Winning Characteristics in the Olympics

Noah Obuya and Tamya Davidson

```
library(tidyverse)
```

```
-- Attaching packages --------------------------------------- tidyverse 1.3.2 --
v ggplot2 3.3.6     v purrr   0.3.4
v tibble  3.1.8     v dplyr   1.0.9
v tidyr   1.2.0     v stringr 1.4.1
v readr   2.1.2     v forcats 0.5.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

```
library(tidymodels)
```

```
-- Attaching packages --------------------------------------- tidymodels 1.0.0 --
v broom        1.0.1     v rsample      1.1.0
v dials        1.0.0     v tune         1.0.1
v infer        1.0.3     v workflows    1.1.0
v modeldata    1.0.1     v workflowsets 1.0.0
v parsnip      1.0.2     v yardstick    1.1.0
v recipes      1.0.2
-- Conflicts ------------------------------------------ tidymodels_conflicts() --
x scales::discard() masks purrr::discard()
x dplyr::filter()   masks stats::filter()
x recipes::fixed()  masks stringr::fixed()
x dplyr::lag()      masks stats::lag()
x yardstick::spec() masks readr::spec()
x recipes::step()   masks stats::step()
* Use suppressPackageStartupMessages() to eliminate package startup messages
```

```r
library(formatR)
library(MASS)
```

```
Attaching package: 'MASS'

The following object is masked from 'package:dplyr':

    select
```

```r
library(nnet)
library(car)
```

```
Loading required package: carData

Attaching package: 'car'

The following object is masked from 'package:dplyr':

    recode

The following object is masked from 'package:purrr':

    some
```

```r
library(lme4)
```

```
Loading required package: Matrix

Attaching package: 'Matrix'

The following objects are masked from 'package:tidyr':

    expand, pack, unpack
```

```r
library(glmnet)
```

```
Loaded glmnet 4.1-6
```

```
olympics <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday
```

```
Rows: 271116 Columns: 15
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (10): name, sex, team, noc, games, season, city, sport, event, medal
dbl  (5): id, age, height, weight, year

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
olympics <- olympics %>%
 na.omit

olympics04 <- olympics %>%
  filter(year == 2004)

olympics08 <- olympics %>%
  filter(year == 2008)

olympics12 <- olympics %>%
  filter(year == 2012)

olympics16 <- olympics %>%
  filter(year == 2016)
```

Introduction and Data

Research Question

**What are the most influential characteristics (between height and weight when it comes to predicting gold medals?

The data that we chose was olympics data from TidyTuesday's github repository (https://github.com/rfordatascience/tidytuesday/b lob/master/data/2021/2021-07-27/readme.md). The dataset was created in May 2018, and the data were collected by scraping www.sports-reference.com. The data contains 271,116 observations of 15 variables. The variables of interest in our research include sex, age, height , weight , noc (country) , year , season, and medals (Gold, Silver, Bronze). Based on these variables , we will answer the question of what

are the most influential variables that influence an athlete receiving a gold medal, and if these variables change over time. From the dataset we will only observe the more recent olympic games (including the years 2004, 2008, 2012, 2016). As part of our data-cleaning process, we have created 4 different dataset subsections for each of these years. Additionally, there were many NA values corresponding to medals. Since this is our variable of interest, we will drop all NA values corresponding to medals. After doing this we are left with a case study of 39,783 observations of 15 variables. The motivation behind this project is to analyze what athletes can do to better prepare for the Olympic games, and see which factors are more influential than others.

Variables of Interest

Variables of Interest

Sex - Sex Assigned at Birth of the Olympian

Age - Age of the Olympian in years

Height - Height of the Olympian in centimeters (cm)

Weight - Weight of the Olympian in kilograms (kg)

NOC - Country as assigned by the National Olympic Committee

Methodology

We will be fitting an ordinal model. We are doing this because the outcome variable, Medal, is not only categorical, but ordered. Within the three separate categories of medal, we can assume that an Olympian who received a gold medal has done better than an Olympian who received a silver medal. The same can be said for a silver medal vs. a bronze medal and a gold medal vs. a bronze medal.

The assumptions of an ordinal model include the following:

1.The dependent variable are ordered

2.One or more of the independent variables are either continuous, categorical or ordinal.

3.There is no multicollinearity.

4.There are proportional odds.

The first two assumptions are already taken care of through the nature of our model. It is clear that our dependent variable, Medal, is ordered. We have continuous and categorical

independent variables (sex and country being categorical, height, weight and age being continuous). For the multicollinearity assumption, we will be examining correlation plots between each variable and Variance Inflation Factor (VIF) tests to ensure multicollinearity does not exist.

In cleaning our dataset, we removed all the NA values corresponding to our variable, Medal. There were NA values corresponding to our Medal variable because data were Missing at Random (MAR). We know that, for each event, there can only be three placements: Gold, Silver and Bronze. This is our observed data, because we have a variable corresponding to the event. Our missing data are all the NA values for medal that we initially had in our dataset. Therefore our missing data are related to observed data, because we know those who didn't receive a medal within their events (NA value) just weren't able to place within the three placements that are offered for each event.

We fit an ordinal model with no transformations or interactions and tested the assumptions to see how it would perform as an initial model; something we could base our improvements off of. This ordinal model included our variables of interest, with Medals as our outcome. We first ran LASSO on this model, selecting important variables from the output LASSO gave. Then, we ran all-subset selection on the variables that were chosen by our LASSO output. We opted not to use forward selection or backward elimination because we knew our variables were highly correlated, which we cover in the discussion of our transformations and interactions.

The predictor variables we are considering for the model are sex, age, height, weight and NOC (country). We considered interactions terms between sex * weight and sex * height.

Exploratory Data Analysis

```
olympics04 %>%
  count(medal) %>%
  mutate(per = n/sum(n)) %>%
  ggplot()
```

```
olympics08 %>%
  count(medal)%>%
  mutate(per = n/sum(n))
```

```
# A tibble: 3 x 3
  medal      n   per
  <chr>  <int> <dbl>
1 Bronze   706 0.347
2 Gold     664 0.326
3 Silver   665 0.327
```

```
olympics12 %>%
  count(medal) %>%
  mutate(per = n/sum(n))
```

```
# A tibble: 3 x 3
  medal      n   per
  <chr>  <int> <dbl>
1 Bronze   669 0.349
2 Gold     622 0.325
3 Silver   624 0.326
```

```
olympics16 %>%
  count(medal) %>%
  mutate(per = n/sum(n))
```

```
# A tibble: 3 x 3
  medal      n   per
  <chr>  <int> <dbl>
1 Bronze   700 0.348
2 Gold     662 0.329
3 Silver   652 0.324
```

In 2004, the number of bronze medals handed out to individuals was 676 which was 33.8% of the total medals, the number of silver medals was 660 which was 33% of the total medals , and the number of gold medals was 664 which was 33.2% of the total medals .

In 2008, the number of bronze medals handed out to individuals was 706 which was 34.7% of the total medals , the number of silver medals was 665 which was 32.7% of the total medals, and the number of gold medals was 664 which was 32.6% of the total medals.

In 2012, the number of bronze medals handed out to individuals was 669 which was 35% of the total medals , the number of silver medals was 624 which was 32.6% of the total medals, and the number of gold medals was 622 which was 32.4% of the total medals.

In 2016, the number of bronze medals handed out to individuals was 700 which was 34.8% of the total medals, the number of silver medals was 652 which was 32.3% of the total medals, and the number of gold medals was 662 which was 32.8% of the total medals.

```
ggplot(olympics , mapping = aes(x = weight , y = sex )) +
  geom_boxplot() +
  theme_minimal() +
  labs(x = "Weight", y = "Sex", title = "Distribution of weights of
       athletes by sex", subtitle = "Men have a higher median weight
       than women across all olympic games")
```