Introduction

Due to the COVID-19 outbreak, the 2020 Olympic Summer Games were postponed for the first time in history. As some members of our group are college athletes themselves, they understand the training, preparation, and heartbreak that occur behind the scenes in this type of situation. As we began this project, we hoped to understand what factors might impact Olympic outcomes. We initially chose an Olympic medal dataset which includes information about the Olympic games from 1896 to 2016. It provides information on all the athletes participating in the Olympics, the countries they represent, events they participate in, years they appear in the Olympics, and medals they win.

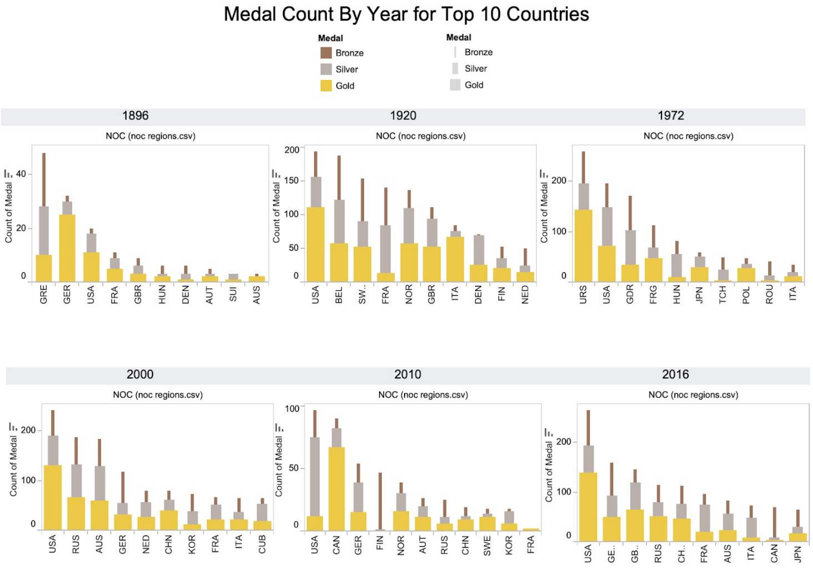
In the first stage of analysis, we answered several different questions about the dataset through various data visualizations. These questions included:

1. Is there a home field advantage? Does the country who hosts win more than usual?
2. Which sports tend to have medal winners from the same country every year and which sports tend to have a greater variety of winners?
3. How do the top medaling countries 100 years ago compare to the top countries 50 years ago? 20 years ago? 10? Now?
4. Do athletes in different sports peak in performance at different times in their lives?

However, as we continued in our analysis we were drawn to analyzing which factors impact how many Olympic medals a country wins. In order to leave space to explain this analysis, we will share our visualization and answer to question 3, as it is most relevant to the current question. If you are interested in knowing the answers to all four questions we addressed previously, you can find the report here.

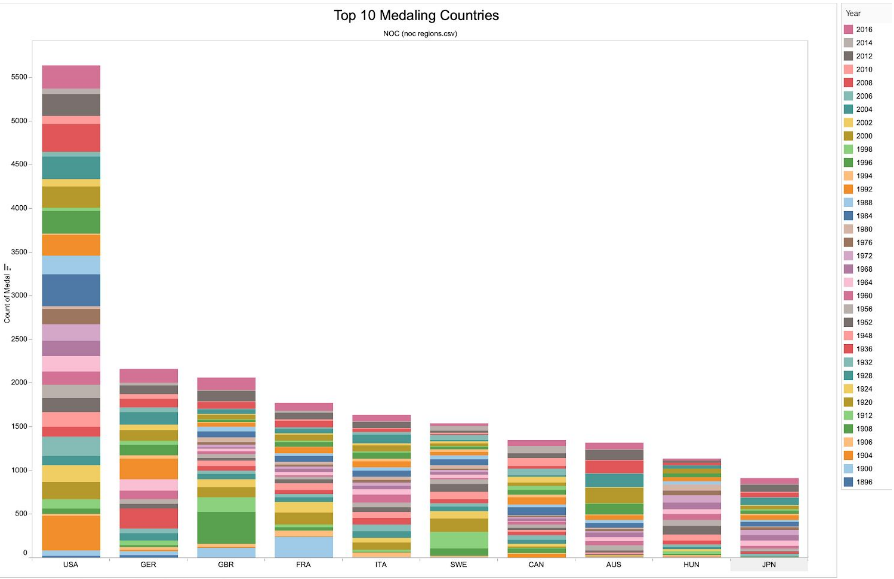
Data Visualization: Top Medaling Countries Over Time

We created a bar chart and filtered the countries to only show the top 10 medaling countries each year. We then made the different bar charts into a dashboard to allow the viewer to compare side by side the top ten medaling countries from 1896 onward.



***NOTE: chart represents counts and differences between countries in the year. So shape and distribution can be compared across charts, but not size of bar (due to difference scales on vertical axis!)*

From the visualization, one can see the top five medaling countries each year were different, except with the USA consistently being one of the top medaling countries. Therefore, the top medaling countries may shift over time but it is likely that the USA will be one of the top five medaling countries. The visualization also shows the total medals distribution by country and the number of gold medals does not predict overall total medal count placement.



We chose to visualize this question further with another bar graph (seen below) with a cumulative medal count throughout all the years. In this visualization, the years are different colors, the countries are sorted in descending order by total medal count, and the horizontal axis is filtered by the top fifteen medaling countries overall.

From the visualization, one can immediately see that the USA is the highest medaling country overall. This supports our other visualization where we saw that the USA was one of the top five medaling countries each year. The current visualization shows diversity in wins across countries, but consistency in top results and participation-similar countries are consistently in the top 10 medaling countries with some movement around the top, except for the USA. This second visualization is particularly pertinent as we shift to focus on which factors impact how many medals a country wins in the following data analysis portion of the paper.

Data Analysis

As we transitioned our focus from exploration of the data set in the visualization stage to analysis of which factors impact the number of medals a country wins, we faced several design decisions. First, we chose to focus on data from the 2016 Summer Olympics in Rio, Brazil. We chose to focus on this year for two reasons. First, it is recent enough that drastic changes in countries' indicators (GDP, life expectancy, etc) worldwide will be minimized. Secondly, this Olympic games followed the Ebola crisis, providing unique parallels to the upcoming games that will be following the COVID-19 pandemic.

Because our initial dataset was confined to data about the games themselves, we gathered data and built our own dataset that included more detailed information about each country that participated in those games. We pulled data from several different sources (see bibliography) for a variety of factors that could impact how many medals a country wins. Those columns in the dataset are: population, GNI, GDP, inflation rate, life expectancy, fertility rate, social development (HDI), Academy Awards for Best International Feature Film from that country that year, number of athletes that participated in the games, number of times they have hosted, political system, type of head of state.

Our analysis will be split into 5 different parts. We will run a linear regression, check that assumptions about the residuals are satisfied, check for collinearity, build an AIC model using cross validation techniques, and finally forecast future medal wins using the original dataset. The code used throughout this analysis can be referenced in this python file.

Linear Regression

Linear regression is a process through which one can try to explain changes in a response (dependent) variable based on one or more explanatory (independent) variables. This process is primarily done with numerical predictors but categorical data can also be used to explain changes between groups. Linear regression uses least squares estimation to fit a model to a set of data. What this means is that we try to minimize the difference between the actual values within the model and the values we would predict with a regression equation of the form:

$$\hat{y}_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_n x_{ni}$$

Where:

- $\hat{y}_i$ represents the ith predicted observation
- β_0 is the intercept of the equation, the predicted value for y if all variables are equal to zero
- β_1 ,..., β_n are the coefficients of our n predictor variables. Each can be positive or negative and represents the change in our prediction based on a 1 unit increase in value for each $x_1, ..., x_n$ within our model. In the case of categorical predictors they will represent change in intercept moving from one level of the category to another rather than change in slope.

In the case of our model specifically, we try to model the number of medals a country will win in a given year (our response/dependent variable) by a variety of numerical and categorical factors. We identified these factors based on the fact that they may be indicators of development within a country which means the country can focus more resources on the games and thus win more medals. We also chose some factors which we think would have nothing to do with medal wins (like number of academy awards) to demonstrate what to do with an insignificant predictor in linear regression.

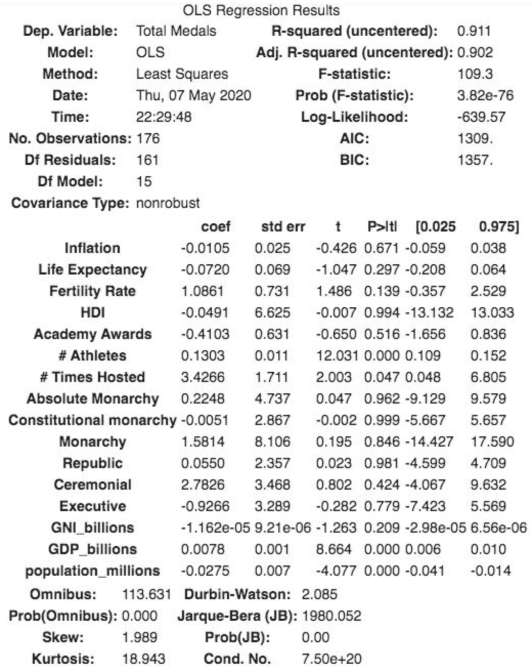
The explanatory variables we chose are as follows:

1. Population. We expressed population in millions so that it is more comparable to other values in our dataset..
2. GNI (Gross National Income), the total amount of money earned by a nation's people and businesses. We expressed GNI in billions to make the values more comparable to others in the dataset.
3. GDP (Gross Domestic Product), the total monetary or market value of all the finished goods and services produced within a country's borders. We chose to express GDP in billions to make the values more comparable to others in the dataset.
4. Inflation rate, the percentage increase in the Consumer Price Index from year to year.
5. Life expectancy, the average time people are expected to live within a country.
6. Fertility rate, the average number of children per woman in each country.
7. Social development (HDI), a summary measure of average achievement in key dimensions of human development: a long and healthy life, being knowledgeable and having a decent standard of living.
8. Academy Awards for Best International Feature Film from that country that year. This is meant to be an insignificant variable.
9. Number of athletes that participated in the games, as the title implies.
10. Number of times they hosted in the past, as the title implies.
11. Political system, a categorical predictor for each country which is either Republic, Constitutional Monarchy, or Absolute Monarchy.
12. Type of head of state, a categorical predictor for each country with either executive or ceremonial. It explains if the head of state is ceremonial (Queen Elizabeth) or has actual power in the government (Donald Trump).

We chose predictors 1 to 7 and 9 and 10 (above) as they related to economic prosperity and may be an indicator of how much resources a country has to dedicate to the games. For predictor 8, we wanted to see what an unrelated variable would be in our model. For predictors 11 and 12, we wanted to include a categorical predictor within our model in addition to seeing if politics had any influence on medaling.

Methods:

To evaluate the relationships between these variables, we used LinearRegression packages available in sklearn. This allows us to fit our data very easily. One major consideration we had to make was coding up our categorical variables as 0-1 dummy variables which allows them to be included in a regression equation and interpreted more easily. Upon adjusting the data and fitting the model we can see the following output on the left. From this we can make a few initial observations. First, our adjusted R-squared value is .902. This means that roughly 90% of the variability in the number of medals a country wins is accounted for by variables within our model! It may be the case, however, that the model is overfit, especially since adjusted R-squared is lower than normal R-squared. Additionally, we have an F-Statistic of 109.3 and a corresponding F-probability of 3.82×10^{-76} which indicates that some of our predictors are in fact significant. We can even look at the p-values of some predictors like number of athletes that participate, GDP, and population and see that they are close to zero, indicating they may be significant. Others like the type of government or the Academy award wins per country have p-values of close to 1, indicating they are likely insignificant. Before we make any conclusions, however, we must check the assumptions associated with linear regression in order to ensure our findings are valid.

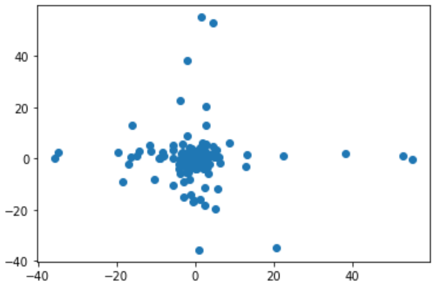


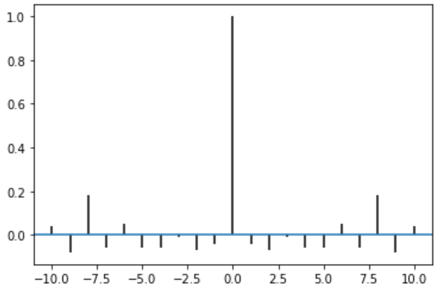
```
                         OLS Regression Results
   Dep. Variable:      Total Medals        R-squared (uncentered):      0.911
       Model:          OLS             Adj. R-squared (uncentered): 0.902
      Method:          Least Squares              F-statistic:           109.3
        Date:          Thu, 07 May 2020        Prob (F-statistic):       3.82e-76
        Time:          22:29:48                Log-Likelihood:           -639.57
No. Observations:      176                          AIC:                 1309.
   Df Residuals:       161                          BIC:                 1357.
     Df Model:         15
Covariance Type:       nonrobust
                          coef      std err    t      P>|t|  [0.025     0.975]
        Inflation       -0.0105    0.025    -0.426  0.671 -0.059     0.038
     Life Expectancy    -0.0720    0.069    -1.047  0.297 -0.208     0.064
      Fertility Rate     1.0861    0.731     1.486  0.139 -0.357     2.529
           HDI          -0.0491    6.625    -0.007  0.994 -13.132    13.033
     Academy Awards     -0.4103    0.631    -0.650  0.516 -1.656     0.836
        # Athletes       0.1303    0.011    12.031  0.000 0.109      0.152
      # Times Hosted     3.4266    1.711     2.003  0.047 0.048      6.805
   Absolute Monarchy     0.2248    4.737     0.047  0.962 -9.129     9.579
Constitutional monarchy -0.0051    2.867    -0.002  0.999 -5.667     5.657
        Monarchy         1.5814    8.106     0.195  0.846 -14.427    17.590
        Republic         0.0550    2.357     0.023  0.981 -4.599     4.709
       Ceremonial        2.7826    3.468     0.802  0.424 -4.067     9.632
       Executive        -0.9266    3.289    -0.282  0.779 -7.423     5.569
      GNI_billions      -1.162e-05 9.21e-06 -1.263  0.209 -2.98e-05 6.56e-06
      GDP_billions       0.0078    0.001     8.664  0.000 0.006      0.010
   population_millions  -0.0275    0.007    -4.077  0.000 -0.041     -0.014
    Omnibus:       113.631  Durbin-Watson:    2.085
Prob(Omnibus):     0.000    Jarque-Bera (JB): 1980.052
     Skew:         1.989        Prob(JB):     0.00
   Kurtosis:       18.943       Cond. No.     7.50e+20
```

Check Assumptions of Linear Regression

The statistics that we calculated above are only valid if the following assumptions are satisfied. The residuals must be:

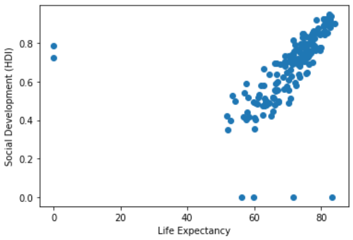
1. Mutually independent: we will check for mutual independence in two different ways first by plotting errors and the other by using the autocorrelation function.

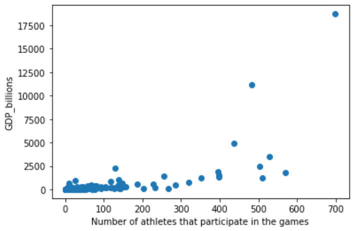


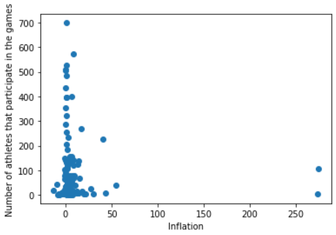


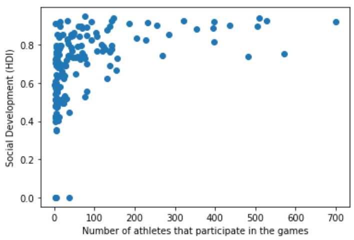
The above graph to the left demonstrates mutual independence since there's no pattern. The same result is shown on the graph on the right by using the autocorrelation functions and since very few autocorrelations are outside the test bounds this again indicates mutual independence.

2. Independent of the covariates: we will check for independence of the covariates by ensuring there is no relationship between the residuals and X. This means the graphs for each of the covariates should show no pattern which indicates it is linear in the predictors and therefore independent of the covariates. Since we had 12 covariates in this dataset, 144 plots were generated (we will show 4).



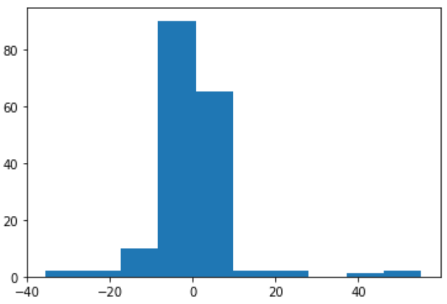


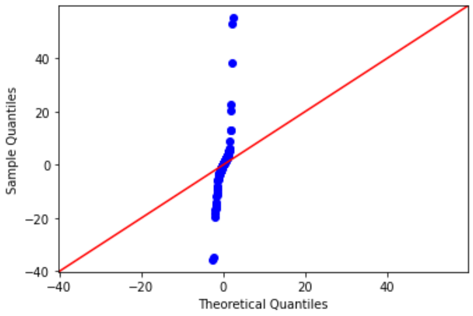




As you can see from above not all the covariates are independent of one another. For example, the upper left plot shows life expectancy versus social development (HDI); rather than randomized plot, there is a linear trend. Due to this result we will need to test for collinearity in our covariates and determine which of the covariates are significant and which are not.

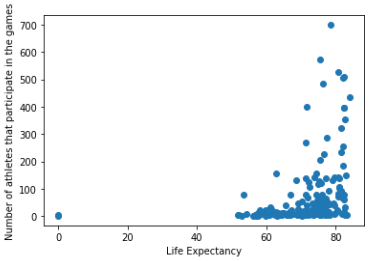
3. Normally distributed: in order to check that our residuals are normally distributed we will construct a Q-Q plot. If the Q-Q plot is nearly a straight line the residuals are normally distributed; if not then a pattern exists such as skewness or heavy-tails.

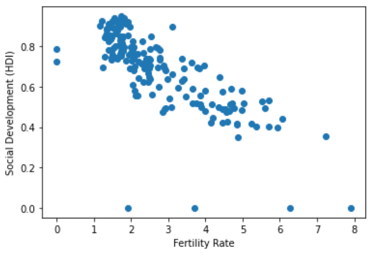


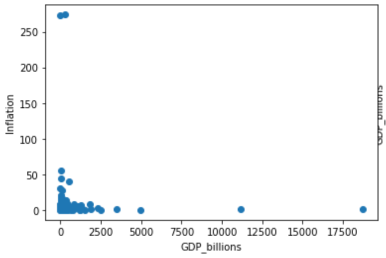


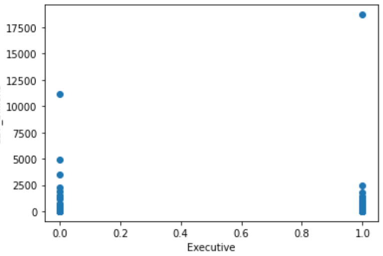
As you can see above, the residuals are not normally distributed, because the Q-Q plot is concave-convex which means the data is heavy-tailed. While having a normal distribution of errors is vital when making predictions, it is not necessary to estimate coefficients in an ordinary least squares regression since the error distribution averages out in the least squares estimate.[1]

4. Have constant variance: to check for constant variance we will plot the absolute value of the residuals versus the fitted values. A constant variance is indicated if the distribution does not depend on the predictors.









As seen above in many instances such as Fertility Rate vs. HDI or Executive government vs. GDP the distribution does depend on the predictors. Therefore there is a non-constant variance or “heteroscedasticity.”

Because three of the four assumptions were violated, we know that some of the selected covariates are dependent in some way on other covariates. Our next step is to employ methods of model selection to find a model that does not violate the assumptions and can more accurately represent the relationships in the data. The next two sections will address this by selecting a model through cross validation and minimizing the Aikake Information Criteria (AIC).

Cross Validation

Because our data set is a limited data sample, we decided to use cross validation to retain the integrity of the statistical measures reported by our model. Cross validation is a resampling procedure used to assess how the results from a statistical analysis will generalize to an independent data set.[2] This is done by splitting the data set into a training set and testing set before you begin. Test sets are usually around 20% of the original data set, something we replicated when we split our data set. The training data is then used to select

[1] Vasishth. “Statistical Modeling, Causal Inference, and Social Science.” *Statistical Modeling, Causal Inference, and Social Science:* 4 June 2007, statmodeling.stat.columbia.edu/2007/06/04/when_to_worry_a/.

[2] “Cross-Validation (Statistics).” *Wikipedia*, Wikimedia Foundation, 23 Apr. 2020, en.wikipedia.org/wiki/Cross-validation_(statistics).

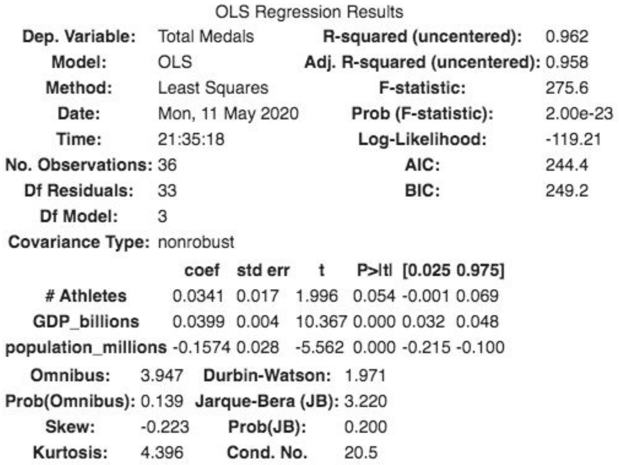
and fit the model. Finally, the testing data is used to predict the independent variable. These predictions are then compared against the actual values of the independent variable in the testing data to assess the quality of fit of the model. This process allowed us to draw insight from the p-values of the model. To summarize, the process is delineated in the following three steps:

1. Split the data into a training set (80%) and testing set (20%)
2. Use the training set to select and fit a model
3. Use the testing data to predict outcomes and assess the quality of fit

We split our data into training and testing sets and then used the training set to fit a model. During step 2, or model selection, we used the method of minimizing the AIC by dropping variables with big p-values, as explained in the next section.

Model Selection

The AIC is a common way to report how well the model represents the data. This method of model selection minimizes the AIC by dropping variables with big p-values, leaving a model with only statistically significant variables. Using the linear regression model calculated above, we fit the model using the training data, dropped the variables with large p-values, refit the model, and recalculated the AIC for the new model. If the new model had a lower AIC than the previous model, we took the new model and repeated this process. This process continued to repeat until only variables significant at the $\alpha = 0.5$ level remained. We then used only these statistically significant variables and refit the model on the testing data, as outlined in step 3 of cross validation. The statistics from the model built from the testing data and significant variables are reported in the image to the right. The p-values suggest that the variables significant at the 0.05 level are the number of athletes, GDP, and population.



```
                        OLS Regression Results
  Dep. Variable:    Total Medals        R-squared (uncentered):     0.962
        Model:      OLS            Adj. R-squared (uncentered):     0.958
       Method:      Least Squares                F-statistic:       275.6
         Date:      Mon, 11 May 2020       Prob (F-statistic):      2.00e-23
         Time:      21:35:18                  Log-Likelihood:       -119.21
No. Observations:   36                                  AIC:        244.4
   Df Residuals:    33                                  BIC:        249.2
     Df Model:      3
Covariance Type:    nonrobust
                      coef   std err    t     P>|t|  [0.025  0.975]
     # Athletes      0.0341  0.017    1.996   0.054  -0.001  0.069
    GDP_billions     0.0399  0.004   10.367   0.000   0.032  0.048
population_millions -0.1574  0.028   -5.562   0.000  -0.215  -0.100
    Omnibus:       3.947   Durbin-Watson:    1.971
Prob(Omnibus):     0.139  Jarque-Bera (JB):  3.220
      Skew:       -0.223     Prob(JB):       0.200
    Kurtosis:      4.396     Cond. No.       20.5
```

Initially, our original linear regression model included all 12 variables and had an AIC of 1309. After using cross validation and the methods of model selection described above, we found a new model that used only 3 variables and had an AIC of 244.4. This new model will hereafter be referred to as the AIC model. The AIC model summary seen above is expressed by this equation:

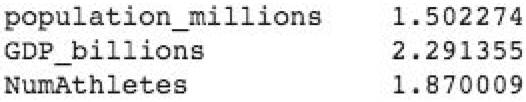
$$\# \, medals = 0.0341(\# \, athletes) + 0.0399(GDP \ in \ billions) - 0.1574(population \ in \ millions)$$

Check Assumptions

Because we found a new model, we decided to recheck the assumptions to understand if the statistics found in the above AIC model were valid. The analysis followed the same steps outlined above in our previous discussion of checking assumptions. Under the new AIC model, the residuals were mutually independent and had constant variance. The 3 covariates were also independent from each other. The only assumption that was violated was that the residuals were once again heavy-tailed. However, as explained earlier, this should not impact the ability to find regression coefficients. Because we had found that some of the covariates of the first linear model were dependent on each other, we decided to double check the independence on the covariates in the new AIC model with a second method. This method of checking collinearity using variance inflation factors (VIF) will be explained in the next section.

Collinearity

Collinearity means high correlations between the predictors, so if two predictors are highly correlated, then it is difficult to separate their effects on the response variable.[3] When the effects are difficult to separate it can be hard to decide which variable is important and it can lead to uninterpretable models with too many features or the wrong features.[4] Thus, collinearity is a problem for inference, figuring out which variables are important, but not a problem for prediction. Luckily, collinearity can be detected with variance inflation factors (VIF). VIF is the increase in variance due to collinearity. If the VIF is greater than ten, then the covariates are likely to be collinear, so the smaller the VIF the better. We began by checking the VIF for all the variables in the new AIC model. The image to the right gives the VIF values for each of the covariates in the AIC model.

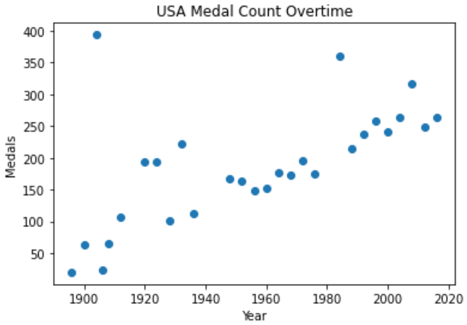


```
population_millions    1.502274
GDP_billions           2.291355
NumAthletes            1.870009
```

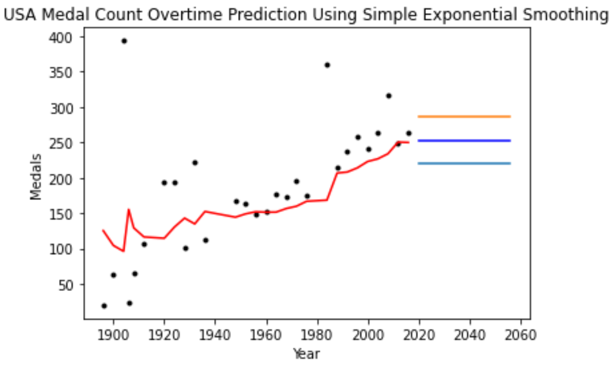
A VIF value near 1 indicates that there is not a problem with collinearity. If the VIF is greater than 10, there may be a problem with collinearity. As you can see in the image above, the VIF for each of the variables is small, so these variables are not significantly collinear. Since these variables are not significantly collinear, we did not try to change anything with orthogonality. Thus, we conclude our model selection portion with the AIC model explained above.

Forecasting

After evaluating what factors we found to have the most profound impact on Olympic performance based solely on the 2016 Olympic data, we decided to forecast the number of medals the United States would win based on its medal counts from 1896-2016. Because the USA boycotted the Olympics in 1980 due to the political landscape at the time, we dropped the data from that year in this analysis. We will forecast the United States number of medals for what is now the 2021 Summer Olympic Games using two methods: Simple Exponential Smoothing (a forecasting method without a trend) and the Holt Method(a forecasting method with a linear trend). The left plot shows a scatter plot of the USA Medal Counts overtime.

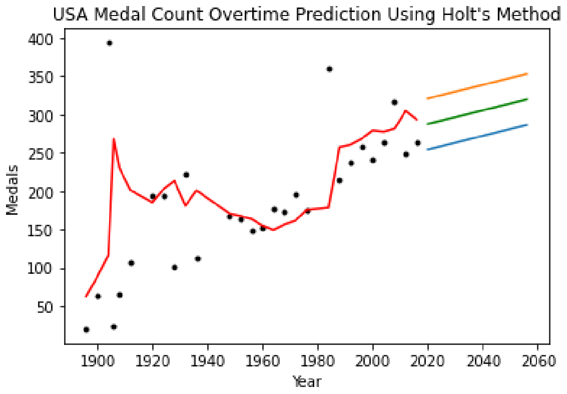


We began this analysis by fitting the data using simple exponential smoothing. The method of exponential smoothing is applicable to this dataset in that in its prediction it is a weighted sum of past observations and the model explicitly uses an exponentially decreasing weight for past observations. That being said, the medal counts from the 1896 Olympics where there were less events and therefore less medals to be won will not be weighted as heavily as past Olympics like 2016. We optimized the model instead of selecting a smoothing level and then forecasted based on the fitted values for this model for the next 10 Olympic games as well as computed a 95% confidence interval for the forecast (which can be seen on the right).



The red line represents our results from a simple exponential smoothing fit and the blue line is the forecast for the USA future medal counts. Based on this model the predicted medal count for the USA in the 2021 Olympic Games is 252 medals, which is less than the Olympics in 2016 (264 medals). As you can see that although the model does not weigh all past events equally, past events such as the USA hosting the 1984 Summer Olympics had a significant impact on the forecast. Therefore we decided to try a different forecasting method: Holt's Method.

[3] Class Slideshow
[4] Class Slideshow



In Holt's method we assume the data had a trend but no seasonality. This is a double exponential smoothing method because we introduce another term to take into account the possibility of a series exhibiting some form of trend. Similarly to above, we optimized the model instead of selecting a smoothing level and then forecasted based on the fitted values for this model for the next 10 Olympic games as well as computed a 95% confidence interval for the forecast (seen to the right).

The red line represents our results from a simple exponential fit and green line is the forecast for the USA future medal counts. Based on this model the predicted medal count for the USA in the 2021 Olympic Games is 287 medals, which is larger than the number of medals won in the 2016 Olympic Games. This forecasting method appears to be better for our dataset since it does show a linear trend despite a few outliers. It is interesting to note that the most significant of these two outliers, in 1904 and 1984, represent unusually high medal wins in years that the USA hosted the Olympics. In a future exploration, it could be interesting to explore adding a categorical "hosting" term to the forecast to be better able to predict these outliers. For more information on the impact of hosting, read our initial data visualization report**.**

Conclusion

Over the course of this project, we began with a multitude of questions regarding data from the Olympics and narrowed down to just one: What factors can explain the number of medals a country wins in an Olympic games? Although many of the factors included in this analysis had an impact, the most significant were GDP in billions, population in millions, and the number of athletes that participated in the games from that country. To come to this conclusion, we looked at medaling data from 2016 from our original dataset and built a linear regression model around it. We gathered data from various sources on predictors that we hypothesized might be able to explain some of the variability we see in the number of medals that a country wins in a given year. All data analyzed was from 2016.

To answer this question, we decided on using a Linear Regression model. In other words, we were trying to predict the number of medals won by a country based on the factors we identified. After initially fitting our model, we could do some preliminary analysis of our predictors but needed to check the assumptions and fit of our model in order to validate our conclusions. The statistics calculated from our model would only be valid if all four assumptions held. However, after checking our assumptions we found that the residuals were mutually independent, but not independent in their covariates, not normally distributed and did not have a constant variance. To address this we decided to use model selection to find a model that did not violate the assumptions. We decided to use the method of model selection that minimizes the Akaike Information Criteria (AIC). This is accomplished by iteratively dropping variables with large p-values, simplifying the model and minimizing the AIC. After running this process, we found a model with only 3 covariates: GDP in billions, population in millions, and the number of athletes that participated in the games from that country. We checked the assumptions again and the only assumption that did not hold was that our residuals were not normally-distributed. However, this is only a problem for prediction and does not impact the ability to find regression coefficients. After checking the covariates for the new AIC model, we concluded that the variables were not significantly collinear. Therefore, we decided to conclude with the sound AIC model. Lastly, we forecasted the medal count for USA for the 2021 Summer Olympics using simple exponential smoothing and Holt's method.

Bibliography

Datasets:

https://www.kaggle.com/ammon1/demographic

- Description: Used to source population, inflation, life expectancy, and fertility rate.

https://www.kaggle.com/sudhirnl7/human-development-index-hdi#HDI.csv

- Description: Used to source Human Development Index.

https://www.kaggle.com/fmejia21/demographics-of-academy-awards-oscars-winners

- Description: Used to source Academy Awards for Best International Feature Film.

https://www.kaggle.com/lewisduncan93/the-economic-freedom-index

- Description: Used to source GDP.

https://www.kaggle.com/heesoo37/120-years-of-Olympic-history-athletes-and-results

- Description: The original dataset. Contains 120 years of Olympic history ranging in data points from the names of athletes and their country of origin to their heights and weights. Used to source the number of athletes that participated in the games and the number of times each country hosted.

Additional Resources:

https://statmodeling.stat.columbia.edu/2007/06/04/when_to_worry_a/

- Description: Used in explanation of the significance (or lack thereof) of heavy-tailed errors.

https://en.wikipedia.org/wiki/Cross-validation_(statistics)

- Description: Used in explanation of cross validation.

Other Reports:

Milestone 1: Data Visualization

- Description: The initial report answering questions about the Olympic medal data using Tableau visualizations.

Milestone 2: Python File

- Description: The python code used to create and analyze the various models put forth in this report.