#—-Trackdown Instructions—-# This is not a common Document. The Document includes properly formatted Markdown syntax and R code. Please be aware and responsible in making corrections as you could break the code. Limit changes to narrative text and avoid modifying R code. Once the review is over accept all changes: Tools -> Review suggested edits -> Accept all. You must not modify or remove these lines, we will do it for you ;) FILE-NAME: project_report.Rmd HIDE-CODE: FALSE #—-End Instructions—-# — title: "STAT 306 (S2024 T2): Final Project (Group B2)" author: - "Jordan Bourak (48510689)" - "Jeffrey Xiong (24893570)" - "Chirag Kharade (84784602)" - "Maahi Gumber (76901438)" output: pdf_document: number_sections: true tables: true —

Introduction

Motivation

Because of the current 2024 Summer Olympics in Paris, we were interested in seeing if we could predict the total number of medals a country would win in the 2024 Summer Olympics by referring to its medal counts as well as other associated data (for example, the country's athletes' average age) in previous summer games. In the end, we decided to choose Canada as the country for our prediction.

Data Source

The data were downloaded manually from Kaggle (Olympic Historical Dataset From Olympedia.org). The user who uploaded the dataset, Joseph Cheng, extracted the data from olympedia.org.

Raw Data Exploration & Manipulation

The data come in six separate files. Four of these files were combined together for our analysis:

- `Olympic_Athlete_Bio.csv`: contains information about all athletes that have competed in the Olympics, including sex, date of birth, and a unique identifier
- `Olympic_Athlete_Event_Results.csv`: contains information about every event that each athlete has participated in, including whether they received a medal and a unique identifier for the athlete, result of the event, and the edition of the Olympics that event was held in.
- `Olympics_Games.csv`: contains information about each edition of the Olympic Games, including the start date and a unique identifier.
- `Olympic_Games_Medal_Tally.csv`: contains medal counts per country for each edition of the Olympic Games.

Medal counts for Canada were extracted from the medal tally dataset. The athlete bio, athlete event results, and Olympic games information datasets were joined together to calculate descriptive measures about Canada's athletes and participation for each edition of the Olympic Games.

The start date of each Olympic games was extracted from the Olympic games information dataset and converted to a decimal of the year.

Entry counts were calculated by counting the number of entries by Canadian athletes and athletes from other countries, for events that Canadian athletes participated in.

Athlete age at the time of event was computed by taking the difference in time (in years) between their date of birth and the start date of the edition of the Olympic Games the event was held in, and then the average was taken, weighted by the number of events an athlete participated in. A small proportion of Canadian athletes had missing values for date of birth (no more than 6% except for the 1904 Summer Olympics with 12.5%) and so were excluded in the calculation for average age.

Female participation was computed by taking the ratio between number of female athlete entries and the total number of athlete entries. Only one Canadian athlete had a missing value for sex and so they were excluded in the calculations.

Variables of Interest

Name	Purpose	Type	Description
Date	explanatory	numerical	The start date of an edition of the Summer Olympic games, represented as a decimal of the year.
Entries	explanatory	numerical	The number of entries in events by Canadian athletes.
Other Entries	explanatory	numerical	The number of entries by other countries in events that Canadian athletes competed in.
Female Participation	explanatory	numerical	The proportion of Canadian athletes competing that are female.
Average Age	explanatory	numerical	The average age of the Canadian athletes competing in years.
Medals	response	numerical	The number of medals awarded to Canadian athletes.

Research Questions

Given the ongoing 2024 Summer Olympics, we are interested in whether the number of medals won by Canada can be predicted using historical data. Specifically, we aim to answer the following question:

Can we predict Canada's Medal count in the 2024 Summer Olympics using the start date of the Olympics, Canada's number of entries in events, entries by other countries in those events, and the average age and proportion of females among Canadian athletes?

Analysis

Preliminary Data Exploration

Table 1: Pairwise sample correlations between variables

	medals	date	entries	other_entries	avg_age	prop_female
medals	1.000	0.460	0.591	0.449	0.181	0.502
date	0.460	1.000	0.823	0.913	-0.078	0.977
entries	0.591	0.823	1.000	0.926	-0.198	0.800
other_entries	0.449	0.913	0.926	1.000	-0.205	0.881
avg_age	0.181	-0.078	-0.198	-0.205	1.000	-0.052
prop_female	0.502	0.977	0.800	0.881	-0.052	1.000

Figure 1 suggests that there is a moderate relationship between number of medals and every explanatory variable, except average age, which appears to be mostly random; the pairwise correlations in Table 1 agree. Furthermore, the number of Canadian entries, number of entries by other countries, female representation, and start date of the Olympics are all quite strongly correlated with each other. Female representation and

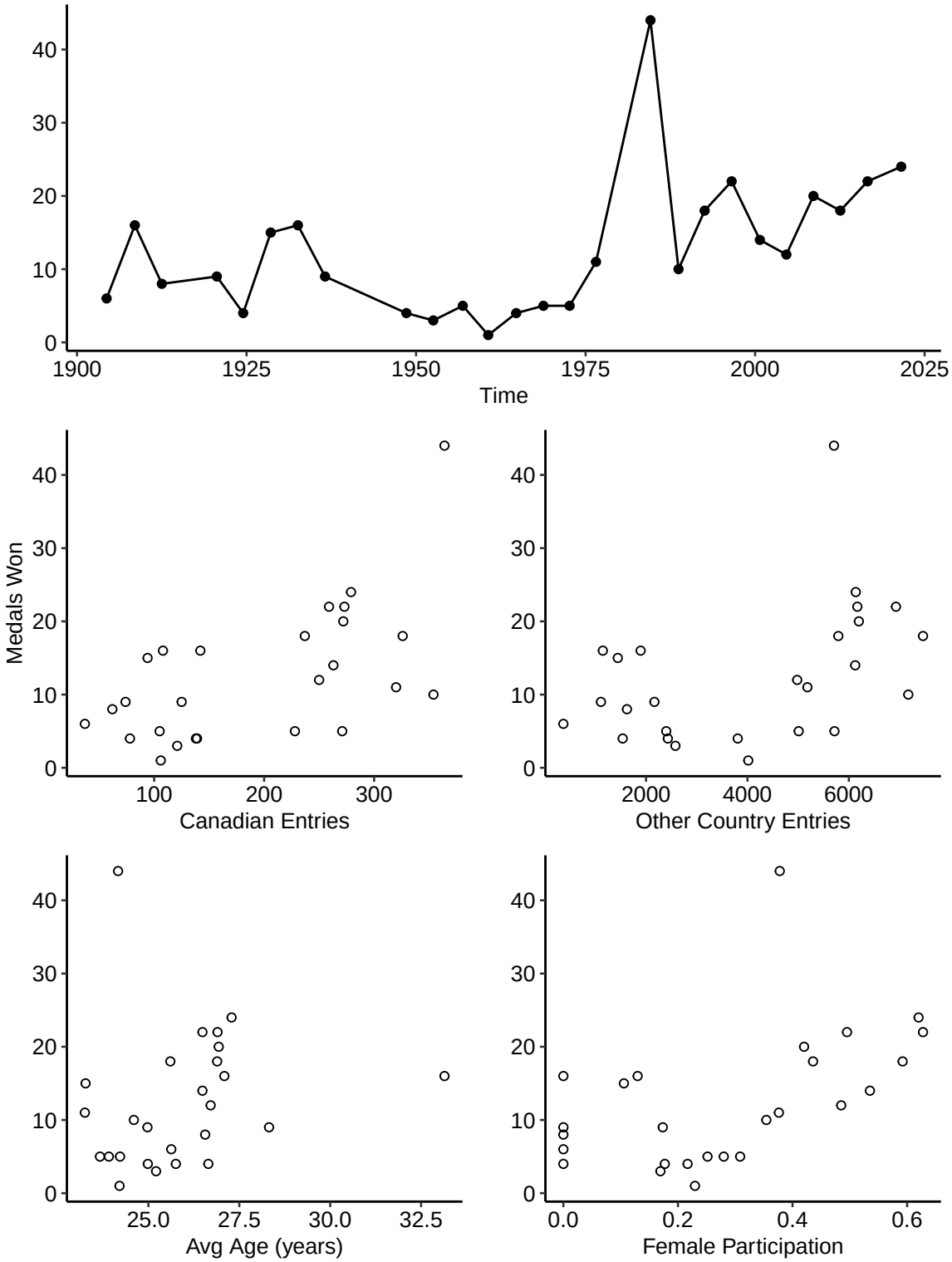

Figure 1: Number of medals won across explanatory variables

start date have an especially high correlation, strongly suggesting that one of them should be excluded to avoid issues associated with multicollinearity. There is one observation where the number of medals won by Canada stands out significantly from the rest of the Olympic games.

Regression Analysis

Due to the time series nature of our data, using techniques such as leave-one-out cross-validation or a variation of using training/holding sets to assess a model's out-of-sample predictive ability is not as straightforward. With only a small number of observations available (26), results from these methods could also potentially be misleading due to the small number of observations available, or not very meaningful if we use a small number of iterations. Hence, two different approaches were taken to arrive at two different models that may be suitable for prediction. Firstly, we choose a model using intuition and prioritizing interpretability. Secondly, we choose a model by choosing a model from among all subsets of the explanatory variables through an exhaustive search.

Intuitive and Interpretable Model

Given that female representation and date are nearly perfectly correlated, we will preemptively drop female representation to avoid multicollinearity issues. Although female representation has a slightly higher sample correlation with medal count, we discard it rather than date as the data are inherently a time series. Hence, as a first attempt, we fit a linear model with all explanatory variables except female representation (Model 1).

Table 2: Regression coefficients for Model 1, $R^2 = 0.516$, $adjR^2 = 0.424$, $\hat{\sigma} = 7.055$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-178.762	186.936	-0.956	0.350
date	0.078	0.100	0.778	0.445
entries	0.120	0.038	3.184	0.004
other_entries	-0.004	0.002	-1.674	0.109
avg_age	1.202	0.739	1.625	0.119

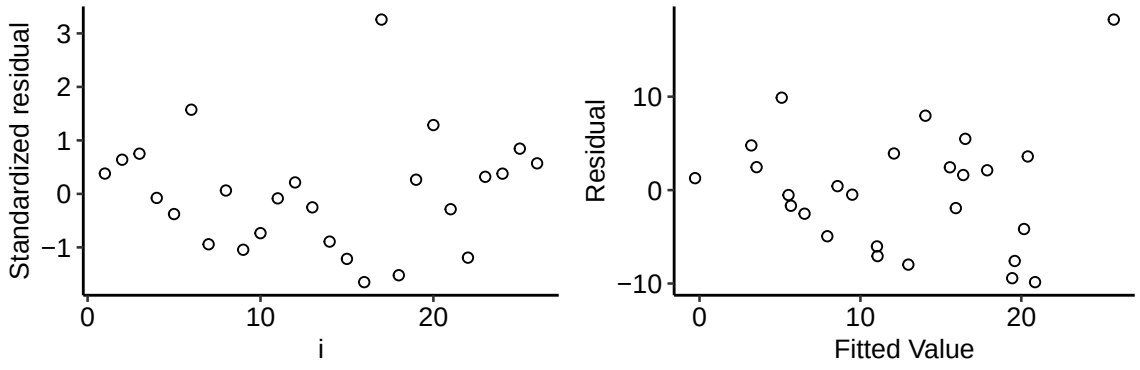

Figure 2: Residual plots for model 1

The standardized residuals are plotted (in order) to investigate the potential "outlier" mentioned earlier and potential serial correlation, as our data are from a time series.

With a sample size of 26, a standardized residual over 3 is concerning. The observation that this residual is associated with is from the 1984 Summer Olympics, which were boycotted by fourteen countries, including the Soviet Union and East Germany, who usually earned a significant proportion of the medals around that time. Because we are only interested in predicting the medal count for the 2024 Summer Olympics, the large standard residual combined with the suggestion of confounding due to the boycotting gives us sufficient reason to drop this edition of the Olympic Games from our sample.

There is potentially some serial correlation present in the data, as the ordered standardized residual plot appears to oscillate over time. We will revisit this potential concern after a refitting of the model.

Intuitively, it may also be worthwhile to add interaction between date and the number of Canada's entries to capture a change in Canadian athlete's overall performance over time. We ignore interaction between date and the number of entries by other countries, as the change in athlete performance over time for every other country combined will likely average out to be negligible.

Hence, we fit a second model (Model 2) with the same explanatory variables, but adding interaction between date and entries and removing the observation associated with the 1984 Summer Olympics.

Table 3: Regression coefficients for Model 2, $R^2 = 0.635$, $adjR^2 = 0.539$, $\hat{\sigma} = 4.653$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	186.082	196.943	0.945	0.357
date	-0.109	0.101	-1.075	0.296
entries	-2.019	0.932	-2.166	0.043
other_entries	-0.001	0.002	-0.519	0.610
avg_age	1.255	0.521	2.408	0.026
date:entries	0.001	0.000	2.207	0.040

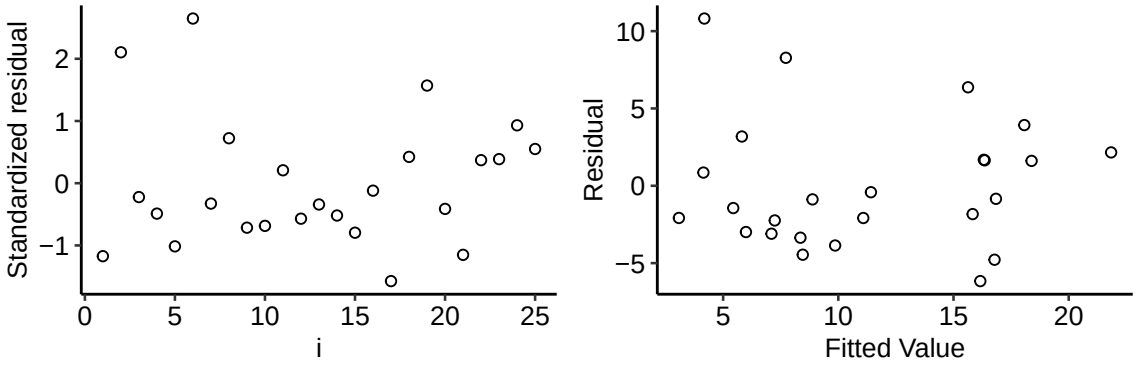

Figure 3: Residual plots for model 2

After the removal of the 1984 Summer Olympics, the model no longer has very large standardized residuals. There are two observations with residuals with magnitude greater than 2, but this is not anything to be significantly concerned about even given our new sample size of 25. Overall, the residuals do not appear to have any obvious pattern.

Model 2 may still have serial correlation, but it is less clear than it was for the previous model. It is still worthwhile to investigate, which we will do by plotting the residuals against each other at lags 1 and 2.

Based on the plots, there does not appear to be any obvious relationship between the residuals at lags 1 or 2. While there is a chance that there is some serial correlation at larger lags that we did not explore in detail, there is very likely not enough degrees of freedom to attempt to account for it in that case.

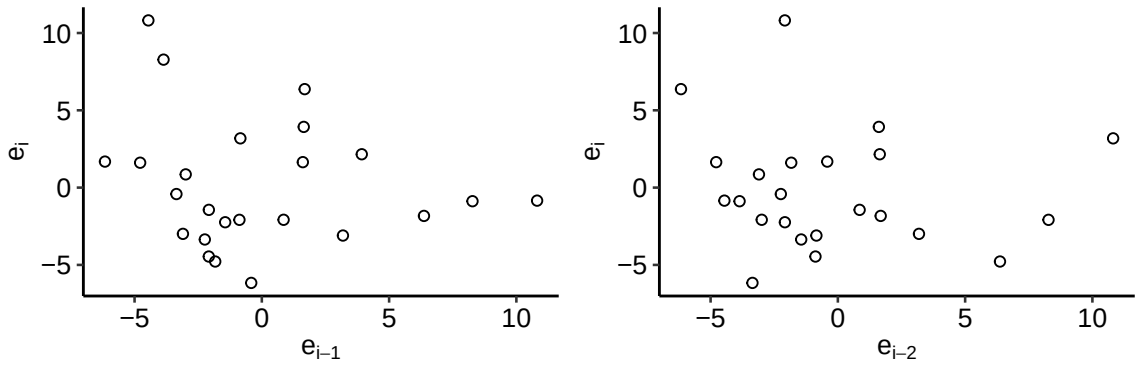

Figure 4: Lagged residual plots for model 2

Regression Subsets

Here we explore alternative models that do not include interaction and only include some subset of the original explanatory variables. Furthermore, we also ignore the collinearity between female representation and date, as our primary motivation is primarily prediction. The discussion in the previous section regarding the confound caused by countries boycotting for the 1984 Summer Olympics does not depend on the existence of the interaction term, nor the collinearity between date and female representation. Thus, removing it will likely result in more accurate predictions for the 2024 Summer Olympics for the models examined here as well.

An exhaustive search was performed over every combination of explanatory variables using `regsubsets` from the `leap` package, yielding the measures and variables displayed in Table 4.

Table 4: Summary of models chosen by regsubsets

R2	adjR2	Cp	p	Intercept	date	entries	other_entries	avg_age	prop_female
0.369	0.341	14.044	2	*					*
0.566	0.527	5.079	3	*				*	*
0.615	0.560	4.364	4	*	*			*	*
0.658	0.589	4.001	5	*	*	*		*	*
0.658	0.568	6.000	6	*	*	*	*	*	*

Based on the table, we suggest the model with 4 variables because it has a C_p value relatively close to p, the highest $adjR^2$, and increasing the number of variables to 5 does not increase R^2 (up to 3 decimal places). The selected model was fit and information about the coefficients are summarized in Table 5. As expected, the collinearity between date and female representation results in a coefficient sign that opposes the correlation between one of the two variables and the number of medals.

Table 5: Regression coefficients for suggested variable subset model, $R^2 = 0.658$, $adjR^2 = 0.590$, $\hat{\sigma} = 4.390$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	450.636	233.532	1.930	0.068
date	-0.254	0.122	-2.091	0.049
entries	0.028	0.018	1.577	0.130
avg_age	1.540	0.451	3.415	0.003
prop_female	53.980	20.456	2.639	0.016

Comparison

Table 6: Model Comparison

	R^2	$adjR^2$	$\hat{\sigma}$	df
Interpretable Model (Model 2)	0.635	0.539	4.653	19
Best Variable Subset	0.658	0.589	4.390	20

Even while using an extra degree of freedom, the more interpretable and intuitive model still explains less variance, has lower $adjR^2$, and higher residual standard deviation.

Discussion

From the previous Mallows' Cp statistic calculations, model 1 appears to be bested by the four-variable model the algorithm chose. The four-variable model has a better Cp statistic as well as a higher adjusted R-squared value than model 1's equivalent values. Instead of excluding female representation, the algorithm chose to exclude the date instead. Because female representation does have a higher correlation with medal count than date, this could explain why the four-variable model performed better. However, because the data is a time series, it makes more logical sense to exclude female representation in order to combat multicollinearity. The adjusted R-squared for the four-variable model is 0.452 and 0.424 for model 1, indicating that the data points do not fit the models all that well and that independent variables do not explain a majority of the variation of the data points. Model 2 has a higher adjusted R-squared value of 0.539, but even then this is still small. This could mean that there is some other variable or variables that have an effect on medal count that we have not taken into account.

Conclusion

Summary

We wanted to see if we could predict the number of medals Canada will win in the 2024 Paris Summer Olympics by using the start date of the games, the number of entries Canada had for events, the number of other entries in the events Canada was participating in, the average age of the Canadian athletes, and the proportion of the Canadian athletes who are female. We ended up discovering that female representation was highly correlated to the dates of the games, so we ended up discarding female representation as an explanatory variable to prevent multicollinearity. We produced two primary models: one without any interaction that omits female representation as an explanatory variable and one that also omits female representation but has an interaction between the date and entries. An exhaustive algorithm did find a four-variable model that outperformed our equivalent model; however, we still found ours more favorable as it made more logical sense to exclude female representation rather than date due to the data being a time series. Although these models were some of the better performing models, the low adjusted R-squared values indicated that no model was a great fit.

Answering Research Questions

To answer the question we posed in the Research Questions section, we technically can try to predict Canada's medal count in the 2024 Paris Summer Olympics by taking into account the start date of the Olympics, Canada's number of entries in events, entries by other countries in those events, the average age of Canadian athletes, and proportion of females among Canadian athletes. However, the correlation between

the proportion of female Canadian athletes and the start date is high, meaning including both in a model would mean multicollinearity will be a problem. Because the data used to make the prediction was collected over the course of years, it will be for the best that the proportion of female Canadian athletes is not an explanatory variable in the model. Even then, the model is not that strong of a tool to predict the amount of medals Canada will win in the Paris 2024 Olympics. This possibly means there are additional variables that we have not taken into account that will help with the prediction.

Limitations and Future Work

While this study provides a preliminary analysis of the factors influencing Canada's Olympic medal counts, there are several limitations and areas for future work:

1. Model Performance: We did not explore the out-of-sample predictive ability of our models in depth due to the time series nature of the data and small number of observations. A potential improvement upon this study would be to explore the effect of small sample size on variations of train/holdout set methods for out-of-sample prediction assessment which are suitable for time series data.

2. Limited Explanatory Variables: The analysis was restricted to a small number of variables, due in part to the small number of observations, potentially omitting significant factors that influence Olympic success. This is supported by the low R^2 values that both of our models had. Future studies could incorporate more diverse variables, such as political stability of a country or socio-economic factors.

3. Expanding scope of study: This study focused exclusively on Canada's performance in the Summer Olympics. Future research could expand the scope to include other countries, as well as explore the Winter Olympics or other international competitions, to determine if similar patterns and predictors apply. Observing trends like increasing number of entries into events and seeing how they affect medal counts could prove insightful. Furthermore, future studies could explore if specific sports such as archery or football have a stronger role in predicting medal counts and if socio-economic factors of countries such as GDP, levels of development, or average levels of education have a measurable effect on their Olympic Game success. Different countries as well as individual athletes incorporate varied training methods and have access to different kinds of equipment. A comparative analysis in this regard could uncover success factors.

4. Individual athlete performance: The data used in this analysis were aggregated at the national level, potentially obscuring individual athlete performance and sport-specific trends. Future research could focus on athlete-level performance metrics and sport-specific analyses. Specifically, a logistic regression model may be of interest.