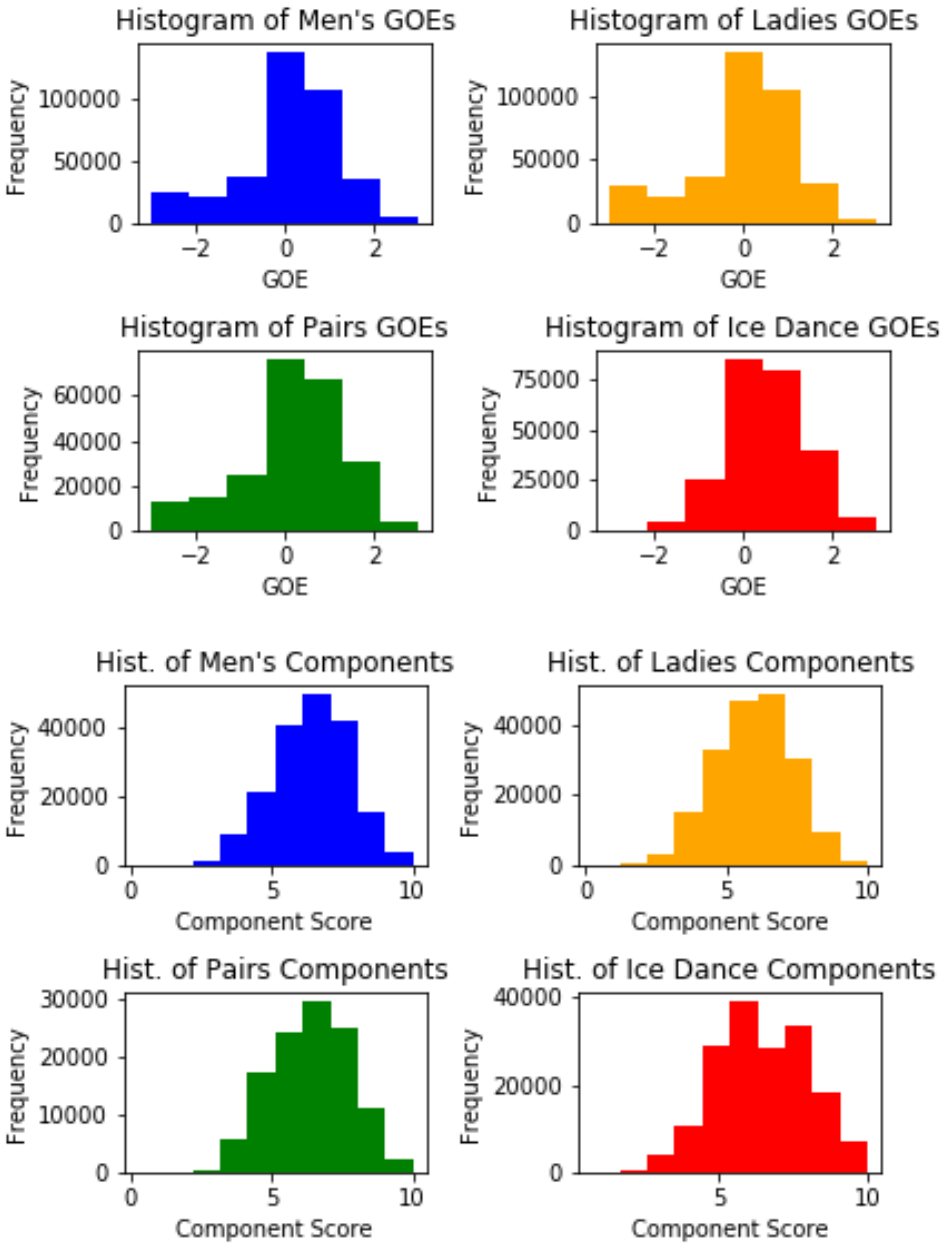

Figure 3.10: Histograms of GOEs and component scores separated into the disciplines. There are no huge differences between distributions, though ice dance seems to exhibit slightly higher scores on average. **Top**: Histograms of GOEs by discipline. **Bottom**: Histograms of component scores by discipline.

Chapter 4

Predictive Models

Fans of figure skating both love and hate predictions. One of the biggest obstacles a skater faces is the mental strength needed to perform consistently, and as a result some skaters are much more consistent than others. Even favorites with the highest score "ceiling" can perform poorly.

When building models, it makes the most sense to consider each discipline separately because of the huge differences in score range. The highest men's scores are significantly higher (total score 300+ points) when compared to the highest ladies' scores (total score 230+) because of the higher technical difficulty in men's programs. Pairs and ice dance skaters perform different types of elements altogether, so the comparison is even weaker there.

In this chapter we describe three predictive models. Each of these models is fitted separately for the four disciplines.

- a linear regression model (OLS) that predicts based on a skater's highest previous score and the skater's starting order in the competition
- Hierarchical Model 1 (HM 1) which predicts group and per-skater point distributions for different element and component types
- Hierarchical Model 2 (HM 2) which modifies the priors of HM 1 and adds predictors based on time

First we give the motivation behind developing these models. Then we describe each model in detail. Then we state and interpret our results. We apply HM 2 and the linear regression to the 2018 World Championships as an example of how real prediction would work. Finally we discuss potential improvements.

4.1 Methods

We attempt prediction using the data we have available, which is all of the detailed historical scores. Figure skating is unique from other sports because the scores of different skaters are roughly independent: because competition is not based on pairwise matches, the majority of a skater's score is determined by what that skater performs.

As in any sport, some athletes are better than others. The simplest predictive model considers historical results: skaters who have scored well in the past will tend to score well in the future. This is due to both an objective difference in quality of skating and the reputation bias of judges remembering good performances from previous competitions.

Thus the first predictive model we attempt in this chapter is a simple linear regression model that uses a skater's best historical score as a predictor of their points earned at a competition. We add the predictor of starting order within a segment, because it is generally believed that scores will be higher towards the end of a competition [24, 25], and because without this additional information we would just rank based on personal best scores.

We use the highest score seen so far rather than the median, mean, or some other combination of a skater's past performances. Judges are more likely to remember standout performances than average ones, so a higher reputation value should generally indicate a higher chance of performing well. It usually takes a breakout performance for a skater to be well-regarded by the skating community. We will further assess the use of this predictor when giving the fitted model.

To build something interesting, and in order to take advantage of the structure of the data, in our subsequent hierarchical models we break down prediction into the pieces of a score that a skater will receive: technical elements and program components.

Skaters perform technical elements, and these technical elements receive a base value. Thus if different skaters perform the same elements to a similar level of quality, they should receive a similar number of points. Some skaters will be better at certain elements than others, but we also want to use the shared information across skaters about elements. This motivates a hierarchical model with grouping by skaters to estimate points received for elements. A hierarchical model helps with the prediction for new skaters when we lack individual data for them.

Hierarchical models can also model program component scores. Program component scores tend to be more consistent for individual skaters, but it also reasonable to think of the component scores of an individual skater as a "sample" from a group distribution that accounts for all skaters.

With hierarchical models, we make the crude assumption that all elements are independent

of each other and of the program components. We also take most distributions to be Normal for simplicity, and because the Normal captures both the distribution of a population's skill level and the distribution of an individual skater's earned points.

For variances, we use Half-Cauchy distributions for a weakly informative prior with the appropriate domain [26, 27].

For each of the models we describe, we must assume the following:

- Within a competition or segment, the performances and scores between skaters are independent of each other, because skaters perform individually.
- A skater's performances are all independent. This is not a fully reasonable assumption, but the factors that affect a skater competition to competition are difficult to model.

For the linear regression model, we assume we know the skating order for each program in advance. This is not completely realistic, as we only know the short program start order shortly beforehand and do not know the free skate start order until after the short program. The typical observation in figure skating is that skaters later in the order score higher, for the following reasons:

- In the short program, start order is seeded by world ranking in international competitions, so higher-ranked skaters will skate later.
- In the free skate, start order is determined based on ranking in the short program. Thus the skaters who skate later will have a higher change of winning due to their leads in the short program. See Chapter 3, Figure 3.7 for visual evidence of this particular trend.
- Judges are also aware of the first two factors, and may award higher scores for later skaters as a result.

To fit linear regressions, we use the `statsmodels` Python package [28]. To fit multilevel models, we use the `pymc3` package [29]. Inspiration for the hierarchical models comes largely from the `pymc3` documentation [26] and Gelman and Hill's (2007) book on multilevel modeling [27].

4.2 Linear Regression (OLS)

As a starting point, we model a skater's score as a simple linear regression:

$$\text{short}_{\text{skater},i} \sim N(\beta_{0,\text{short}} + \beta_{1,\text{short}} \cdot r_{\text{short, skater}} + \beta_{2,\text{short}} \cdot \text{start}_i, \sigma^2_{\text{short}})$$

Discipline	Segment	Number of Observations	Number of Skaters
men	short	170	76
	free	144	62
ladies	short	169	72
	free	144	60
pairs	short	111	47
	free	97	44
dance	short	142	57
	free	120	46

Table 4.1: Number of observations and unique skaters per linear regression. Recall that observations are results from the 2016-2017 season, while reputation predictors are derived from previous seasons.

$$\text{free}_{\text{skater},i} \sim N(\beta_{0,\text{free}} + \beta_{1,\text{free}} \cdot r_{\text{free, skater}} + \beta_{2,\text{free}} \cdot \text{start}_i, \sigma^2_{\text{free}})$$

Essentially each skater's segment score $\text{short}_{\text{skater},i}$, or $\text{free}_{\text{skater},i}$ is Normally distributed based on their "reputation" in the short and the free and their start order in that segment.

- $r_{\text{short, skater}}$ is the skater's "reputation" in the short, or the highest short program score the skater has ever received in our dataset prior to this competition. Similarly $r_{\text{free, skater}}$ is the highest free skate score a skater has ever received. If a skater has no previous observations (i.e. this is their first season appearing in any major senior international competition), then we set these predictors to the median of all of the unique skater reputation values of the given program type.

- start_i is the normalized start order of performance i in its segment of a competition. This will be a value in $(0, 1]$, so a performance that is 6th out of 12 total skaters will have a predictor of 0.5. Because start order reflects world ranking and performance in the short program, this is a helpful predictor that gives more recent information about a skater.

For each of the men, ladies, and pairs discipline, we determine reputation predictors based on historical scores up to and including the 2015-2016 season. We then use 2016-2017 as our training dataset of response variables and fit each of these models. For the dance discipline, we determine reputation predictors based on data from the 2010-2011 season to the 2015-2016 season inclusive, because prior to the 2010-2011 season ice dancers skated three programs [12].

We model the short program and free skate separately to capture the idea of having two separate performances to predict and to account for skaters who may be stronger at either the short program or the free skate.

Table 4.1 indicates the number of observations we have for each linear regression: these are overall total score data points from the 2016-2017 season. The number of skaters indicates the

Discipline	Segment	R^2	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
men	short	0.62	20.82	0.61	12.32
	free	0.58	72.39	0.40	35.29
ladies	short	0.45	21.83	0.53	10.52
	free	0.49	71.83	0.19	35.10
pairs	short	0.51	26.65	0.51	10.42
	free	0.66	57.29	0.39	28.91
dance	short	0.82	11.15	0.81	3.89
	free	0.83	21.52	0.70	12.67

Table 4.2: Estimated coefficients for each linear predictive model. Every coefficient is reported to be significant at the 0.01 level. Recall β_0 is the model intercept. β_1 is the percentage of "reputation" contributing to the score. For instance, in the pairs free we expect teams to score at least 57.29 plus 0.39 times their personal best free skate score. β_2 describes the effect of start order. It is approximately the point increase a skater would receive in going last in the order as opposed to first.

number of unique skaters or teams in the response vector.

We report the fitted coefficients and R^2 for each linear regression in Table 4.2. In all models every predictor is tested to be significant at the 0.01 level. We first note that the R^2 and estimated reputation coefficient $\hat{\beta}_1$ is higher for ice dance compared to the rest of the disciplines. This is because ice dance scores exhibit much less variance due to fewer mistakes in lower-risk elements. We expect a skater to receive $\hat{\beta}_0 + \hat{\beta}_1 \cdot r$ points for a program, plus a boost depending on the start order. We expect a skater who goes last to receive about $\hat{\beta}_2$ more points than if they were to skate first.

Figure 4.1 shows the relationship between the men's short program reputation predictors and predicted and observed outcomes in our training data, the 2016-2017 season. There is a clear linearity with the observed scores (blue data points) and the reputation predictor of highest previous short score. Note the vertical line of dots at $r_{\text{short}} \approx 60$ is the set of skaters with no prior results in our data set, so we predict the median personal best short program score of 60.81.

4.3 Hierarchical Model 1 (HM 1)

We want to take advantage of our structured data to predict the average performance for skaters on different kinds of elements and components. Our strategy will thus be to model the point distributions of different *types* of elements (jumps, spins, etc.) and components (skating skills, performance, etc.) as separate hierarchical models to capture information about different types of elements for all skaters.

We group elements as described in Table 4.3; it is not necessary to understand the groups

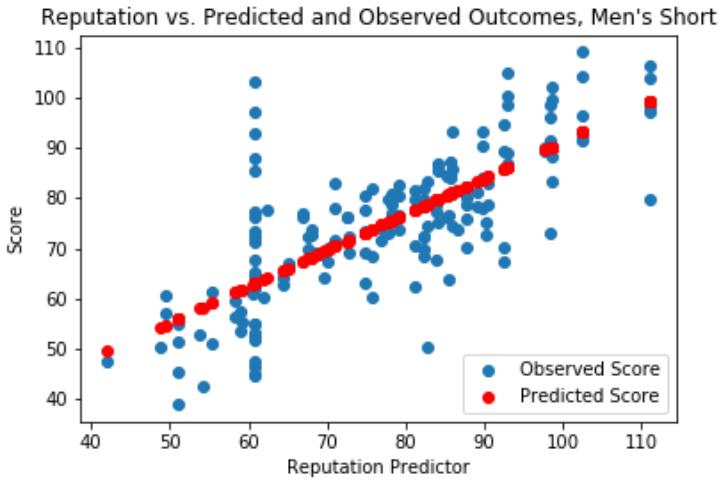

Figure 4.1: Reputation predictors plotted against observed and predicted outcomes for the men's short program in the 2016-2017 season. There is a clear linear relationship. The cluster of points with a reputation predictor of approximately 60 is the set of skaters with no prior results in our data set.

Discipline	Group Name	Description
men, ladies	`ch`	choreography sequence
	`st`	step sequence
	`sp`	spin
	`1j`	single jump
	`2j`	double jump
	`3j`	triple jump
	`4j`	quadruple jump
pairs	`ch`	choreography sequence
	`st`	step sequence
	`sp`	spin
	`li`	lift
	`tw`	twist lift
	`th`	throw jump
	`ds`	death spiral
	`ju`	side-by-side jump
ice dance	`ch`	choreography sequence
	`st`	step sequence
	`sp`	spin
	`li`	lift
	`tw`	twizzles
	`pd`	pattern dance (e.g. Rhumba pattern)
	`l2`	two lifts recorded as one element

Table 4.3: Element group types defined per discipline. For more information on pair and ice dance elements, refer to their Wikipedia pages [30, 31].

except to recognize that they are different categories of elements. We also break down components into the five categories scored: skating skills, transitions, performance, choreography, and interpretation.

For each of these groups, we fit a hierarchical model of the following specifications. Consider element or component type t. Then our model is:

$$\begin{aligned} \mu_{\text{group},t} &\sim \text{prior}_t & \sigma^2_{\text{group},t} &\sim \text{HalfCauchy}(5) && \text{(priors)} \\ \mu_{t,\text{skater}} &\sim N(\mu_{\text{group},t}, \sigma^2_{\text{group},t}) & \sigma_t^2 &\sim \text{HalfCauchy}(5) && \text{(group distributions)} \end{aligned}$$

$$\text{points}_{t,\text{skater}} \sim N(\mu_{t,\text{skater}}, \sigma_t^2)$$

$\mu_{\text{group},t}$ is the aggregated mean of points for element or component type t across all skaters. $\sigma^2_{\text{group},t}$ is the variance of mean points earned for t across skaters. $\mu_{t,\text{skater}}$ is skater's mean points for t, and σ_t^2 is the performance variance for each skater each time they perform element or component t.

We consider each skater's mean points earned for type t to be a draw from a group distribution. Thus this model accounts for (1) skaters' performance on an element or component type generally and (2) differences among skaters on an element or component type. We assume for simplicity that for each element or component type t each skater has the same performance variance σ_t^2, but this is not necessarily the case: some skaters are more consistent than others, and will have a narrower point distribution.

Table 4.4 details each of the group priors. We take uninformative priors on the group parameters, but provide a prior mean based on the typical base value for each kind of element. We specify the prior on `1j` (single jumps) to be an Exponential distribution. This is because single-revolution jumps tend to be invalidated quite a lot, and so we see many observations near zero.

For every component type c, we take the prior on the group mean to be:

$$\mu_{\text{group},c} \sim N(7.0, 10^5)$$

As seen in Chapter 3, the mean of component scores is around 7.0.

Though this model takes significantly more effort to fit using `pymc3` and does not beat linear regression in predicting skater scores, it is informative in the variation we see across skaters for certain element types.

For convenience we do not report the fitted distributions, but consider for example Figure 4.2 which gives the fitted probability distributions of the parameters for spin elements in the men's discipline. We see that $\mu_{\texttt{sp},\text{skater}}$ varies significantly across skaters. The right most

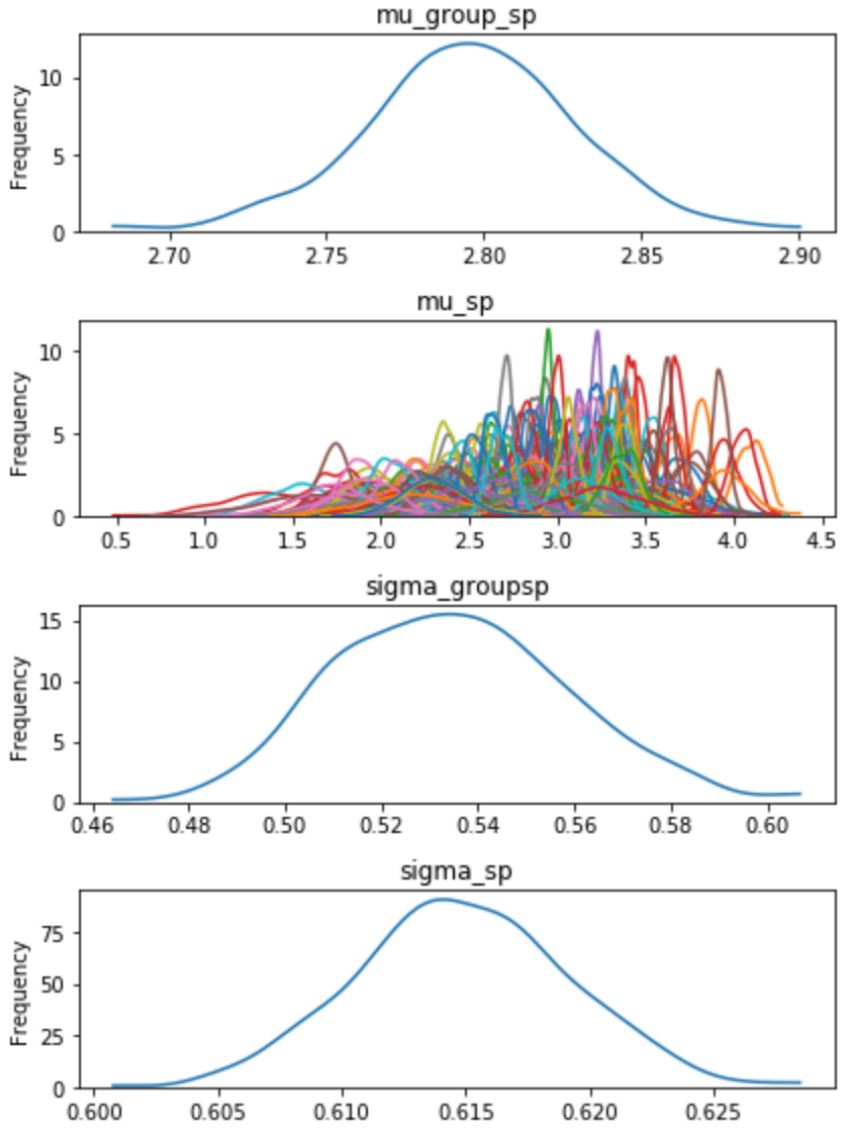

Figure 4.2: An example of the traceplots of the fitted hierarchical models by `pymc3`. These are Hierarchical Model 1 distributions on parameters for points earned on spins in the men's discipline, and illustrate the additional information that a hierarchical model can provide. **Top**: The distribution of $\mu_{\text{group,sp}}$. We believe $\mu_{\text{group,sp}}$ is around 2.8 based on our data, meaning that the average skater earns 2.8 points on average for a spin. **Top-middle**: The distributions of $\mu_{\text{sp,skater}}$ for each skater in our training dataset. Notice there is a wide variation of curve heights and widths. **Bottom-middle**: The distribution of $\sigma_{\text{group,sp}}$. This is our distribution of the standard deviation across the means of spin points earned by skaters. **Bottom**: The distribution of σ_{sp}. This is the estimate of the standard deviation in individual skaters' spin performances around individual skaters' spin means $\mu_{\text{skater,sp}}$.

men, ladies	pairs	ice dance
$\mu_{\text{group,ch}} \sim N(1.0, 10^5)$	$\mu_{\text{groupst}} \sim N(3.0, 10^5)$	$\mu_{\text{group,tw}} \sim N(6.0, 10^5)$
$\mu_{\text{group,st}} \sim N(3.0, 10^5)$	$\mu_{\text{group,sp}} \sim N(3.5, 10^5)$	$\mu_{\text{group,st}} \sim N(7.0, 10^5)$
$\mu_{\text{group,sp}} \sim N(2.5, 10^5)$	$\mu_{\text{group,tw}} \sim N(5.0, 10^5)$	$\mu_{\text{group,pd}} \sim N(4.0, 10^5)$
$\mu_{\text{group,1j}} \sim Expo(1.5)$	$\mu_{\text{group,th}} \sim N(4.0, 10^5)$	$\mu_{\text{group,li}} \sim N(4.0, 10^5)$
$\mu_{\text{group,2j}} \sim N(4.0, 10^5)$	$\mu_{\text{group,li}} \sim N(6.0, 10^5)$	$\mu_{\text{group,l2}} \sim N(8.0, 10^5)$
$\mu_{\text{group,3j}} \sim N(6.0, 10^5)$	$\mu_{\text{group,ds}} \sim N(4.0, 10^5)$	$\mu_{\text{group,sp}} \sim N(4.0, 10^5)$
$\mu_{\text{group,4j}} \sim N(10.5, 10^5)$	$\mu_{\text{group,ch}} \sim N(2.0, 10^5)$	$\mu_{\text{group,ch}} \sim N(2.0, 10^5)$
	$\mu_{\text{group,ju}} \sim N(3.0, 10^5)$	

Table 4.4: The priors on the group means for each element type, $\mu_{\text{group},t} \sim \text{prior}_t$. The prior means are approximately based on typical base values for the different kinds of elements, but we make the priors non-informative by specifying large variances.

curve in the second plot indicates the distribution of the mean points that the best spinner in the men's discipline will receive for a spin. This turns out to be Jason Brown, who is considered to be one of the best spinners in the world.

4.4 Hierarchical Model 2 (HM 2)

One of the issues with HM 1 is that it underpredicts, particular in components marks. This is because it considers each data point to be equally "weighted" — the time at which we observed it does not matter. However, skaters tend to improve over time and typically see their scores go up. Scores have also generally increased dramatically since the introduction of IJS. Consider Shizuka Arakawa's winning score of 191.34 in the 2006 Olympics ladies' event. That score would not have been on the podium at the 2010 Olympics, just four years later. This trend was visualized in Chapter 3, Figure 3.3. Thus we would like to include some sort of time metric in our model to account for more recent data points being more useful, and to capture the trend of improvement over time. Gelman suggests encoding a functional dependence on time lag [32].

Another phenomenon in the judging system is that a skater's program component marks do not exhibit a huge amount of variance across different types of components. Often less than half a point separates the highest component mark from the lowest component mark of a skater in a program, which results in less than a point difference in the overall score. Thus it is more efficient to not split components into their separate types, but predict them as a group.

For a type of score t (including element types and "component" on its own), we can describe HM 2 as follows.

Priors

$$\alpha_{\text{group},t} \sim \text{prior}_t \qquad \sigma^2_{\alpha_t} \sim \text{HalfCauchy}(5)$$
$$\beta_{\text{group},t} \sim N(0, 10^5) \qquad \sigma^2_{\beta_t} \sim \text{HalfCauchy}(5)$$
$$\gamma_{\text{group},t} \sim N(0, 10^5) \qquad \sigma^2_{\gamma_t} \sim \text{HalfCauchy}(5)$$

Group distributions and group-level parameters

$$\alpha_{t,\text{skater}} \sim N(\alpha_{\text{group},t}, \sigma^2_{\alpha_t})$$
$$\beta_t \sim N(\beta_{\text{group},t}, \sigma^2_{\beta_t})$$
$$\gamma_{t,\text{skater}} \sim N(\gamma_{\text{group},t}, \sigma^2_{\gamma_t})$$
$$\sigma^2_t \sim \text{HalfCauchy}(5)$$

Data generating distribution

$$\mu_{t,\text{skater},i} = \alpha_{t,\text{skater}} + \beta_t \log(w_{\text{ijs},i}) + \gamma_{t,\text{skater}} \log(w_{\text{skater},i})$$
$$\text{points}_{t,\text{skater},i} \sim N(\mu_{t,\text{skater},i}, \sigma^2_t)$$

We add the following predictors to our model:

- $w_{\text{ijs},i}$ is the number of weeks since the beginning of our dataset for the given ith data point, which can be interpreted as the number of weeks since the IJS was introduced. This is to capture the gradual increase of scores over time, estimated by β_t.
- $w_{\text{skater},i}$ is the number of weeks since the first data point for a given skater for the ith data point; this can be interpreted as the length of the skater's career at that time. This is to capture the gradual improvement of a skater over time, estimated for each skater by $\gamma_{t,\text{skater}}$.

We use the log number of weeks because the increases are likely to slow over time and plateau, more like a log function than a linear relationship. $\alpha_{t,\text{skater}}$ models the skater's quality for element or component type t at the beginning of their career after removing the effect of increasing scores since the beginning of IJS.

We also tweak our priors on $\alpha_{\text{group},t}$ for each element type t slightly to be more informative, and list them in Table 4.5. There is no need to put such a large variance on priors when we know a reasonable range for the points earned for each kind of element. The variances are different to reflect our different uncertainties for each element type. We also update the prior on the $\alpha_{\text{group,component}}$ parameter:

$$\alpha_{\text{group,component}} \sim 10 \times \text{Beta}(20, 6)$$

men, ladies	pairs	ice dance
$\alpha_{\text{group,ch}} \sim N(2.0, 0.5)$	$\alpha_{\text{group,st}} \sim N(3.0, 0.5)$	$\alpha_{\text{group,tw}} \sim N(6.0, 1.0)$
$\alpha_{\text{group,st}} \sim N(3.0, 0.5)$	$\alpha_{\text{group,sp}} \sim N(3.5, 0.5)$	$\alpha_{\text{group,st}} \sim N(7.0, 1.0)$
$\alpha_{\text{group,sp}} \sim N(2.5, 0.5)$	$\alpha_{\text{group,tw}} \sim N(5.0, 1.0)$	$\alpha_{\text{group,pd}} \sim N(4.0, 0.5)$
$\alpha_{\text{group,1j}} \sim N(0.67, 0.01)$	$\alpha_{\text{group,th}} \sim N(4.0, 1.0)$	$\alpha_{\text{group,li}} \sim N(4.0, 0.5)$
$\alpha_{\text{group,2j}} \sim N(4.0, 0.5)$	$\alpha_{\text{group,li}} \sim N(6.0, 0.7)$	$\alpha_{\text{group,l2}} \sim N(8.0, 0.5)$
$\alpha_{\text{group,3j}} \sim N(6.0, 0.7)$	$\alpha_{\text{group,ds}} \sim N(4.0, 0.5)$	$\alpha_{\text{group,sp}} \sim N(4.0, 0.5)$
$\alpha_{\text{group,4j}} \sim N(10.5, 1.0)$	$\alpha_{\text{group,ch}} \sim N(2.0, 0.5)$	$\alpha_{\text{group,ch}} \sim N(2.0, 0.5)$
	$\alpha_{\text{group,ju}} \sim N(3.0, 1.0)$	

Table 4.5: The table of priors for HM 2 that specifies $\alpha_{\text{group},t} \sim \text{prior}_t$ for each element type t across different disciplines.

Because components are bounded below by 0.00 and above by 10.00, a scaled Beta distribution will be more accurate in capturing the group mean parameter. These parameters will still give us a mean close to 7.0 which was the mean of the original Normal prior, while leaving the variance large enough to be only weakly informative. As evidenced by the plots in Chapter 3, Figure 3.9, in recent years components scored have shown a skewed distribution with a mean of around 7.0 points.

4.5 Results

We fit models on data up to the 2016-2017 season and then test and report error metrics on the 2017-2018 season up to and including the 2018 Four Continents Championships. For testing simplicity, with competitions later in the 2017-2018 season we do not update our models with data from earlier in the season. Both models assume more knowledge than in a true prediction scenario: for the linear regression we assume the knowledge of start orders, and for the hierarchical models we assume the knowledge of the elements that each skater will perform. Later we will attempt prediction on an actual unknown dataset to better assess predictive power.

We use the following error metrics to evaluate models:

- Score Loss: We predict a skater's total score in a competition, and then add the squared difference from the skater's observed total score to our loss. This is least squares loss.
- Average Score Loss: The total score loss divided by the number of results in the sum, then the square root of that. This can be interpreted as the average absolute distance from the true total score for each skater result.

		men	ladies	pairs	ice dance
Number of Test Data Points		142	137	77	109
OLS	Rank Loss	308	336	58	134
	Average Rank Loss	2.2	2.5	0.8	1.2
	Score Loss	50238	58255	9898	10611
	Average Score Loss	18.8	20.6	11.3	9.9
HM 1	Rank Loss	302	436	82	170
	Average Rank Loss	2.1	3.2	1.1	1.6
	Score Loss	91053	91204	34737	46161
	Average Score Loss	25.3	25.8	21.2	20.6
HM 2	Rank Loss	308	408	78	170
	Average Rank Loss	2.2	3.0	1.0	1.6
	Score Loss	43757	82094	16349	20180
	Average Score Loss	17.6	24.5	14.6	13.6

Table 4.6: Summary of our results for the three predictive models, tested on the 2017-2018 season up to and including the Four Continents Championships. Note that one "result" (one test data point) is the placement of a skater in a competition they enter, which includes their placement (rank) and total score. Generally the linear model is more accurate than either of the hierarchical models, and predictions are more accurate for pairs and ice dance.

- Rank Loss: We predict a skater's placement in a competition, and then add the absolute difference from the skater's observed placement in that competition to our loss.
- Average Rank Loss: The total rank loss divided by the number of results in the sum. This can be interpreted as the average number of placements away from the true placement, per skater result.

Though scores are important, we are ultimately interested in predicting the ranked outcomes of competitions. we can predict the ranked outcomes of competitions. Thus introducing rank loss as a metric makes sense. This can also account for variations in judging across competitions: two panels might score skaters differently but rank their scores exactly the same.

We summarize our results in Figure 4.3, which visualizes the relative average loss metrics for each model. We can see that the linear regression gives the smallest loss the majority of the time, and that HM 1 is the worst of the three models at prediction.

In Table 4.6 we give the detailed error metrics of our models. Out of the 8 hierarchical models attempted (HM 1 and HM 2 for each discipline), only one of them, HM 2 for men, outperforms or matches the linear regression on both error metrics. Overall the linear regression performs remarkably well, particularly for pairs and ice dance where we on average are only about one rank off.

Pairs and ice dance are more reputation-driven. The longevity of a couple is important in competitive success. Being on average 10 points off for ice dance and pairs is a large amount

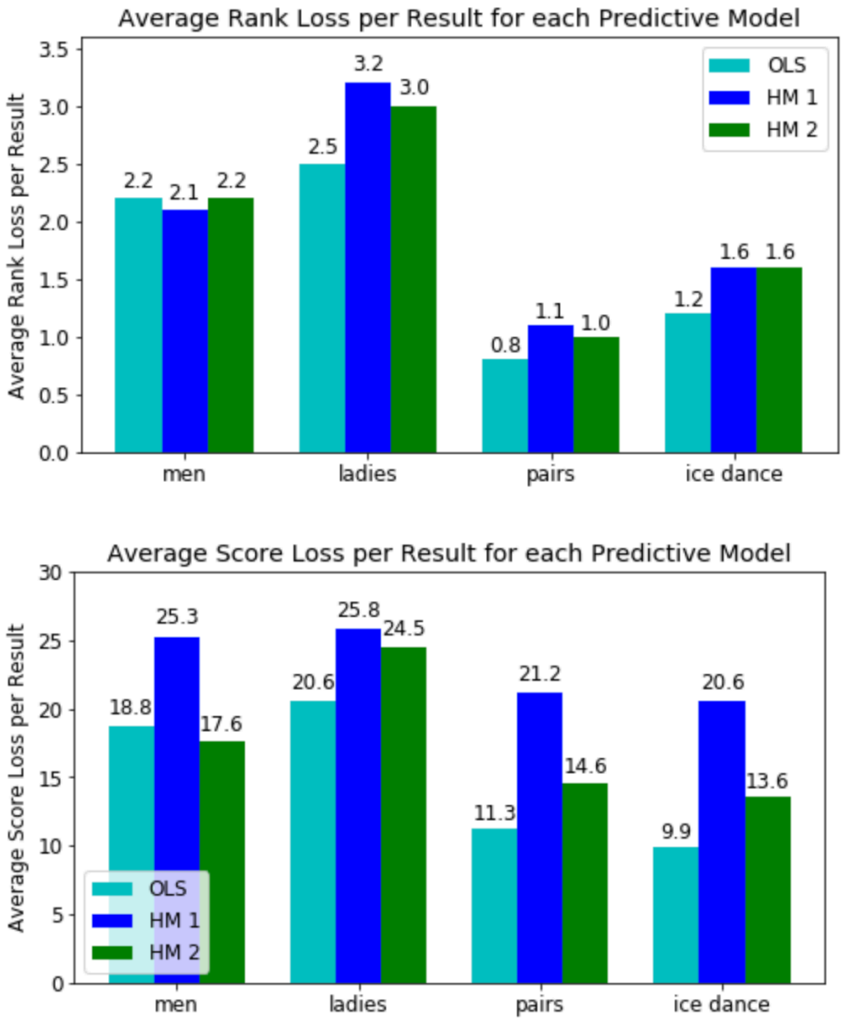

Figure 4.3: Average loss metrics per discipline per predictive model. In general the linear regression outperforms the hierarchical models. HM 2 applied to the men's discipline beats the linear regression. **Top**: Average rank loss gives the average distance from true placement per result. **Bottom**: Average score loss gives our average absolute score error per result.

in the typical total scores and margins of victory seen, but given the linear regression's success with rank prediction, it is likely that it over-predicts or under-predicts for all skaters, while getting their relative placements largely correct.

In the men's and ladies' disciplines the linear regression also performs well, with average rank losses of 2-3 per skater. In general the men's and ladies' disciplines are more volatile because the skaters perform a larger number of risky jump elements to make mistakes on and thus more surprising results. In the men's discipline, predicting a skater's score within 20 points is actually useful because scores for individual skaters range even more than that across competitions.

Where the hierarchical models saw the most success was in the men's discipline. This is where the collective knowledge of different element types was probably the most useful. The top competitors in the men's discipline are currently pushing the envelope forward in terms of technical difficulty, with the best in the world performing five or more quadruple jumps in a single competition. These jumps can really impact a skater's score, so accounting for the element breakdown gave us more success in the men's discipline.

In both men's and ladies' we predict poorly for skaters not previously seen in the dataset. The entire men and ladies' podium from the 2017 Junior World Championships made their senior debuts in the 2017-2018 season, and one would expect those skaters to be much more successful as a new skater than an average or median across all skaters would predict. Because we did not update the hierarchical models or the linear regression's reputation predictors throughout the season for testing purposes, the reported error metrics are for models that do not have complete information in this sense.

4.6 Predicting the 2018 World Championships

We applied the linear regression model and HM 2 to the 2018 World Championships to get a sense of what real-time prediction would be like. Because the linear regression takes start orders as predictors, we had to wait for the start orders to be released before and after the short program. The linear regression thus had a lot of information about how the final result would play out because the free skate start order is seeded based on the short program result. For HM 2, we took the elements that the skater performed at their most recent competition and predicted based on the assumption that the skater would execute those elements, which is actually a reasonable assumption given that "momentum" is considered to be important throughout a season.

Thus we report our results in Table 4.7 with the caveat that the HM 2 prediction was completed entirely before the competition, while the linear regression had to be updated after the

Discipline, # Skaters	Predictive Model	Score Loss	Average Score Loss	Rank Loss	Average Rank Loss
men	OLS	8420	15.1	104	2.8
37	HM 2	188489	71.4	194	5.2
ladies	OLS	16857	21.3	108	2.9
37	HM 2	92784	50.1	168	4.5
pairs	OLS	2007	8.5	62	2.2
28	HM 2	67822	49.2	80	2.9
dance	OLS	739	4.9	38	1.2
31	HM 2	34692	33.5	48	1.5

Table 4.7: The error metrics of linear regression and HM 2 in predicting the 2018 World Championships. The linear regression is more successful than HM 2, but has more information as it predicts while the competition is happening to receive skating order as a predictor. The large losses seen for HM 2 are due to mistakes predicting the skaters that qualify to the free skate; the linear regression receives this information and does not make these mistakes.

short programs. We see the recurring trend that the linear regression is remarkably accurate. In particular, for the ice dance discipline it was only on average 5 points off per skater. The extremely high score losses by HM 2 are due to mistakes in determining free skate qualification: if HM 2 predicted that a skater would not qualify to the free skate, then we predicted that skater's total score as only the short program score. Thus if HM 2 made mistakes on qualification, that would introduce score losses of at least 100^2 for two skaters, both the skater that HM 2 expected to qualify but did not and the one that HM 2 did not expect to qualify but did. This is another example of the additional information the linear regression had available. We also again see the trend of the linear regression being extremely successful at predicting the pairs and ice dance disciplines, which are more reputation-driven.

This world championships was a reminder of how unpredictable skating can be, particularly for the men and ladies. We report the medalists and predicted medalists in Table 4.8. Note that both models predicted Alina Zagitova to win, because she has the highest scoring potential and because she placed well in the short program. However, she fell three times in the free skate and dropped to 5th overall. Both models also predicted Boyang Jin to win the bronze. However, he also had a disastrous free skate falling five times, and dropped to 19th. These were huge shocks, both for the skating community and for our models.

Something worth pointing out in Table 4.8 is that HM 2 correctly predicted $2/3$ of the men, pairs, and dance podium completely prior to the competition's start. Thus we should not discount its predictive power entirely, since in this situation it generated its final predictions before the start of the competition.

	Rank	Outcome	OLS Prediction	HM 2 Prediction
men	1	Nathan Chen	Nathan Chen	Nathan Chen
	2	Shoma Uno	Shoma Uno	Shoma Uno
	3	Mikhail Kolyada	Boyang Jin	Boyang Jin
ladies	1	Kaetlyn Osmond	Alina Zagitova	Alina Zagitova
	2	Wakaba Higuchi	Carolina Kostner	Satoko Miyahara
	3	Satoko Miyahara	Satoko Miyahara	Maria Sotskova
pairs	1	Aljona Savchenko/ Bruno Massot	Aljona Savchenko/ Bruno Massot	Aljona Savchenko/ Bruno Massot
	2	Evgenia Tarasova/ Vladimir Morozov	Evgenia Tarasova/ Vladimir Morozov	Evgenia Tarasova/ Vladimir Morozov
	3	Vanessa James/ Morgan Cipres	Vanessa James/ Morgan Cipres	Xiaoyu Yu/ Hao Zhang
dance	1	Gabriella Papadakis/ Guillaume Cizeron	Gabriella Papadakis/ Guillaume Cizeron	Gabriella Papadakis/ Guillaume Cizeron
	2	Madison Hubbell/ Zachary Donohue	Madison Hubbell/ Zachary Donohue	Madison Chock/ Evan Bates
	3	Kaitlyn Weaver/ Andrew Poje	Kaitlyn Weaver/ Andrew Poje	Kaitlyn Weaver/ Andrew Poje

Table 4.8: The observed ("Outcome") and predicted medalists at the 2018 World Championships. Both models predicted Boyang Jin and Alina Zagitova to be on the podium, but these two skaters had disastrous free skates and dropped out of contention. Despite underperforming in error metrics compared to the linear regression, HM 2 predicted the men, pairs, and dance podiums very successfully.

4.7 Discussion

In this chapter we explored three models of predicting future competitions based on past results. The first and most successful was a linear regression that used reputation and start order predictors. The other two models were hierarchical models that fit distributions on points earned per elements and per component.

Though the linear regression model was more successful in prediction, the hierarchical model provides an interesting framework and more information about skater potential as a whole. It explicitly models improvement over time, while the linear regression is largely stagnant. The hierarchical models also do not require the input information of start order required by the linear regression.

The hierarchical models could be improved by better categorization of elements. In Table 4.3 we described the element groups used for HM 1 and HM 2, but refining further into jump types would be a more sensible model for prediction in the men and ladies discipline.

The linear regression model was particularly successful in ice dance and pairs, suggesting that these two disciplines are both (1) much less volatile and (2) much more based on prior reputation.

Ultimately we suffer most from a lack of predictors outside of historical data. Factors like injury, coaching changes, and equipment issues are influential. However, the lack of English-language coverage on figure skating (the sport is extremely popular in Japan) renders the collection of this data very difficult, as there are few online outlets that do any reporting on skating, let alone consistent reporting on athlete health and other factors.

When predicting a new season, we have no information about the new skaters who have not competed on the senior level before. More nuanced predictions could use junior-level and smaller international competition results to gain more information about these skaters, who often have more scoring potential than our models think when they have no information on them.

Accurate prediction is a difficult task, but this chapter introduced a framework for future work to build off of in attempting to develop more accurate models.

Chapter 5

Judging Bias

As evidenced by Chapter 4, accurate prediction is difficult and simplistic models can outperform more nuanced ones. We thus turn to assessing the judges. As a judged sport, skating has been and will continue to be plagued by controversy. Some even assert that skating does not deserve the label of "sport" due to its reliance on judges [33, 34].

Do judges score skaters from their own country more generously than they score skaters from other countries? has been asked before in the literature. This chapter will attempt to answer this question and quantify its impact on competition results. We apply the following methods to address bias:

- A nonparametric test of the independence of judge nationality and scoring above or below the median
- An ordinary least squares regression to measure the amount of bias introduced by a judge scoring a skater from their own country
- A ridge regression to measure judges' biases for or against specific countries, and to estimate unbiased standings

5.1 Literature Review

In the 6.0 system, judges assigned each skater two scores out of 6.0 evaluating technical merit and presentation. These two scores were summed to calculate a judge's total score for a skater. However, this score was only important in how it placed the skater in a judge's *individual ranking of the skater* [11]. For simplicity, imagine a competition with three judges and two

	Judge A	Judge B	Judge C
Skater 1	10.2	10.3	10.2
Skater 2	10.1	10.4	10.1

Table 5.1: Simplistic sample scoring data under the 6.0 system (note each judge gives a skater a total mark out of 12.0). Because Judge A and Judge C both "prefer" Skater 1 to Skater 2, Skater 1 is ranked ahead of Skater 2.

skaters, as depicted in Table 5.1. Then what matters is that Judge A prefers Skater 1 to Skater 2, Judge B prefers Skater 2 to Skater 1, and Judge C prefers Skater 1 to Skater 2. These are the judges' individual *preference orders*, and the variations in the 6.0 system varied in how they aggregated these preference orders. Ties were not allowed within a judge's preference order. The 6.0 system thus used a *social ranking rule* to aggregate the preferences of individual judges.

Bassett and Persky (1994) proved that the original 6.0 system was the only social ranking rule that satisfied desirable properties of majority rule: in viewing the rating of skating as a subjective matter, the original 6.0 system was relatively fair and limited possible strategic manipulation. They also simulated judges' marks as objective measurements of performances plus noise and showed that the 6.0 system captured the true rankings effectively as well [35].

Wu and Yang (2004) compared variations of the 6.0 system for robustness to manipulation by judges. In 1998 the ISU moved to the one-by-one version of the 6.0 system. In this system, skaters were ranked by the number of pairwise wins when compared to other skaters in judges' rankings. Wu and Yang assessed the one-by-one system, the original 6.0 system, and a new rule that they proposed to be more robust to manipulation [36].

With the introduction of the IJS in 2004, skating scoring abandoned the use of a social ranking rule altogether. Thus the literature on the former judging system as a social ranking rule is less relevant for figure skating's future, but interesting in seeing how the judging system changed from being essentially a voting system to one that assigns a score based on more detailed, arguably objective rules.

Zitzewitz (2006) performed analysis on actual competition scores, a collection of events judged under the 6.0 system around the time of the 2002 Winter Olympics. He found statistically significant evidence of nationalistic judging bias: judges would score skaters from their own country an average 0.166 points higher on a score scale with 12.0 maximum points [37].

Emerson and Arnold (2011) performed one of the first analyses of competitions judged under the new system. They examined data from the 2009 European Championships and the 2010 Olympic Games, both of which occurred during a time period when the ISU randomly picked a sub-panel of the given judges' scores to count towards the final score. At the 2009 European Championships, judges' columns on scorecards were presented in the same ran-

dom order for all skaters in a segment. They applied a variety of statistical tests to show that the results at the 2010 Olympics had columns of the selected judges permuted between scorecards, and raised concerns over the difficulty of performing analysis on such randomized columns [21].

Skating scoring has undergone improvements to remedy some of these issues. The introduction of IJS introduced accountability for scores in providing much more detail into the reasoning behind scores. The practice of choosing a random sub-panel of judges to calculate the final score in the original version of IJS has since been abandoned. The ISU also removed the anonymization of judges for the 2016-2017 season, so we can finally associate scores to judges under the new scoring system.

A recent statistical analysis by journalists at BuzzFeed News investigated national biases of judges selected for the Olympics. The BuzzFeed analysis computed the average number of points higher that individual judges score skaters from their own countries, and showed one competition where 4th and 5th place would have swapped after removing the scores of a nationalistic judge [38]. We take a more holistic approach to look for overall trends, and come to a slightly different conclusion. A later analysis by the same journalists evaluated how nationalistic judges may have changed standings at the 2018 Olympics [39]. We will discuss their results when we perform our own bias evaluation of the Olympics.

It is not clear how to measure "bias" when the only data we have on what skaters performed is in the form of potentially biased judges' scores. Thus each of our methods must attempt to control for "true quality" in some way. Campbell and Galbraith (1996) introduced a nonparametric test of bias that they applied to Olympic figure skating data under the 6.0 system [40]. We will explain this test and apply it to our data. Emerson et al. (2009) introduced a regression model for breaking down judge-country bias applied to diving data that we will modify and apply to our data [41].

5.2 Methods

When assessing nationalistic bias, we must rely on the aggregated scores to infer true skating quality as a neutral benchmark to use in assessing judging bias. Prior to the 2016-2017 season, judges' individual scores were randomized on the final scorecard so there is no way to match scores to judges. Thus we only use the 2016-2017 and 2017-2018 data in this chapter.

Recall that skaters perform technical elements and receive a grade of execution (GOE) for each element in a program. The judges also score each skater on the five categories of program components, which we will refer to as "components." Thus we have (1) GOEs and (2) component scores for individual judges and performances. GOEs are integers between -3 and

3 inclusive, and component scores are marked on increments of 0.25 between 0.00 and 10.00 inclusive. Thus we should treat these two separately because their distributions clearly differ. Components tend to exhibit less variation across performances for a single skaters.

We perform a nonparametric test and apply linear regressions in this chapter to investigate judging bias by nationality. Our data is still contained in `pandas` dataframes [19] and we use Python 2.7 to evaluate the nonparametric tests. For the vanilla linear regressions we use the `statsmodels` package [28]. We use the `sklearn` implementation of ridge regression [42].

5.3 Nonparametric Test

GOEs and component scores are difficult to model due to their discrete nature. The domain of GOEs consists of only 7 values, so working with the most common continuous models for inference is very difficult. Component scores have a wider range in being able to take on $4 \times 10 = 40$ different values, but suffer from a different problem of skewness: there are very few scores under 5.00, and the top skaters earn multiple 10.00's.

Campbell and Galbraith (1996) introduced a *nonparametric test of nationalistic judging bias*. They applied this test to Olympic data under the 6.0 system, and the method is compelling because it does not assume a distribution on the judges' scores [40]. Readers should refer to Campbell and Dufour (1995) and Campbell and Calbraith (1996) for correctness proofs [40, 43].

We generalize the description of Campbell and Galbraith's method to work for IJS as well. Consider the scheme of one competition segment where there are J judges. Each judge will score all n quantities in that segment. In the 6.0 system, these quantities are technical merit and artistry for each of the skaters in the segment. Under IJS, these quantities are elements and component types for each skater. Let $\vec{a}_j$ be the n-dimensional vector of judge j's scores for all n quantities to be scored. Let $\vec{\alpha}$ be the median of $\vec{a}_1, \ldots, \vec{a}_J$ such that α_i is the median of $a_{1i}, \ldots, a_{Ji}$. Then construct $\vec{v}_j = \vec{a}_j - \vec{\alpha}$ for each judge j. This is the vector of differences between judge j's scores and the median of all the judges' scores.

Now define $\vec{c}_j$ to be a vector of -1's and 1's. c_{ji} is 1 if the skater associated with the ith quantity represents the same country as judge j and -1 if the countries do not match. Then under the null hypothesis of independence of judge scoring and nationality, c_{ji} is independent of v_{ji}. The test statistic is as follows:

$$S = \sum_{i,j} I(v_{ji} \neq 0) \cdot I(v_{ji} \cdot c_{ji} > 0),$$

Let $N = \sum_{i,j} I(v_{ji} \neq 0)$. Under the null hypothesis:

$$S \sim \text{Binomial}(N, 0.5)$$

We throw out all observations where a judge scores exactly the median of all the scores of the quantity i. $I(v_{ji} \cdot c_{ji} > 0)$ is an indicator of whether the signs of v_{ji} and c_{ji} match. Under the null v_{ji} and c_{ji} are independent each with median 0, so $I(v_{ji} \cdot c_{ji} > 0) \sim Bernoulli(0.5)$. Campbell and Dufour (1995) prove this result holds even when we remove the instances where $v_{ji} = 0$, which is less obvious [43]. The null distribution holds under very general assumptions, allowing us to assess the results across competitions and segments as well as GOE and components scores in a single test.

Thus if we want to test whether judge-skater nationality matching and scoring are independent, we can compute the test statistic for the dataset in question and see if it lies in a range of values that is very unlikely under the null Binomial distribution. The departure we would expect from the null model would be higher values of S because that means a judge scores a skater from their own country higher and skaters from other countries lower. Thus if we compute the test statistic to be s, we will consider the upper extreme of the Binomial and our p-value will be $P_{H_0}(S \geq s)$

Across all element and component marks of the 2016-2017 season and the 2017-2018 season up to Four Continents 2018, we initially have 247278 total data points, but 138831 of these are cases where $v_{ji} = 0$, leaving us with 108847 cases where $v_{ji} \neq 0$.

We report some of our test statistics in Table 5.2. We also report $B = S/N$ ("percent bias") as a indication of whether our suspicion that we will see higher values of S is founded; recall we add 1 to S if either (1) a judge scores a skater from their own country higher than the median or (2) a judge scores a skater not from their own country lower than the median. It is not a formal measurement of how much bias exists, but can be used to gain a rough sense of which way the correlation leans.

We break down our data by discipline, segment, and GOEs vs. components. In every single case, we have a p-value that is essentially 0, rejecting the null hypothesis that judge scoring is independent of judge-skater nationality matching using any selective inference procedure. In all rows our observed B is greater than 0.5, suggesting the reasonable alternative that if a judge has the same nationality as a skater, they are more likely to score that skater higher.

However, all values of B are not too far from 0.5. This suggests strong evidence for a small amount of bias. Ice dance notably has the highest B values.

Data Slice	p-value	S	N	B
all	10^{-16}	61106	108447	0.56
short GOEs	10^{-16}	11069	19841	0.56
short components	10^{-16}	17744	31391	0.57
free GOEs	10^{-16}	17140	30704	0.56
free components	10^{-16}	15153	26511	0.57
men	10^{-16}	17958	32992	0.54
ladies	10^{-16}	17497	31510	0.56
pairs	10^{-16}	11794	20807	0.57
ice dance	10^{-16}	13857	23138	0.60

Table 5.2: Results for the nonparametric test of whether the judge-skater country matching is independent of scoring. $B \triangleq S/N$ gives a crude indication of whether judges are being more nationalistic than not. We report test statistics on slices of our data, broken down by GOEs, components, segments, and discipline.

5.3.1 Assessing the 2018 Olympics

We would like to answer more refined questions about our data. Most of the B values reported in Table 5.2 were very close to 0.5. Perhaps in a single competition there is more or less "bias" than in the aggregated data.

Because the Olympics are always a source of intrigue, we apply the nonparametric test to the 2018 Olympic data sliced in a variety of different ways. Table 5.3 shows all of the different slices tested. We split competitors into the top 6 or bottom $1/3$ of final placements. We split by program, as well as GOEs vs. component scores. With so many hypothesis tests, we need to perform some sort of selective inference. To test at a total significance level of $\alpha = 0.05$, we can apply Bonferroni's procedure and test each hypothesis at the level $\alpha/50 = 0.001$. In total there are fewer than 50 hypothesis tests, but we choose 50 so our threshold is a round number to compare to.

We still have extremely low p-values and not too extreme B observations. However, in men's, ladies, and pairs we have some cases not reported as significant, particularly with the bottom $1/3$ of competitors where bias would not make a difference in medal standings. The men's short program does not see to have been affected by nationalistic judging, nor the ladies' or pairs' frees.

In ice dance there are higher B and lower p-values for all of the slices, even those that only contained the bottom $1/3$ of competitors. This again suggests that nationality and scoring are more positively correlated in ice dance. Possible explanations include:

- It is easier for a judge to give a biased score in ice dance because there are fewer obvious mistakes.
- Judges have stronger preferences for particular ice dance styles from their home coun-

Discipline	Description	p-value	B	S	N
men	everything	10^{-8} (*)	0.55	1645	2989
	top 6 competitors	10^{-12} (*)	0.63	437	692
	bottom 1/3 competitors	0.13	0.52	379	729
	short	0.17	0.51	772	1508
	short GOEs	0.53	0.50	316	635
	short components	0.09	0.52	456	873
	free	10^{-12} (*)	0.59	873	1481
	free GOEs	10^{-6} (*)	0.57	490	853
	free components	10^{-8} (*)	0.61	383	628
ladies	everything	10^{-12} (*)	0.57	1583	2802
	top 6 competitors	10^{-6} (*)	0.59	367	620
	bottom 1/3 competitors	0.12	0.52	363	696
	short	10^{-10} (*)	0.58	830	1429
	short GOEs	10^{-6} (*)	0.59	328	554
	short components	10^{-6} (*)	0.57	502	875
	free	0.0001 (*)	0.55	753	1373
	free GOEs	0.017	0.54	389	722
	free components	0.0011	0.56	364	651
pairs	everything	10^{-10}	0.57	1224	2161
	top 6 competitors	10^{-5} (*)	0.58	368	637
	bottom 1/3 competitors	0.067	0.53	291	548
	short	10^{-9} (*)	0.59	655	1119
	short GOEs	10^{-6} (*)	0.61	280	463
	short components	0.0001 (*)	0.57	375	656
	free	0.0013	0.55	569	1042
	free GOEs	0.054	0.53	313	588
	free components	0.0028	0.56	256	454
dance	everything	10^{-16} (*)	0.61	1291	2126
	top 6 competitors	10^{-13} (*)	0.66	316	478
	bottom 1/3 competitors	10^{-6} (*)	0.59	357	609
	short	10^{-9} (*)	0.59	605	1024
	short GOEs	10^{-6} (*)	0.62	224	360
	short components	10^{-5} (*)	0.57	381	664
	free	10^{-16} (*)	0.62	686	1102
	free GOEs	10^{-9} (*)	0.61	366	596
	free components	10^{-10} (*)	0.63	320	506

Table 5.3: Results for the nonparametric test of whether the judge-skater country matching is independent of scoring, at the 2018 Olympics. We test each slice at the 0.001 level, and (*) marks significance at this level. We consistently see lack of significance in the bottom 1/3 of competitors. Ice dance reports many more significant results than the other three disciplines, suggesting further investigation. For p-values less than 10^{-4}, we report only the degree of magnitude.

try.

- In ice dance, it is harder to break out of the status quo and move up in the world ranking, so judges want to do more to boost their home teams.

5.4 Ordinary Least Squares: Judge and Skater Effects

The nonparametric tests showed strong evidence of a small correlation between judge nationality and scoring. However, we could not measure the amount of bias (as shown by our attempts to interpret B values), nor pinpoint where the bias was coming from (particular judges or countries). We thus move on to parametric techniques to answer these questions.

The immediate parallel to the nonparametric test is to regress a judge's score on a predictor of a judge being from the same country as a skater. However, clearly some skaters simply perform better elements and have higher quality than others. Thus we want to control for skater quality.

Another issue with the nonparametric test is that some judges may score all skaters higher than average, and perhaps more generous judges are more likely to come from a particular country. Thus also want to control for the fixed effect of judge leniency.

With the regression, we also run into the problem of our GOEs and components having dramatically different distributions. Along with this difference in domains, components scores are typically more closely tied to a skater than element GOEs. Any skater can earn a -3 GOE on an element for falling, but not every skater can earn 9.50 in components. Thus we must perform separate regressions for GOEs and components.

We use the following ordinary least squares regression:

$$s_{i,j,k} = \alpha_k + \mu_j + \beta \cdot I(C(j) = C(k)) + \varepsilon_{i,j,k},$$

where $s_{i,j,k}$ is the score by judge j for the kth skater's ith GOE or component (but recall we fit the model separately for each). $C(x)$ indicates the country of person x, so $I(C(j) = C(k))$ is an indicator of judge j and skater k's nationalities matching. β will attempt to measure the amount of bias that judges hold for skaters of their own nationality. α_k will measure the skater fixed effect, attempting to measure the points that skater k "deserves." μ_j will measure the relative leniency of judge j. $\varepsilon_{i,j,k}$ will be independent Normal noise that will absorb both skater performance variations and small judging variations.

Table 5.4 reports the estimate coefficients and their p-values, as well as the R^2 for the model. β in both regressions is clearly significant, but the amount of bias appears to be small. Recall GOEs are integers, so $\hat{\beta} = 0.256$ only shows a slight tendency to score higher. Components

Score Type	$\hat{\beta}$	p-value	R^2
GOEs	0.256	$< 10^{-20}$	0.24
components	0.277	$< 10^{-20}$	0.87

Table 5.4: Reported results of the linear regression accounting for skater effects, judge effects, and same-country bias. The coefficient estimates suggest that a judge will tend to score a skater from their own country $\hat{\beta}$ points higher than other judges not from the country would score that same skater.

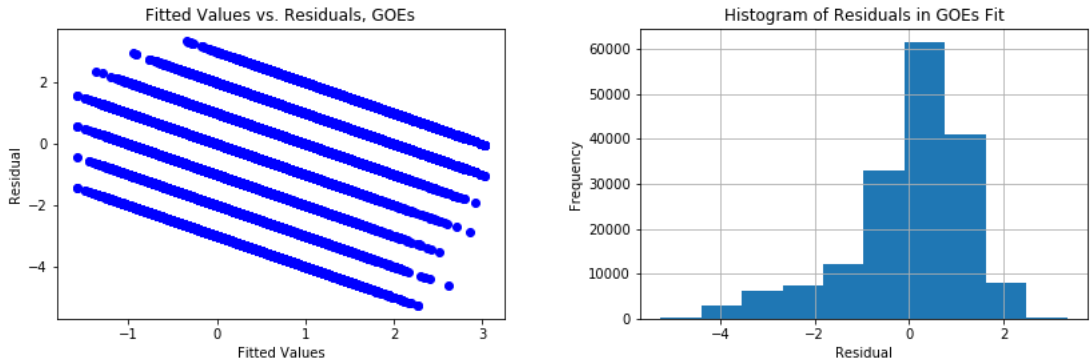

Figure 5.1: Diagnostic plots for the linear regression assessing bias in GOEs. Both plots suggest problems with linear regression assumptions. **Left**: Fitted values vs. residuals. There is a clear downward trend, showing a bias in our fit. **Right**: Histogram of residuals. There is clear non-Normality because the distribution is skewed.

are delimited at levels of 0.25, so the result is slightly more interpretable in that a judge will score a skater from the same country on average 0.277 higher than the mark the skater "deserves."

Immediately the R^2 for the GOEs fit is concerning. Is concluding the existence of a small nationalistic bias in GOEs really an appropriate conclusion if we are not sufficiently explaining the data? Some diagnostic plots in Figure 5.1 further this concern. First, the fitted values vs. residuals plot shows a bias in our fit. The straight lines arise from the discrete nature of our response variable, the GOEs. There is a clear downward trend, with larger fitted values resulting in smaller residuals.

The residual histogram also shows a clear lack of normality, with an obvious left skew. Ultimately this simplistic model that controls for skater effect does not work for GOEs due to their limited range and discrete nature, as well as the drastic variation across GOEs for an individual skater. This is reflected in the R^2: just knowing a skater won't tell us how well they're going to perform their elements. Even the best skaters will fall and receive -3 GOEs sometimes. For GOEs, then, it makes more sense to use the nonparametric test because it makes fewer assumptions and does not attempt to measure the skater effect.

For the components fit, we removed all data points with scores less than or equal to 2.00.

Skater	Judge	Scores for this Component	Segment
Marissa CASTELLI / Mervin TRAN	Christiane MOERTH	7.75, 6.50, 6.75, 7.00, 7.25, 7.00, **0.25**, 7.50, 7.50	GP of France 2016, pairs short
Chafik BESSEGHIER	Matjaz KRUSEC	7.00, 6.75, 7.50, 7.75, 7.25, 7.50, **0.25**, 7.25, 7.25	Europeans 2017, men's free
Kaitlyn WEAVER / Andrew POJE	Jean SENFT	9.75, 9.50, 9.00, 9.25, 9.50, 8.50, 9.50, **0.25**, 8.75	GP France 2017, ice dance short
Camille RUEST / Andrew WOLFE	Doug WILLIAMS	7.00, 7.00, 6.75, **0.50**, 5.50, 6.75, 6.25, 7.00, 7.00	Four Continents 2018, pairs free

Table 5.5: Removed outliers for all components linear regressions. The outlier score is bolded and underlined, and clearly these data points were scoring or reporting mistakes and not intentionally low scores.

These data points are reported in Table 5.5 for completeness. We bolded the score below 2.00 in the list of all the judges' scores for that component mark. Clearly in the context of the other scores (all above 5.00), these individual data points are score entering or reporting mistakes and not intentional.

The components fit is more promising because of a much higher R^2 value as well as a more interpretable coefficient on the scale of scores. Because the scores are on increments of 0.25, $\hat{\beta} = 0.277$ says that on average a judge from the same country as the skater will score that skater more than one increment higher than judges not from the same country will score the skater. The diagnostic plots in Figure 5.2 look significantly better. There is no visual evidence of heteroskedasticity or bias in the fitted values vs. residuals plot, and the relationship looks random. This residual histogram provides evidence that the residuals are Normal with small variance.

5.5 Ridge Regression: Specific Bias Effects

Given evidence that there is a small but significant nationalistic bias, we would like to dissect it further to see where these biases arise.

Emerson et al. (2009) gave a framework for assessing nationalistic judging bias in diving scores that we apply to our setting:

$$s_{i,j} = \lambda_i + \mu_j + \beta_{j,C(i)} + \varepsilon_{i,j}$$

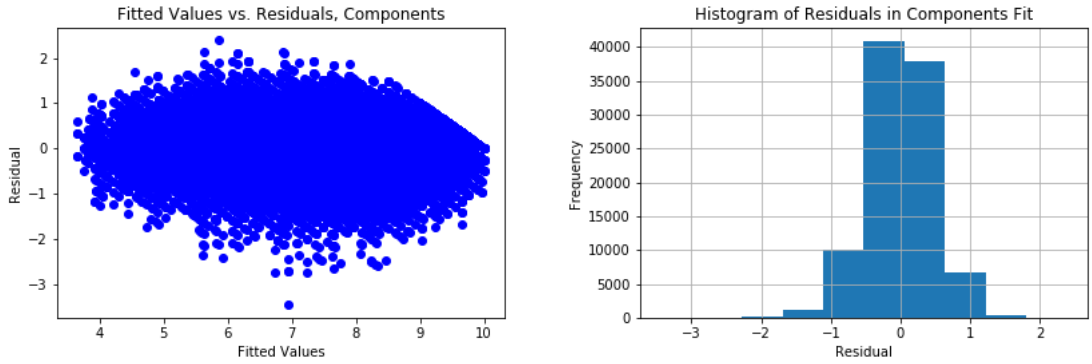

Figure 5.2: Diagnostic plots for the linear regression assessing bias in components. **Left**: Fitted values vs. residuals plot. There is no evidence of heteroskedasticity or bias. **Right**: Histogram of residuals, which looks sufficiently Normal.

Translating this to the context of skating, $s_{i,j}$ is the score of the jth judge for the ith component to be scored. λ_i should be interpreted as the "true quality" of the ith quantity. μ_j is the judge effect for judge j and captures the tendency of a judge to score higher or lower on average. $\beta_{j,C(i)}$ is an interaction term between judge j and the country of the skater who performed quantity i. $\varepsilon_{i,j}$ is a Normally distributed error term which is assumed to be independently identically distributed for all data points. Emerson et al. (2009) applied this to an Olympic diving competition and then go on to assess what the "true results" should have been based on the estimated λ_i [41].

The addition of the interaction term and the quality term offer a more detailed look into exactly what biases are cropping up. We have no metric for exactly what the "true quality" of a quantity is, so the best we can do is use all of the judges' scores to estimate it.

We experiment with this model applied to a single event: an obvious choice is the 2018 Olympics. Because the panels for each discipline at the Olympics are mostly disjoint, we will fit a separate model per discipline. We will also continue to focus on component scores.

A major issue that arises with this model applied to a single competition is collinearity. Within a single discipline at a single competition, there can be at most 3 skaters from the same country. Across both segments, each skater will be scored on components a total of 10 times. The predictors used to estimate λ_i, the quality of one component type of one skater, are indicators. The predictors used for the judge-skater country interaction are indicators per judge, per country. The indicators for a particular skater's components will be strongly correlated with the judge indicators for their country. If a skater is the only competitor from their country, then the sum of their quality indicators will be exactly equal to the sum of their judge-country indicators.

Discipline	Training Data Points	Testing Data Points	Train R^2	Test R^2
men	2152	278	0.94	0.92
ladies	2153	277	0.96	0.94
pairs	1528	182	0.94	0.95
ice dance	1753	227	0.97	0.95

Table 5.6: Summary of training and testing data for ridge regression on the 2018 Olympic individual judges' scores. R^2 does not drop significantly for the test dataset, suggesting that we predict relatively well.

For instance, when we apply this model to the 2018 Olympics ice dance competition, for every judge j on the panel we get $\hat{\beta}_{j,\text{USA}} > 0$. It is unreasonable to say that *all* the judges are biased positively towards the American ice dancers. In reality, American ice dancers all have very high quality components, which is getting mistakenly absorbed into the bias coefficients. Since we have no other data than "biased" judge scores to estimate λ_i, we must assume that $\beta_{j,c}$ are roughly centered around 0 for any fixed c.

To ameliorate this effect, we explicitly code in the country term and use ridge regression:

$$s_{i,j} = \lambda_i + \mu_j + \delta_{C(i)} + \beta_{j,C(i)} + \varepsilon_{i,j}$$

Now the country effect is deliberately encoded in $\delta_{C(i)}$. Because ridge regression penalizes the magnitude of the coefficient vector, it will prefer to put weight in δ_c rather than $\beta_{j,c}$ because there are at least 9 times as many $\beta_{j,c}$ terms (one for each judge) as there are δ_c terms for each country c.

Though in using ridge regression we lose the ability to get accurate p-values, we can get a better estimate of "true quality" with $\lambda_i + \delta_{C(i)}$, which removes the judge effects μ_j and the bias terms $\beta_{j,C(i)}$. To avoid overfitting, we split our data into a training set and a test set, and ensure that the test set contains at least one data point for (1) each skater and (2) each judge-skater country interaction term (and as a result, each judge and each country). We report the number of observations in the training and testing datasets as well as R^2 on each set in Table 5.6. The differences between training and testing R^2 are small and do not reflect a lack of fit.

We then determine the scores based on the estimated "true qualities" of components $\lambda_i + \delta_{C(i)}$. Assuming the technical score stayed the same, Table 5.7 shows the comparison of "true" point totals (with judge effects μ_j and judge-country biases $\beta_{j,C(i)}$ removed) and observed point totals for the top 20 ice dance teams at the Olympics, and the ranks that would be associated with each. Note that only two placements change: 16th and 17th place swap, and the point difference was originally less than 1 point, which is a very small margin in skating. That margin could arise from judging variance or observational error.

The top 10 teams in this event only come from 5 different countries: France, Canada, Rus-

Observed Rank	"True" Rank	Observed Score	"True" Score	Skater
1	1	206.07	204.38	Tessa VIRTUE / Scott MOIR
2	2	205.28	203.34	Gabriella PAPADAKIS / Guillaume CIZERON
3	3	192.59	191.59	Maia SHIBUTANI / Alex SHIBUTANI
4	4	187.69	187.76	Madison HUBBELL / Zachary DONOHUE
5	5	186.92	185.67	Ekaterina BOBROVA / Dmitri SOLOVIEV
6	6	185.91	185.16	Anna CAPPELLINI / Luca LANOTTE
7	7	181.99	181.12	Kaitlyn WEAVER / Andrew POJE
8	8	176.91	176.46	Piper GILLES / Paul POIRIER
9	9	175.58	175.10	Madison CHOCK / Evan BATES
10	10	173.47	173.24	Charlene GUIGNARD / Marco FABBRI
11	11	170.32	169.41	Penny COOMES / Nicholas BUCKLAND
12	12	168.33	167.97	Sara HURTADO / Kirill KHALIAVIN
13	13	162.24	162.20	Tiffani ZAGORSKI / Jonathan GUERREIRO
14	14	161.35	161.22	Natalia KALISZEK / Maksym SPODYRIEV
15	15	160.63	160.32	Kana MURAMOTO / Chris REED
17	**16**	149.59	151.17	Marie-Jade LAURIAULT / Romain LE GAC
16	**17**	150.49	150.37	Kavita LORENZ / Joti POLIZOAKIS
18	18	147.74	148.47	Yura MIN / Alexander GAMELIN
19	19	147.18	147.24	Alisa AGAFONOVA / Alper UCAR
20	20	142.57	143.33	Lucie MYSLIVECKOVA / Lukas CSOLLEY

Table 5.7: Comparison of observed ranks and score outcomes as compared to those estimated by ridge regression based on the "true quality" of components $\lambda_i + \delta_{C(i)}$ in the ice dance competition of the 2018 Olympics. Note that only one placement changes: 16 and 17 swap (bolded and underlined). We only include the top 20 teams. This assumes that technical score stays the same.

	FRA	CAN	RUS	USA	ITA
Christine HURTH (France)	-0.04	-0.15	0.02	0.14	0.08
Leanna CARON (Canada)	-0.05	**0.52**	-0.11	0.20	0.07
Maira ABASOVA (Russia)	-0.13	-0.05	0.07	-0.17	-0.18
Sharon ROGERS (USA)	0.02	-0.02	0.10	**0.33**	0.17
Walter ZUCCARO (Italy)	0.20	-0.06	-0.13	0.05	0.09

Table 5.8: Estimated bias terms $\hat{\beta}_{j,c}$ for judges and countries of the top 10 ice dance teams at the 2018 Olympics. Columns are biases for/against skater countries. Rows are biases of particular judges. Underlined entries are where judge and skater countries match. Bolded entries are where the coefficient is larger than 0.25, enough to warrant looking at. These results are consistent with favoring skaters of the same country, and shows very small biases for other country pairs.

Discipline	Skater	Country	Original Rank	"True" Rank	Observed Score	"True" Score
men	Nathan CHEN	USA	5	4	297.35	296.89
	Boyang JIN	CHN	4	5	297.77	296.40
	Keiji TANAKA	JPN	18	16	244.83	245.24
	Michal BREZINA	CZE	16	17	246.07	244.42
	Misha GE	UZB	17	18	244.94	243.30
ladies	Loena HENDRICKX	BEL	16	15	171.88	171.91
	Gabrielle DALEMAN	CAN	15	16	172.46	171.75

Table 5.9: Placements that would change at the 2018 Olympic games in the men, ladies, and pairs event with judge-country biases removed for skaters who qualified to the free skate. Note that in all of these cases, the skaters were originally separated by less than 1.5 points. There were no swaps in the pairs event.

sia, USA, and Italy. Each of these countries also had one judge represented on the panels, so we look at the interaction terms for each of these judges and countries in Table 5.8. The estimated biases are in general very small in magnitude, which explain why the standings in Table 5.7 do not change for the top 10. The underlined entries in Table 5.8 indicate the biases when the judge and skater countries match. All but one of these entries indicates a small positive bias, which is consistent with the results of the nonparametric test of bias. The two bolded entries are the ones that are above 0.25, the scale of precision for component scores; the Canadian judge and the American judge show stronger nationalistic preferences for skaters from their own country.

Using the same ridge regression methodology on the other three disciplines, the only movements in skaters that qualified to the free skate are those in Table 5.9. Free skate qualification is top 24 for the men and ladies and top 16 for the pairs after the short program. None of the changes occurred in the medals, and all occurred when the original scores were less than 1.5 points apart originally. No swaps occurred in the pairs discipline. The most high-stakes swap occurred between Boyang Jin and Nathan Chen, originally in 4th and 5th places respectively.

Journalists at BuzzFeed News also looked for standings swaps at the 2018 Olympics. They performed an analysis considering what would have happened if the scores of certain judges were thrown out, and claimed that two swaps in the standings would have occurred. They claimed that if the French and Canadian ice dance judges had their scores replaced or thrown out, then the French team would have won the gold instead of the Canadian team [39]. Our bias removal did not change the placements of the top two teams as shown in the "true standings" we reported in Table 5.7. An important distinction to draw between methodologies is that we remove *all* estimated biases, not just those of particular judges. One caveat in our approach is that we do not recalculate the technical score. The general belief in the skating community is that the Canadians did deserve the gold, because the French team had a costume malfunction in the short program that prevented them from performing their best.

Our results and BuzzFeed's results actually match in the placement swap of Nathan Chen and Boyang Jin. BuzzFeed calls out Chinese judge Weiguang Chen for overscoring Boyang Jin, and using the same replacement technique, claim that Nathan Chen would have placed 4th ahead of Boyang Jin [39]. We see the same swap in our bias-removed results, and the estimated nationalistic bias term for the judge in question is remarkably high:

$$\hat{\beta}_{\text{Weiguang Chen, CHN}} = 0.67.$$

However, even in this situation the two skaters' "true" scores were less than half a point apart, which is still potentially attributable to small observational variation.

5.6 Discussion

We applied three methods to determine the existence of nationalistic bias in figure skating scores in the 2016-2017 and 2017-2018 seasons. All three methods showed strong evidence for a small amount of bias.

The nonparametric tests strongly reject the lack of correlation between judge and skater nationality matching and scoring. We applied this test to slices of the 2018 Olympic scores to see if there were biases in particular aspects of the competition. Across the men's, ladies, and pairs disciplines, we failed to reject the lack of correlation in data slices that only contained the scores for the bottom $1/3$ of competitors. We rejected all of the null hypotheses in the ice dance discipline, suggesting that the most bias exists in ice dance.

We then applied a linear regression controlling for skater and judge effects to test the significance of the predictor of whether a judge and a skater were from the same country. We found this technique insufficient for grades of execution. Judges on average scored skaters

more than 0.25 higher on component marks if they were from the same country.

We then applied a ridge regression to estimate per-country biases in components for each judge. At the 2018 Olympics, only a few placements would have changed and none of the podiums would have changed. These placement changes were all originally very close in scoring, and because judging is imperfect, could have been attributed to observational error and not bias.

These results do not indicate a major flaw in the judging system. No more than two judges of the same nationality ever sit on the same judging panel. Even in our examples where placements did swap, the original point margin was small enough that minute observational errors could have made a difference. Judges can only be so accurate, and it is impossible to distinguish between close performances every time.

These results along with anecdotal evidence suggest that in the majority of cases, medal standings are decided by far more points than nationalistic bias would sway. The judging system puts multiple judges on a panel in order to control for effects such as nationalism. Upsets in figure skating require more than one judge. One of the biggest scandals after the introduction of IJS was in 2014, when Russia's Adelina Sotnikova beat out reigning Olympic champion Yuna Kim for the gold in the 2014 Olympics by a margin of more than 5 points [44]. That upset victory required the entire judging panel to grant Sotnikova the victory, not just the two judges called into question by the media (one was pictured hugging Sotnikova after the win). Due to the anonymity of scoring in 2014, it is difficult to assess this result statistically.

Chapter 6

Conclusions

This thesis collected and parsed all figure skating scores during the period 2005-2018 into a format amenable to analysis for the first time, explored models for prediction using this historical data, and analyzed judging bias based on nationalism.

6.1 Data

The data collection and parsing methods transformed PDF scoresheets into formats conducive to data analysis. The parsing methods should prove relatively robust barring any major changes to the scoring system or scoresheet formats. There are a variety of other trends beyond prediction and bias that can now be explored in these 10+ years of scoring data.

6.2 Prediction

We explored two types of predictive models: a simple linear regression based on a skater's best historical score and start order, and hierarchical models that fit expected point values for element and component types. The linear model performed remarkably well in comparison to the more complicated models, though the hierarchical models provide more information about the details of a skater's performance.

There are many improvements that can be made to these models. They both assume a lot of information: the linear model requires starting orders, which are unknown until a time close to the competition, and the hierarchical models assume a knowledge of what elements a skater will execute, though we showed with our Worlds prediction that we could guess based on the last program a skater performed.

The linear regression model outperformed the hierarchical models in our experiments, but the linear regression model cannot give a prediction until the day before competition due to requiring starting order as a predictor.

The hierarchical models for the single's disciplines would probably be improved by a further dissection of element types. We currently fit models for single, double, triple, and quadruple jumps, but would probably be better served by splitting based on different types of jumps (axels, lutzes, etc.) and accounting for the larger point worth of combination jumps.

The other main issue with our predictive models is their relative lack of predictors. We aimed to do the best prediction we could with historical data. However, inevitably other factors like injury, illness, coaching changes, and the like would come into play.

Another change we could make to the hierarchical models is to change the variance term to be per-skater. Some skaters are much more consistent than others on particular jumps. In theory we could take our hierarchical models and generate probabilities of one skater placing in front of another.

The success of the linear regression model also suggests that a simple model may be the best to predict figure skating, since it an unpredictable sport. One potential direction for future work would be to develop some sort of simple world ranking system that represents current scoring potentials of individual skaters. If this ranking could be easily updated after each competition, then we would have a much cleaner methodology for maintaining our current belief of figure skating scoring potential. Consider, for example, the Elo rating system for chess [45]. Unfortunately many sports rankings system do not carry over easily to figure skating, as most sports see observations of pairwise wins, losses, or ties.

6.3 Judging Bias

We also took advantage of the detailed scoring data in order to address the question of nationalistic judging bias. We found strong evidence of a small nationalistic bias using a nonparametric test and two regression tests.

An obvious next step would be to apply the ridge regression model to larger sets of data to evaluate specific judge-country biases over time and look for other instances where rankings would change.

We only scratched the surface of looking at judging bias. There are many other questions we could ask which were beyond the scope of the main question this thesis set out to answer:

- After the removal of anonymous judging in the 2016-2017 season, do we see any changes in scoring of skaters of particular nationalities?

- Are there particular judges showing more or less nationalistic bias?
- What is the average degree of consensus among judges? Does this imply that the scoring of elements and components is subjective or objective in nature?
- Are there biases in validating certain technical elements, such as underrotations on jumps for skaters?

The ISU will make changes to the scoring system now that the Olympic cycle has just ended. One change that has been approved is to increase the possible GOEs to range from -5 to +5, giving 11 possible values instead of the current 7 [46]. Will this increase the effect that a single judge has on the score? Will it decrease the degree of consensus among judges? Will it increase the amount of nationalistic bias we see in the scores?

6.4 Implications

We have rich data on skater performances of the past decade. The hope is to inspire further analyses of figure skating to better inform athletes, judges, officials, fans, and sports statisticians.

As a whole, there is little literature on judged sports such as gymnastics, diving, and halfpipe snowboarding, just to name a few more. They pose a challenge different from sports where victories are determined by pairwise matches.

The prediction techniques presented in this thesis can also serve as a framework for other judged sports. Gymnastics and diving also break down their scores into difficulty and execution, and different athletes have different scoring potential. They also have separate disciplines for men and women, and diving has synchronized (pair) events. A multilevel modeling approach grouping by athletes could be attempted for these judging systems.

Gymnastics, diving, and other scored sports also suffer from a similar lack of information of "true quality" when trying to answer questions about scoring trends like judging bias and score inflation. Analyses showing the existence or lack of judging bias can help grant judged sports more credibility or highlight the need for improvement.

Appendix A

List of URLs for All Score Data

2005-2006

- http://www.isuresults.com/results/gpusa05/
- http://www.isuresults.com/results/gpcan05/
- http://www.isuresults.com/results/gpchn05/
- http://www.isuresults.com/results/gpfra05/
- http://www.isuresults.com/results/gprus05/
- http://www.isuresults.com/results/gpjpn05/
- http://www.isuresults.com/results/gpf0506/
- http://www.isuresults.com/results/ec2006/
- http://www.isuresults.com/results/fc2006/
- http://www.isuresults.com/results/owg2006/
- http://www.isuresults.com/results/wc2006/

2006-2007

- http://www.isuresults.com/results/gpusa06/
- http://www.isuresults.com/results/gpcan06/
- http://www.isuresults.com/results/gpchn06/
- http://www.isuresults.com/results/gpfra06/
- http://www.isuresults.com/results/gprus06/
- http://www.isuresults.com/results/gpjpn06/
- http://www.isuresults.com/results/gpf0607/
- http://www.isuresults.com/results/ec2007/
- http://www.isuresults.com/results/fc2007/
- http://www.isuresults.com/results/wc2007/

2007-2008

- http://www.isuresults.com/results/gpusa07/
- http://www.isuresults.com/results/gpcan07/
- http://www.isuresults.com/results/gpchn07/
- http://www.isuresults.com/results/gpfra07/
- http://www.isuresults.com/results/gprus07/
- http://www.isuresults.com/results/gpjpn07/
- http://www.isuresults.com/results/gpf0708/
- http://www.isuresults.com/results/ec2008/
- http://www.isuresults.com/results/fc2008/
- http://www.isuresults.com/results/wc2008/

2008-2009

- http://www.isuresults.com/results/gpusa08/
- http://www.isuresults.com/results/gpcan08/
- http://www.isuresults.com/results/gpchn08/
- http://www.isuresults.com/results/gpfra08/
- http://www.isuresults.com/results/gprus08/
- http://www.isuresults.com/results/gpjpn08/
- http://www.isuresults.com/results/gpf0809/
- http://www.isuresults.com/results/ec2009/
- http://www.isuresults.com/results/fc2009/
- http://www.isuresults.com/results/wc2009/

2009-2010

- http://www.isuresults.com/results/gpfra09/
- http://www.isuresults.com/results/gprus09/
- http://www.isuresults.com/results/gpchn09/
- http://www.isuresults.com/results/gpjpn09/
- http://www.isuresults.com/results/gpusa09/
- http://www.isuresults.com/results/gpcan09/
- http://www.isuresults.com/results/gpf0910/
- http://www.isuresults.com/results/ec2010/
- http://www.isuresults.com/results/fc2010/
- http://www.isuresults.com/results/owg2010/
- http://www.isuresults.com/results/wc2010/

2010-2011

- http://www.isuresults.com/results/gpjpn2010/
- http://www.isuresults.com/results/gpcan2010/
- http://www.isuresults.com/results/gpchn2010/
- http://www.isuresults.com/results/gpusa2010/
- http://www.isuresults.com/results/gprus2010/
- http://www.isuresults.com/results/gpfra2010/
- http://www.isuresults.com/results/gpf1011/
- http://www.isuresults.com/results/wc2011/
- http://www.isuresults.com/results/ec2011/
- http://www.isuresults.com/results/fc2011/

2011-2012

- http://www.isuresults.com/results/gpusa2011/
- http://www.isuresults.com/results/gpcan2011/
- http://www.isuresults.com/results/gpchn2011/
- http://www.isuresults.com/results/gpjpn2011/
- http://www.isuresults.com/results/gpfra2011/
- http://www.isuresults.com/results/gprus2011/
- http://www.isuresults.com/results/gpf1112/
- http://www.isuresults.com/results/ec2012/
- http://www.isuresults.com/results/wc2012/
- http://www.isuresults.com/results/fc2012/

2012-2013

- http://www.isuresults.com/results/gpusa2012/
- http://www.isuresults.com/results/gpcan2012/
- http://www.isuresults.com/results/gpchn2012/
- http://www.isuresults.com/results/gprus2012/
- http://www.isuresults.com/results/gpfra2012/
- http://www.isuresults.com/results/gpjpn2012/
- http://www.isuresults.com/results/gpf1213/
- http://www.isuresults.com/results/ec2013/
- http://www.isuresults.com/results/fc2013/
- http://www.isuresults.com/results/wc2013/

2013-2014

- http://www.isuresults.com/results/gpusa2013/
- http://www.isuresults.com/results/gpcan2013/
- http://www.isuresults.com/results/gpchn2013/
- http://www.isuresults.com/results/gpjpn2013/
- http://www.isuresults.com/results/gpfra2013/
- http://www.isuresults.com/results/gprus2013/
- http://www.isuresults.com/results/gpf1314/
- http://www.isuresults.com/results/ec2014/
- http://www.isuresults.com/results/fc2014/
- http://www.isuresults.com/results/owg2014/
- http://www.isuresults.com/results/wc2014/

2014-2015

- http://www.isuresults.com/results/gpusa2014/
- http://www.isuresults.com/results/gpcan2014/
- http://www.isuresults.com/results/gpchn2014/
- http://www.isuresults.com/results/gprus2014/
- http://www.isuresults.com/results/gpfra2014/
- http://www.isuresults.com/results/gpjpn2014/
- http://www.isuresults.com/results/gpf1415/
- http://www.isuresults.com/results/ec2015/
- http://www.isuresults.com/results/fc2015/
- http://www.isuresults.com/results/wc2015/

2015-2016

- http://www.isuresults.com/results/season1516/gpusa2015/
- http://www.isuresults.com/results/season1516/gpcan2015/
- http://www.isuresults.com/results/season1516/gpchn2015/
- http://www.isuresults.com/results/season1516/gpfra2015/
- http://www.isuresults.com/results/season1516/gprus2015/
- http://www.isuresults.com/results/season1516/gpjpn2015/
- http://www.isuresults.com/results/season1516/gpf1516/
- http://www.isuresults.com/results/season1516/ec2016/
- http://www.isuresults.com/results/season1516/fc2016/
- http://www.isuresults.com/results/season1516/wc2016/

2016-2017

- http://www.isuresults.com/results/season1617/gpusa2016/
- http://www.isuresults.com/results/season1617/gpcan2016/
- http://www.isuresults.com/results/season1617/gprus2016/
- http://www.isuresults.com/results/season1617/gpfra2016/
- http://www.isuresults.com/results/season1617/gpchn2016/
- http://www.isuresults.com/results/season1617/gpjpn2016/
- http://www.isuresults.com/results/season1617/gpf1617/
- http://www.isuresults.com/results/season1617/ec2017/
- http://www.isuresults.com/results/season1617/fc2017/
- http://www.isuresults.com/results/season1617/wc2017/

2017-2018

- http://www.isuresults.com/results/season1718/gprus2017/
- http://www.isuresults.com/results/season1718/gpcan2017/
- http://www.isuresults.com/results/season1718/gpchn2017/
- http://www.isuresults.com/results/season1718/gpjpn2017/
- http://www.isuresults.com/results/season1718/gpfra2017/
- http://www.isuresults.com/results/season1718/gpusa2017/
- http://www.isuresults.com/results/season1718/gpf1718/
- http://www.isuresults.com/results/season1718/ec2018/
- http://www.isuresults.com/results/season1718/fc2018/
- http://www.isuresults.com/results/season1718/owg2018/
- http://www.isuresults.com/results/season1718/wc2018/

Appendix B

Technical Scraping Details

All of the data and code are available in a public GitHub repository:

https://github.com/mengyazhu96/figure-skating-analysis

B.1 Modeling Competition Data

The structure of the figure skating season, disciplines, and segments suits object-oriented programming quite well. The human actors were modeled as follows.

- A `Skater` consists of a name, a country represented, and a discipline type. Skaters on occasion do change countries to represent, but most often this is one skater in a pair or ice dance couple switching countries in order to skate with another partner. In these cases, it is more reasonable to consider this new pair as a separate `Skater`.
- An `Official` is a member of a judging panel that is certified by the ISU, and consists of a name, a country that sponsors the official, and a function. This function delimits whether an official is a judge or a member of a technical panel. Officials on occasion can also switch countries.
- A `Panel` is a collection of officials that scores a particular segment of a competition. As detailed in Chapter 2, a panel will contain `Officials` that serve as either judges or members of the technical panel.

Historical scores are modeled using the following nested object types:

- A `Season` is a season of figure skating competition. A `Season` contains all of the `Events` that occurred in that season.

- An `Event` is a single competition such as the 2016 World Championships. It is always associated with a particular season, and will contain each of the four `Disciplines` that occurred at that event.
- A `Discipline` contains the results for a particular discipline at a particular event. This will summarize the skater entries in the event as well as the overall results of the event. It will contain the one to three `Segments` that constituted the competition.
- A `Segment` contains the results for one segment of a discipline at a particular event, such as the pairs short program at the 2017 World Championships. It will also be associated with the `Panel` that judged it. It will contain the `Scorecards` that breakdown each program score during the segment.
- A `Scorecard` contains all of the details in the PDF scorecard shown in Chapter 3, Figure 3.1: the `Skater`, the start number if recorded, the segment rank, and all of the score breakdowns. It contains a list of `Elements` and `ProgramComponents`, which each store the detailed information of each element and program component. It also stores any deductions and the reason for those deductions in a dictionary.

B.2 Acquiring and Parsing Data

B.2.1 Retrieval

Because the same important competitions repeat every year, acquiring the published information about scores was straightforward. The ISU has also been relatively consistent with its results URL formats (see Appendix A), so retrieving the information was a matter of an HTTP GET request for the event page, parsing through the important links using `BeautifulSoup` [18], and downloading any important links.

The data includes all downloaded data to check for parsing consistency. These downloaded files include:

- HTML files detailing judging panels, overall results, and entries.
- PDF files containing the official detailed score breakdown.

As mentioned previously, the only source of complete detailed score breakdown (scorecards) was in PDF format, so extracting this information was the main challenge of data collection. The majority of the data retrieval logic can be found in the `fetch_info` method of the `Event` type.

B.2.2 Parsing

HTML files were mostly straightforward to parse, as the format was very consistent and contained in tables. However, the PDF scorecards proved a substantial challenge due to small variations in scorecard formatting across competitions and the lack of any guiding lines on the PDFs.

The PDFs also do not contain guiding lines that assist in interpreting the scores as a table. For instance, each element should really be thought of as a row in a CSV file so that the element number, element name, base value, GOE, individual judges' GOEs, and element score can be separated clearly. However, this is not the case, and there are situations where the row alignment is off.

When first approaching this challenge, we explored a few Python PDF parsing options such as PDFMiner [47]. However, documentation was rather limited and sample usages not particularly enlightening. Their PDF to TXT conversion was also unreliable in the spaces introduced in between columns, as the library focused on text extraction from PDFs (such as articles saved as PDFs).

We thus explored PDF parsing libraries that focused on tabular data, because this was the main challenge in getting the useful information out of the PDF. We came across Tabula, a Java library for "extracting tables from PDF files" [48]. Tabula's PDF to CSV parsing turned out to be much more consistent than what we tried with PDFMiner.

The final workflow of extracting data from PDFs thus started by converting each scorecard PDF using Tabula (`Event.pdfs_to_csvs`). This resultant CSV did not delimit table columns perfectly and also contained unnecessary information, so we had to do extensive further parsing. This further parsing (`Segment.parse_raw_csv`) employed extensive regular expression matching, which was flexible enough to extract all the data starting in the 2005-2006 season. IJS was also used in the 2004-2005 season, but scorecards would contain one more judge than reported in the panel description. Since this was the first year for IJS and likely would not contain any extremely useful information as a transition period, we chose to leave out the 2004-2005 season.

The finalized parsing logic is mainly per segment (`Segment.parse_raw_csv`), as there is one PDF file per segment. The logic essentially reads the Tabula-generated CSV line by line and matches it against regular expressions for useful information, adding this information to an in-progress scorecard while parsing. Once a new scorecard is detected, the current scorecard is completed.

Each parsed scorecard was also checked for consistency to make sure that the element scores summed up to the technical scores, the program component scores matched the summary reported, and so on. The final parsed data contains no mistakes aside from differences of at

most 0.03 due to rounding differences. In figure skating, medalists are very rarely decided by less than one tenth of a point.

Each parsed scorecard is written to its own CSV to facilitate easy extraction, and the data types for seasons, events, and so on can easily read this data in. This makes it easy to create CSVs for data analysis. For instance, to create a data frame of every single element performed since 2005, we can iterate through all the seasons, events, disciplines, segments, and scorecards to create a row for each element.

One issue that we had to resolve were quirks in judging. At the 2015 Four Continents Championships, one judge did not mark a GOE for the last place skater in the free skate, Harry Hau Yin Lee [49]. For this judge, we recorded the GOE as 0, which was the median and mode of the other judges' scores. At Skate America 2016 (the Grand Prix in the U.S.), judge number 6 was removed from the pair's short program [50]. Multiple ice dance scorecards such as the short program at Skate Canada 2015 had excess columns [51]. For these situations of excess columns or excluded judges, we assumed the judges were never in the data at all and removed the columns altogether.

B.3 Resolving Names

Aside from the changes mentioned in Chapter 3, one other obstacle to performing per-skater analysis was matching skaters and judges across competitions. The ISU was not consistent in how names were formatted, except for always having the last name capitalized. Even then, a skater with a multi-word last name such as Kevin VAN DER PERREN would have certain words in their last name capitalized inconsistently. Skaters' names could also be reported as "first last" or "last first," for example Evan Lysacek could be reported as Evan LYSACEK or as LYSACEK Evan. Skaters also sometimes changed the spelling of their names. For example, Aljona Savchenko, the 2018 Olympic pairs champion, changed the official ISU spelling of her name from Aliona to Aljona in 2018.

To resolve this, we took the data frame containing all results for each discipline and collected the set of names that Python considered to be unique. We did not change any cases in this set of names, because the uppercasing distinguished between first and last name. For each pair of names x, y with $x \neq y$, we applied the Levenshtein algorithm to compute the edit distance between x, y [52]. We also swapped the first and second words (first and last or last and first names) of y and computed the edit distance between this modified name and x in order to catch cases such as Mao ASADA and ASADA Mao. We took the similar pairs reported (an edit distance of less than 7 to be overcautious in finding duplicates) and manually iterated through them to choose a single version for each skater. We chose the format Mao ASADA to

use consistently. We also had to manually check for names with more than two words (more than one space), such as Ingrid Charlotte WOLTER. This was manageable because each discipline ended up with about 300 unique names or fewer.

We employed the same strategy for judge names, but judge countries also needed to be determined. At championship events (Europeans, Four Continents, Olympics, and Worlds), judges' nations are reported as ISU. We thus had to map each judge to every nation they've ever been reported as, and hoped that this would give the actual country of origin. However, many judges still ended up with no nation reported as they had only ever been reported as representing the ISU. The website "Skating Scores" has judge countries for judges in the past few years, so we used their information to fill in the missing judge countries [53]. Four judges changed countries of representation in the data:

- Irina MEDVEDEVA represented Ukraine until the 2014-2015 season, when she began representing Azerbaijan.
- Garry HOPPE represented Israel until the 2007-2008 season. After no appearances in the 2008-2009 season, he began representing Great Britain for the 2009-2010 season.
- Irina ABSALIAMOVA represented Belarus in the 2005-2006 season. After no appearances in the 2006-2007 season, she began judging for Armenia in 2007-2008.
- Lolita LABUNSKAIYA represented Ukraine in the Grand Prix of Russia in 2005. A judge by the same name represented Russia in the 2011-2012 season. Because of the time gap, it is not immediately clear even after a Google search whether the Ukrainian and the Russian judges are the same person. However, the Ukrainian judge judged only one early competition so this should not be a huge issue.

References

[1] Ahiza Garcia. U.S. figure skating used to be wildly popular. What happened? *CNN Money*, February 2018. URL: http://money.cnn.com/2018/02/13/news/figure-skating-popularity-us-olympics-pyeongchang/index.html.

[2] Jarrett Bell. Number crunching a growing craze in the NFL. *USA TODAY*, May 2013. URL: https://www.usatoday.com/story/sports/2013/05/15/advanced-statistics-nfl/2164723/.

[3] NBA Stats. URL: https://stats.nba.com/.

[4] Sameer K. Deshpande and Shane T. Jensen. Estimating an NBA player's impact on his team's chances of winning. *Journal of Quantitative Analysis in Sports*, 12(2):51–72, March 2016.

[5] Hot-hand fallacy, January 2018. Page Version ID: 822379628. URL: https://en.wikipedia.org/w/index.php?title=Hot-hand_fallacy&oldid=822379628.

[6] Kathleen Elkins. 5 years ago, US Olympian Adam Rippon was broke and would 'steal all the apples' at his gym. *CNBC*, February 2018. URL: https://www.cnbc.com/2018/02/12/adam-rippon-was-once-so-broke-he-stole-apples-from-the-gym.html.

[7] Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553, November 2009. URL: http://linkinghub.elsevier.com/retrieve/pii/S0167923609001377, doi:10.1016/j.dss.2009.05.016.

[8] Ryan Rodenberg. Perception ? Reality: Analyzing Specific Allegations of NBA Referee Bias. *Journal of Quantitative Analysis in Sports*, 7(2), 2011. URL: http://www.degruyter.com/view/j/jqas.2011.7.2/jqas.2011.7.2.1326/jqas.2011.7.2.1326.xml?rskey=g7ESEv&result=1&q=referee+bias, doi:10.2202/1559-0410.1326.

[9] How a Wine Competition Works, January 2017. URL: https://www.newyorkwines.org/awards-how-a-competition-works.

[10] 2002 Winter Olympics figure skating scandal, February 2018. Page Version ID: 827319486. URL: https://en.wikipedia.org/w/index.php?title=2002_Winter_Olympics_figure_skating_scandal&oldid=827319486.

[11] 6.0 system, February 2018. Page Version ID: 826760376. URL: https://en.wikipedia.org/w/index.php?title=6.0_system&oldid=826760376.

[12] ISU Judging System, March 2018. Page Version ID: 832344149. URL: https://en.wikipedia.org/w/index.php?title=ISU_Judging_System&oldid=832344149.

[13] Four Continents Figure Skating Championships, March 2018. Page Version ID: 829554028. URL: https://en.wikipedia.org/w/index.php?title=Four_Continents_Figure_Skating_Championships&oldid=829554028.

[14] Ryan Mitchell. *Web Scraping with Python*. O'Reilly Media, Inc., July 2015. URL: http://proquest.safaribooksonline.com/9781491910283.

[15] Python 2.7.13. URL: https://www.python.org/.

[16] John D. Hunter Droettboom, Michael. matplotlib: Python plotting package. URL: http://matplotlib.org.

[17] Kenneth Reitz. requests: Python HTTP for Humans. URL: http://python-requests.org.

[18] Leonard Richardson. beautifulsoup4: Screen-scraping library. URL: http://www.crummy.com/software/BeautifulSoup/bs4/.

[19] The PyData Development Team. pandas: Powerful data structures for data analysis, time series,and statistics. URL: http://pandas.pydata.org.

[20] GPF 2015-2016 Men Short Program Scores. URL: http://www.isuresults.com/results/season1516/gpf1516/gpf1516_Men_SP_Scores.pdf.

[21] John W. Emerson and Taylor B. Arnold. Statistical Sleuthing by Leveraging Human Nature: A Study of Olympic Figure Skating. *The American Statistician*, 65(3):143–148, 2011. URL: http://www.jstor.org/stable/24591407.

[22] ISU Grand Prix and Junior Grand Prix Final - Ice Dance. URL: http://www.isuresults.com/results/gpf0910/CAT008RS.HTM.

[23] XXI Olympic Winter Games 2010 - Ice Dance. URL: http://www.isuresults.com/results/owg2010/CAT004RS.HTM.

[24] Wändi Bruine de Bruin. Save the last dance II: Unwanted serial position effects in figure skating judgments. *Acta Psychologica*, 123(3):299–311, November 2006. URL: http://www.sciencedirect.com/science/article/pii/S0001691806000187, doi:10.1016/j.actpsy.2006.01.009.

[25] PCS inflation in the last group, March 2015. URL: https://www.goldenskate.com/forum/showthread.php?54711-PCS-inflation-in-the-last-group.

[26] A Primer on Bayesian Methods for Multilevel Modeling — PyMC3 3.3 documentation. URL: http://docs.pymc.io/notebooks/multilevel_modeling.html.

[27] Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, New York, 2007.

[28] Skipper Seabold Perktold, Josef. statsmodels: Statistical computations and models for Python. URL: http://www.statsmodels.org/.

[29] Thomas Wiecki. pymc3: PyMC3. URL: http://github.com/pymc-devs/pymc3.

[30] Pair skating, February 2018. Page Version ID: 826016976. URL: https://en.wikipedia.org/w/index.php?title=Pair_skating&oldid=826016976.

[31] Ice dancing, March 2018. Page Version ID: 832352071. URL: https://en.wikipedia.org/w/index.php?title=Ice_dancing&oldid=832352071.

[32] Andrew Gelman. Time-series regression question, June 2005. URL: http://andrewgelman.com/2005/06/21/timeseries_regr/.

[33] Bernie Lincicome. It's a quadrennial duty to inform that figure skating is not a sport. *Chicago Tribune*, February 2018. URL: http://www.chicagotribune.com/sports/international/ct-spt-lincicome-figure-skating-olympics-20180217-story.html.

[34] Lee Moran. Olympians Shut Down Local Fox Anchor Who Said Figure Skating Is 'Not A Sport'. *Huffington Post*, February 2018. URL: https://www.huffingtonpost.com/entry/figure-skating-not-a-sport-local-anchor_us_5a83e965e4b0adbaf3d8cd01.

[35] Gilbert W. Bassett and Joseph Persky. Rating Skating. *Journal of the American Statistical Association*, 89(427):1075–1079, 1994. URL: http://www.jstor.org/stable/2290937, doi:10.2307/2290937.

[36] Samuel S. Wu and Mark C. K. Yang. Evaluation of the Current Decision Rule in Figure Skating and Possible Improvements. *The American Statistician*, 58(1):46–54, 2004. URL: http://www.jstor.org/stable/27643498.

[37] Eric Zitzewitz. Nationalism in Winter Sports Judging and Its Lessons for Organizational Decision Making. *Journal of Economics & Management Strategy*, 15(1):67–99, March 2006. doi:10.1111/j.1530-9134.2006.00092.x.

[38] John Templon and Rosalind Adams. Top-Level Figure Skating Judges Consistently Favor Skaters From Their Home Countries. *BuzzFeed*, February 2018. URL: https://www.buzzfeed.com/johntemplon/the-edge.

[39] John Templon and Rosalind Adams. How Figure Skating Judges May Have Shaped The Olympic Podium. *BuzzFeed*, February 2018. URL: https://www.buzzfeed.com/johntemplon/by-voting-for-their-own-figure-skating-judges-may-have.

[40] Bryan Campbell and John W. Galbraith. Nonparametric Tests of the Unbiasedness of Olympic Figure-Skating Judgments. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 45(4):521–526, 1996. URL: http://www.jstor.org/stable/2988550, doi:10.2307/2988550.

[41] John W. Emerson, Miki Seltzer, and David Lin. Assessing Judging Bias: An Example From the 2000 Olympic Games. *The American Statistician*, 63(2):124–131, May 2009. doi:10.1198/tast.2009.0026.

[42] Andreas Mueller. scikit-learn: A set of python modules for machine learning and data mining. URL: http://scikit-learn.org.

[43] Bryan Campbell and Jean-Marie Dufour. Exact Nonparametric Orthogonality and Random Walk Tests. *The Review of Economics and Statistics*, 77(1):1–16, 1995. doi:10.2307/2109988.

[44] Juliet Macur. Adelina Sotnikova's Upset Victory Is Hard to Figure. *The New York Times*, February 2014. URL: https://www.nytimes.com/2014/02/21/sports/olympics/adelina-sotnikovas-upset-victory-is-hard-to-figure.html.

[45] Chess rating system, February 2018. Page Version ID: 824906265. URL: https://en.wikipedia.org/w/index.php?title=Chess_rating_system&oldid=824906265.

[46] Philip Hersh. ISU official: 'Radical change' could be on the way. *Icenetwork*, September 2017. URL: http://web.icenetwork.com/news/2017/09/11/253667206.

[47] Yusuke Shinyama. pdfminer: PDF parser and analyzer. URL: http://euske.github.io/pdfminer/index.html.

[48] tabula-java: Extract tables from PDF files, March 2018. original-date: 2014-05-22T03:11:57Z. URL: https://github.com/tabulapdf/tabula-java.

[49] FC2015 Men Free Skating Scores, 2015. URL: http://www.isuresults.com/results/fc2015/fc2015_Men_FS_Scores.pdf#page=12.

[50] GPUSA2016 Pairs Short Program Scores, 2016. URL: http://www.isuresults.com/results/season1617/gpusa2016/gpusa2016_Pairs_SP_Scores.pdf.

[51] GPCAN2015 Ice Dance Short Dance Scores, 2015. URL: http://www.isuresults.com/results/season1516/gpcan2015/gpcan2015_IceDance_SD_Scores.pdf.

[52] Levenshtein distance, March 2018. Page Version ID: 830427260. URL: https://en.wikipedia.org/w/index.php?title=Levenshtein_distance&oldid=830427260.

[53] Skating Scores: Latest Figure Skating Scores, Rankings & Statistics, 2018. URL: http://skatingscores.com.