Predicting Gymnastics Scores

Alyssa Weber

November 19, 2022

Introduction

For the past 10 years I have been managing a competitive gymnastics team for level three through five. It is my job to train these athletes and the coaching staff in order to run a safe, competitive program. Each year, our goal is to rank within the top five at state. In order to do so, team scores need to project to be somewhere between 110-113 points, and then also be achieved on the day of of the competition. I have been collecting competitive data on my gymnasts for years and want to know what information is locked inside the spreadsheets. I plan on using data science techniques to not only clean up data, but to also find trends and predictive insights on an individual and team basis. If I could unlock predictive tendencies from simple data, the entire coaching community would want in on the power of data science.

Research questions

1. Is there a relationship between individual scores earned at the first meet of the year and scores earned at the state meet?
2. Is there a relationship between team scores earned at the first meet of the year and scores earned at the state meet?
3. Does a team have limited or unlimited growth in overall scores?
4. Is it better to set a lineup from the previous meet scores or from personal best scores?
5. Do the predictive tendencies hold true on other sets of data (different levels, different years)?

Approach

The first step will be to clean up the data. There are many null values in both the columns and rows that need to be removed. The columns will need some renaming to make them easier to work with. I also plan on breaking up the data sets since they include both team and individual scores. Organizing the data will help make the future steps more concise and accurate. In order to look for predictive tenancies, I plan on working with scatter plots, linear models, and regressions. I will start by fully visualizing the data, then move into summary statistics. I will look for correlations between early data and state data. The use of confidence intervals and hypothesis testing will be key in order to see if the results are significant. I will make sure sure that the data fits a linear model well, before drawing any conclusions.

How your approach addresses (fully or partially) the problem.

The steps of cleaning, visualizing, then testing will be the best approach to find results. The ultimate goal is to be able to look at early data and use it to predict future results. The approach above encompasses many

of the techniques learned through the Statistics for Data Science Course and should suffice to either reveal significant results, or open the doors for more data related questions.

Data

The data is original to myself. The current data was collected over the course of September through December of 2021 with the purpose of tracking personal best scores for the level 4 gymnasts whom I coach. Each value was entered by hand. The original file is a spreadsheet with six main sheets.

1. Meet 1
2. Meet 2
3. Meet 3
4. Meet 4
5. State
6. Personal Bests

Each sheet contains the data for 13 individual gymnasts including their score for each of the four gymnastics events, their rank among their teammates, the place they were awarded in their age divisions, their all-around score, and whether or not they achieved qualifying scores for state and/or the next level. Each sheet also contains team totals for each event and the overall score. Missing values in the original data set were left blank and indicated with a NA when imported into R.

The original data sets can be found here: Level 4 2021 Competition Season Score-sheet

Required Packages

- ReadXL - to open and read excel files
- GGPlot2 - to produce high quality graphs and plots
- dplyr - to clean and merge the data

Plots and Table Needs

In order to visualize the data, I will mainly make use of scatter plots with regressions. However, box plots and histograms may come into play as well. I am hopeful that regression models will be a key component of the research. If this is the case, I will need a table or data frame that holds the diagnostics needed for determining outliers and independence.

How to import and clean my data

To import the data I will need to set the working directory to the source of the project file. I will download the package readxl in order to read excel spreadsheets. Since the sheet consists of six separate sheets of data, each one will need to be loaded separately.

```
setwd("C:/Users/lyssi/OneDrive/Documents/GitHub/Weber-DSC520/Final Project")
library(readxl)
excel_sheets("Level 4 2021 Competition Season Scoresheet.xlsx")
```

```
##  [1] "InHouse"                   "Meet1"
##  [3] "Meet2"                     "Meet3"
##  [5] "Meet4"                     "State"
##  [7] "Personal Bests"            "2019 groups for offseason"
##  [9] "Extra1"                    "Extra2"
## [11] "Blank score sheet"
```

```
inhousemeet <- read_excel("Level 3 2022-2023 Competition Season Scoresheet.xlsx"
                          ,sheet = "InHouse")
```

```
## New names:
## * `Place` -> `Place...3`
## * `RS rank` -> `RS rank...4`
## * `Place` -> `Place...6`
## * `RS rank` -> `RS rank...7`
## * `Place` -> `Place...9`
## * `RS rank` -> `RS rank...10`
## * `Place` -> `Place...12`
## * `RS rank` -> `RS rank...13`
## * `Place` -> `Place...15`
## * `RS rank` -> `RS rank...16`
## * `` -> `...17`
```

```
meet1 <- read_excel("Level 4 2021 Competition Season Scoresheet.xlsx",
                    sheet = "Meet1")
```

```
## New names:
## * `Place` -> `Place...3`
## * `RS rank` -> `RS rank...4`
## * `Place` -> `Place...6`
## * `RS rank` -> `RS rank...7`
## * `Place` -> `Place...9`
## * `RS rank` -> `RS rank...10`
## * `Place` -> `Place...12`
## * `RS rank` -> `RS rank...13`
## * `Place` -> `Place...15`
## * `RS rank` -> `RS rank...16`
## * `` -> `...17`
```

```
meet2 <- read_excel("Level 4 2021 Competition Season Scoresheet.xlsx",
                    sheet = "Meet2")
```

```
## New names:
## * `Place` -> `Place...3`
## * `RS rank` -> `RS rank...4`
## * `Place` -> `Place...6`
## * `RS rank` -> `RS rank...7`
## * `Place` -> `Place...9`
## * `RS rank` -> `RS rank...10`
## * `Place` -> `Place...12`
## * `RS rank` -> `RS rank...13`
## * `Place` -> `Place...15`
## * `RS rank` -> `RS rank...16`
## * `` -> `...17`
```

```
meet3 <- read_excel("Level 4 2021 Competition Season Scoresheet.xlsx",
                    sheet = "Meet3")
```

```
## New names:
## * `Place` -> `Place...3`
## * `RS rank` -> `RS rank...4`
## * `Place` -> `Place...6`
## * `RS rank` -> `RS rank...7`
## * `Place` -> `Place...9`
## * `RS rank` -> `RS rank...10`
## * `Place` -> `Place...12`
## * `RS rank` -> `RS rank...13`
## * `Place` -> `Place...15`
## * `RS rank` -> `RS rank...16`
## * `` -> `...17`
```

```
meet4 <- read_excel("Level 4 2021 Competition Season Scoresheet.xlsx",
                    sheet = "Meet4")
```

```
## New names:
## * `Place` -> `Place...3`
## * `RS rank` -> `RS rank...4`
## * `Place` -> `Place...6`
## * `RS rank` -> `RS rank...7`
## * `Place` -> `Place...9`
## * `RS rank` -> `RS rank...10`
## * `Place` -> `Place...12`
## * `RS rank` -> `RS rank...13`
## * `Place` -> `Place...15`
## * `RS rank` -> `RS rank...16`
## * `` -> `...17`
```

```
statemeet <- read_excel("Level 4 2021 Competition Season Scoresheet.xlsx",
                    sheet = "State")
```

```
## New names:
## * `Place` -> `Place...3`
## * `RS rank` -> `RS rank...4`
## * `Place` -> `Place...6`
## * `RS rank` -> `RS rank...7`
## * `Place` -> `Place...9`
## * `RS rank` -> `RS rank...10`
## * `Place` -> `Place...12`
## * `RS rank` -> `RS rank...13`
## * `Place` -> `Place...15`
## * `RS rank` -> `RS rank...16`
## * `` -> `...17`
## * `` -> `...18`
```

```
personalbests <- read_excel("Level 4 2021 Competition Season Scoresheet.xlsx",
                    sheet = "Personal Bests")
```

```
## New names:
## * `Place` -> `Place...3`
## * `RS rank` -> `RS rank...4`
## * `Place` -> `Place...6`
## * `RS rank` -> `RS rank...7`
## * `Place` -> `Place...9`
## * `RS rank` -> `RS rank...10`
## * `Place` -> `Place...12`
## * `RS rank` -> `RS rank...13`
## * `Place` -> `Place...15`
## * `RS rank` -> `RS rank...16`
## * `` -> `...17`
```

The cleaning process will consist of creating a new data frame that houses the data needed for this research. First, I will build a data frame that holds the data from meet 1 and the state meet. I will only pull the event, all around, and mobility score data. Rows 13-19 do not contain gymnast data and will be deleted.

```
meet1scores <- data.frame(meet1$Name, meet1$Vault, meet1$Bars, meet1$Beam, meet1$Floor, meet1$AA, meet1$
meet1scores <- meet1scores[-(13:19),]

statemeetscores <- data.frame(statemeet$Name, statemeet$Vault, statemeet$Bars, statemeet$Beam, statemee
statemeetscores <- statemeetscores[-(13:19),]
```

The column names are too long and should be condensed. All scores from meet 1 will be abbreviated with 'm1' while state meet data will be abbreviated with 'state'. The letters v, ba, be, f, aa, and mn will represent each of the following categories respectively; vault, bars, beam, floor, all around, and mobility score.

```
names(meet1scores) <- c('name', 'm1v', 'm1ba', 'm1be', 'm1f', 'm1aa', 'm1ms')

names(statemeetscores) <- c('name', 'statev', 'stateba', 'statebe', 'statef', 'stateaa', 'statems')
```

The state all around scores imported as characters. They will need to be converted to a double with 3 decimals.

```
statemeetscores$stateaa <- as.numeric(statemeetscores$stateaa)
```

Next, I would like to combine the meet 1 data frame with the state meet data frame corresponding by name.

```
scoresdf <- merge(meet1scores,statemeetscores,by="name")
```

Gymnast EB, #5 did not compete all four events at meet 1 or state. The scores of 0.000 will potentially skew the results and should be removed.

```
scoresdf <- scoresdf[-(5),]
```

What does the final data set look like?

Here is the head of the resulting data set:

```
head(scoresdf)
```

```
##   name  m1v  m1ba  m1be   m1f   m1aa m1ms statev stateba statebe statef stateaa
## 1   AC 7.75 7.750 8.725 9.050 33.275   NO  8.800   8.875   8.625  9.225  35.525
## 2   AD 8.10 8.600 8.625 8.200 33.525   NO  8.825   9.200   9.300  9.175  36.500
## 3   AH 7.75 7.350 9.050 8.200 32.350   NO  7.850   8.575   8.925  8.650  34.000
## 4   CT 7.35 8.475 8.900 8.400 33.125   NO  8.350   9.175   8.700  9.100  35.325
## 6   EG 8.00 6.950 7.950 8.150 31.050   NO  8.050   8.700   9.000  8.625  34.375
## 7   HS 8.35 8.550 8.400 8.825 34.125  YES  8.475   9.075   8.575  8.900  35.025
##   statems
## 1     YES
## 2     YES
## 3     YES
## 4     YES
## 6     YES
## 7     YES
```

What information is not self-evident?

At this point, the new data frame does not contain any of the personal best data or meets in between the first meet and the state meet. It is not evident at this point if scores increased or not. There may also not be enough data to draw valid conclusions.

What are different ways you could look at this data?

This data can look at each event's correlation between the first meet and the state meet. We can also look at a particular case's (gymnast's) growth or decline. We can see how many gymnasts earned a mobility score that did not earn it at the beginning of the season.

How do you plan to slice and dice the data?

The data is already merged together in a format that will be sufficient for generating models. As previously mentioned, key columns with complete data were taken from two separate pages of the original spreadsheet. They were merged back together based on gymnast names. Null rows and rows with team data rather than individual data were deleted.

How could you summarize your data to answer key questions?

I could start with a summary function which would give me 5-number summaries and averages of each variable. From a quick glance, it looks as if the all around scores have increased by about two points from the first meet to the state meet. This is something that I will want to explore further.

```
summary(scoresdf)
```

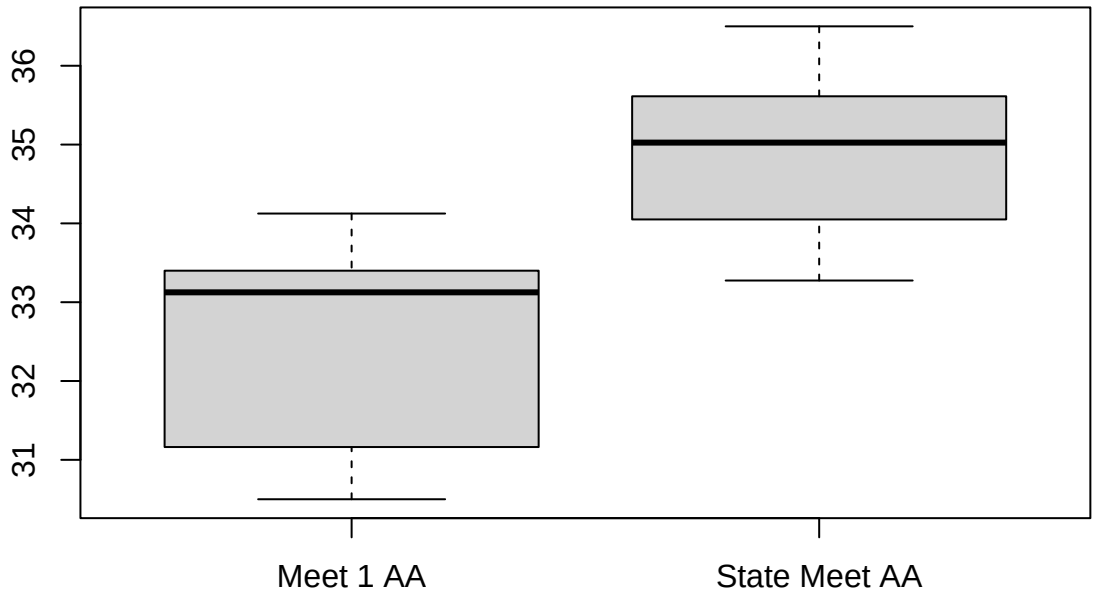
```
##      name                m1v             m1ba            m1be      
##  Length:11          Min.   :6.750   Min.   :6.350   Min.   :7.400  
##  Class :character   1st Qu.:7.725   1st Qu.:7.150   1st Qu.:8.125  
##  Mode  :character   Median :8.000   Median :7.750   Median :8.575  
##                     Mean   :7.909   Mean   :7.714   Mean   :8.384  
```

```
##                     3rd Qu.:8.325   3rd Qu.:8.512   3rd Qu.:8.775
##                     Max.   :8.500   Max.   :9.075   Max.   :9.050
##       m1f             m1aa           m1ms              statev
##  Min.   :8.150   Min.   :30.50   Length:11          Min.   :7.850
##  1st Qu.:8.213   1st Qu.:31.16   Class :character   1st Qu.:8.175
##  Median :8.400   Median :33.12   Mode  :character   Median :8.375
##  Mean   :8.441   Mean   :32.45                      Mean   :8.420
##  3rd Qu.:8.575   3rd Qu.:33.40                      3rd Qu.:8.738
##  Max.   :9.050   Max.   :34.12                      Max.   :8.975
##     stateba         statebe          statef         stateaa
##  Min.   :7.725   Min.   :8.400   Min.   :8.075   Min.   :33.27
##  1st Qu.:8.600   1st Qu.:8.613   1st Qu.:8.550   1st Qu.:34.05
##  Median :8.700   Median :8.925   Median :8.900   Median :35.02
##  Mean   :8.768   Mean   :8.857   Mean   :8.809   Mean   :34.85
##  3rd Qu.:9.125   3rd Qu.:9.025   3rd Qu.:9.137   3rd Qu.:35.61
##  Max.   :9.325   Max.   :9.300   Max.   :9.225   Max.   :36.50
##    statems
##  Length:11
##  Class :character
##  Mode  :character
##
##
##
```

What types of plots and tables will help you to illustrate the findings to your questions?

A box and whisker plot would be a beneficial first chart to create to showcase the spread, min, max, and median scores from the beginning of the season compared to the end. I would also like a scatter plot for each event comparing meet 1 and the state meet. I could make a histogram of a variable to see if the data is normally distributed or not. That would help me decide what tests and models could be built from the data.

```
boxplot(scoresdf$m1aa,scoresdf$stateaa, names = c("Meet 1 AA", "State Meet AA"))
```



Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

I do not plan on incorporating machine learning techniques into my research questions as of yet. I do not believe that my data set is large enough to split into a training and test group at this point. If I were to use these techniques, it would be to predict growth of a gymnast based on the first meet of the year.

Analysis

The analysis of choice will be a simple linear model. The hypotheses being tested include the following:

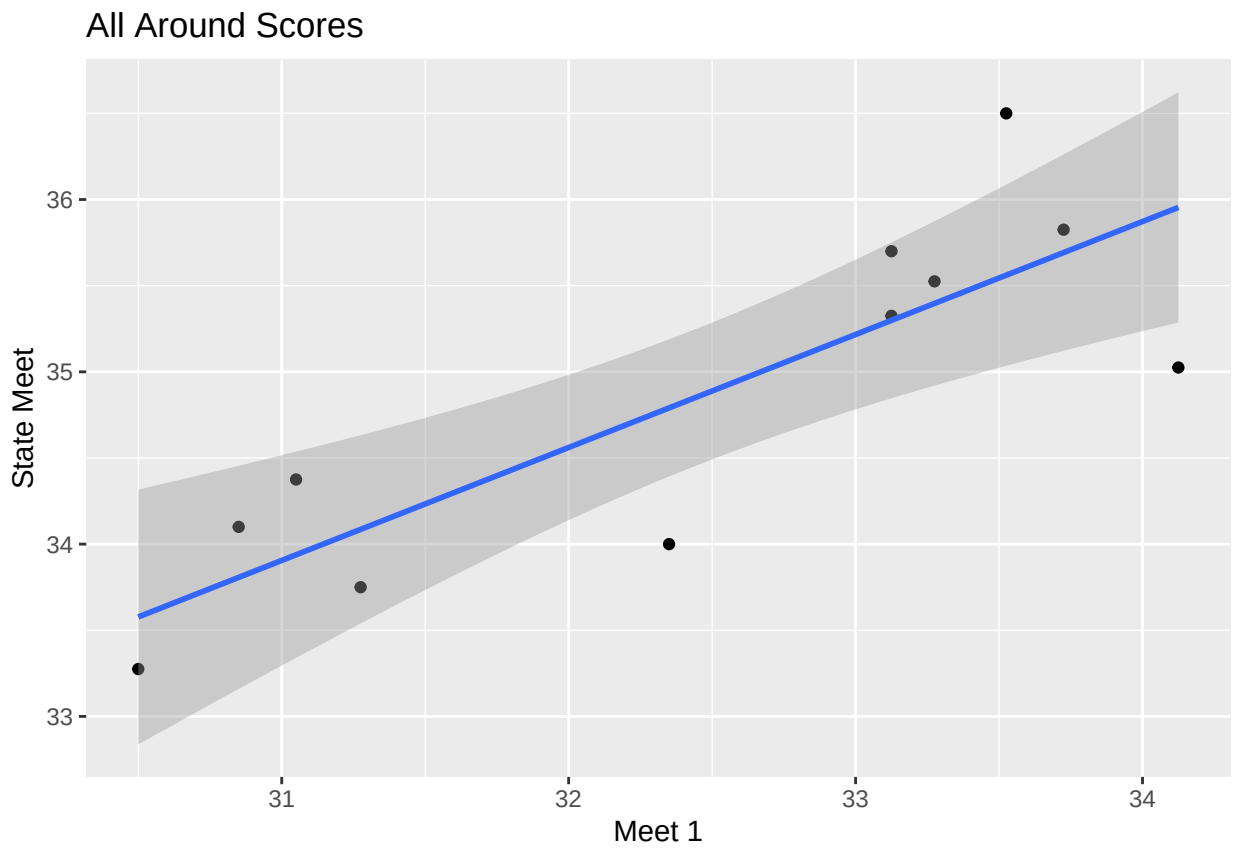
> Null Hypothesis: Meet 1 all-around scores do not impact state meet scores (There is no correlation) Alternative Hypothesis: Meet 1 all-around scores do impact state meet scores

I will first construct a visual.

```
library(ggplot2)

ggplot(scoresdf, aes(x = m1aa, y = stateaa)) + geom_point() +
    labs(x = "Meet 1", y = "State Meet") + ggtitle("All Around Scores") +
    geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Visually, it looks like a linear regression is a good fit. To determine if this model represents the data well, I will extract the intercept and slope and run a summary on the model.

```
scoreslm <- lm(stateaa ~ m1aa, data = scoresdf)
scoreslm
```

```
##
## Call:
## lm(formula = stateaa ~ m1aa, data = scoresdf)
##
## Coefficients:
## (Intercept)         m1aa
##     13.5802       0.6556
```

```
summary(scoreslm)
```

```
##
## Call:
## lm(formula = stateaa ~ m1aa, data = scoresdf)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9293 -0.3191  0.1280  0.3472  0.9391
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  13.5802     4.5971   2.954  0.01611 *
## m1aa          0.6556     0.1416   4.631  0.00123 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5819 on 9 degrees of freedom
## Multiple R-squared:  0.7044, Adjusted R-squared:  0.6716
## F-statistic: 21.45 on 1 and 9 DF,  p-value: 0.001235
```

The p-value is less than 0.05, which is a statistically significant result. We can reject the null hypothesis that Meet 1 all-around scores do not impact state meet scores. The summary also reveals a coefficient of determination, r^2, value of 0.7. From this I can conclude that the meet 1 all-around scores account for 70% of the variability in state meet scores, leaving only 30% of the variation still to be accounted for by other variables.

Implications

Based on the analysis, there is a strong positive correlation between meet 1 all-around scores and the state meet all-around scores. This means that I can use the model above to predict how a gymnast will perform at state based on her first competition of the year.

Limitations

This model was built on a relatively small sample size. The chances that this model will have an accurate success rate on future data sets is small. After the 2022 season has commenced, I plan on using the training data to see how well it performs on a new set of data. I would like to know the amounts of false positive and false negative rates. I can later increase the sample size to see if the results become more accurate.

Concluding Remarks

Having clean data is essential for a smooth, accurate model. A third of this project was parsing through, deleting, adding, combining, and reformatting data. The small sample size in this project was perfect for practicing techniques in RStudio. I look forward to drawing further conclusions from a more advanced database.