Predicting Success for Olympic Track Athletes with a Multilevel Model

Nick Browen, Eric Ortiz

Introduction

Every four years, the Olympics Events garner worldwide attention much attention. We wanted to know what athlete characteristics contribute to their success and how the country they are from affects this. There is so much data readily available about the Olympics, even going back to 1896. We also were excited to be able to combine datasets and information from multiple resources from Olympic data to population and GDP of the countries of the world over time.

We chose to include athlete-level characteristics because the ideal physique for a distance runner is much different than a 100m sprinter and we think this will be useful in explaining differences in finishing time and are curious what other nuances will be revealed. At the country-level, we chose to include explanatory variables because athletes are sent as a team by their country and so the athlete pool, training resources and even quality of life might reveal some trends in predicting finishing times.

When looking through some literature on the subject, we came to realize that more is involved with success in the Olympics than just the variables at the athlete level such as height, weight, gender, and age. We were struck by Xun Bian's paper titled "Predicting Olympic Medal Counts: the Effects of Economic Development on Olympic Performance" in which Olympic medal counts for a country were predicted from variables at the country level such as population, GDP, who the hosting country is, and whether the country is Socialist or not. Further, when we looked at the paper by Filippo Radicchi titled "Universality, Limits and Predictability of Gold-Medal Performances at the Olympic Games", we could see that there is variablility at the athlete level since many athletes compete in more than one Olympics in their lifetime.

We chose to consider factors at both the athlete level and country level in a Hierarchical Model to predict the finishing time of track athletes.

Research Question

How can we best predict an Olympic track athlete's finishing time? What is the relationship between an athlete's finishing time and factors at the athlete level such as sex, age, weight, and height as well as factors at the country level such as the athlete's nationality, their country's GDP, and population?

Materials and Methods

Our data consist of merged datasets from Kaggle.com. First we found a comprehensive dataset of all Olympic medalists, but we subsetted this to include just track athletes that performed in running events (specifically the 10k race and all events shorter than 10k, not including events such as hurdles or steeple chase). This file contains info on athlete level characteristics such as height, weight, and age, but it did not include the finishing times. So we then merged this with another dataset from Kaggle that included finishing time. Then, from Gapminder.com we obtained country level information such as GDP and population and merged this into our dataset as well.

Having synthesized all these datasets together, we converted the event variable that would read like "100 M Men" into a quantitative variable that indicated the distance of the race. Finally, we rescaled the distance variable in order that the intercept would be about the 100m race. Similiarly, the year of the event was rescaled to be number of years since 1896. Height and weight were converted to BMI. Several variations of GDP and population of countries were tried out during the model building process. We first rescaled GDP as GDP in billions of dollars and, as indicated from our Exploratory Data Analysis, population was put on a log scale. GDP per capita (GDP/population) was also calculated. Later, GDP and population were categorized into "small", "medium", "large", and GDP per capita was categorized into "low" and "high". Throughout the model building process, log transformations of the distance of races and finishing times were calculated. All continous explanatory variables were centered, with the exclusion of distance and year.

To quantify the association between the finishing time of races and our predictor variables, we investigated these relationships at the athlete level and then at the country level. Correlation plots and correlation matrices were produced to identify predictor variables that were important to finishing time. To investigate interactions, plots of the finishing time versus a predictor variable were split into panels by another predictor variable were analyzed to determine if any relationships differs across another variable.

Initially, a two level random intercepts model was fit predicting finishing time (seconds) allowing the country the athlete is from to be the random. Quickly, we included the distance of the race as a predictor variable to account for the obvious variation in finishing times. Then, random slopes were included for distance. After verifying this was helpful, we added all predictor variables and interactions that our Exploratory Data Analysis indicated would be useful in predicting the finishing time. Then using t-tests, we systematically removed insignificant terms and refitted the model. Throughout this process is when several variables were converted into more useful variables such as height and weight into BMI, GDP into GDP into billions of dollars, and GDP per capita converted to a categorical variable. AIC and log-likelihood ratio tests were used to compare models. We briefly attempted log transformations on distance and finishing time, in an effort to remedy the effects of using distance (a somewhat categorical variable with large spacing between values) as a quantitative variable.

When merging these data sets together we also ran into missing values, namely for countries that were not included in Gapminder's country GDP and Population data or countries that changed their name at some point (for example Soviet Union to Russia). Where possible, we were able to search out the countries that changed name and correct for that error. However, we chose to omit countries that we had no country-level data on.

Results

In our final data set, we ended up with:

- 585 total observations (completed track events by a medaling athlete)
- 410 total athletes
- 45 total countries

Country	United States	United Kingdom	Jamaica	Kenya	Ethiopia	Finland
Number of Athletes	177	54	45	39	34	33

The following variables are used in our final model to predict `timeSecs`:

- `dist100`: Distance of the event, subtracting 100 to make our intercept (the 100m Dash) meaningful
- `c_BMI`: Centered BMI of the athlete in meters/cm^2
- `year1896`: Year of the event, centered at 1896 to make the intercept meaningful
- `sex`: Sex of the athlete
- `gdpPerCap_`: GDP per Capita of the country represented, where a GDP per Capita of greater than \$10,000 is considered "high" and a GDP per Capita of less than \$10,000 is considered "low"

Distance of Event	100	200	400	800	1500	5000	10000
Number of Observations	109	98	85	75	76	68	74

GDP/Capita of Country	high	low
Number of Observations	273	312

Sex of Athlete	M	W
Number of Observations	413	172

Final Model

Level 1 Equation:

$$FinishingTime_{ij} = \beta_{0j} + \beta_{1j}(dist100_{ij}) + \beta_1(cBMI_{ij})$$
$$+\beta_2(year1896_{ij}) + \beta_3(sex_{ij}) + \beta_4(sex_{ij} * dist100_{ij}) + \epsilon_{ij}$$

Level 2 Equation:

$$\beta_{0j} = \beta_{00} + \beta_{01}(gdpPerCap_{ij}) + u_{0j}$$
$$\beta_{1j} = \beta_{10} + \beta_{11}(gdpPerCap_{ij}) + u_{1j}$$

$$\epsilon \sim N(0, \sigma^2)$$
$$u_{0j} \sim N(0, \sigma^2_{u0})$$
$$u_{1j} \sim N(0, \sigma^2_{u1})$$
$$cov(u_{0j}, u_{1j}) = \tau_{01}$$

Composite Equation:

$$FinishingTime_{ij} = \beta_{00} + \beta_{10}(dist100_{ij}) + \beta_{01}(gdpPerCap_{ij}) + \beta_{11}(gdpPerCap_j * dist100_{ij}) + \beta_1(cBMI_{ij})$$
$$+\beta_2(year1896_{ij}) + \beta_3(sex_{ij}) + \beta_4(sex_{ij} * dist100_{ij}) + u_{0j} + u_{1j}(dist100_{ij}) + \epsilon_{ij}$$

Parameter Estimates

Observations	585
Dependent variable	timeSecs
Type	Mixed effects linear regression

AIC	5151.25
BIC	5203.71

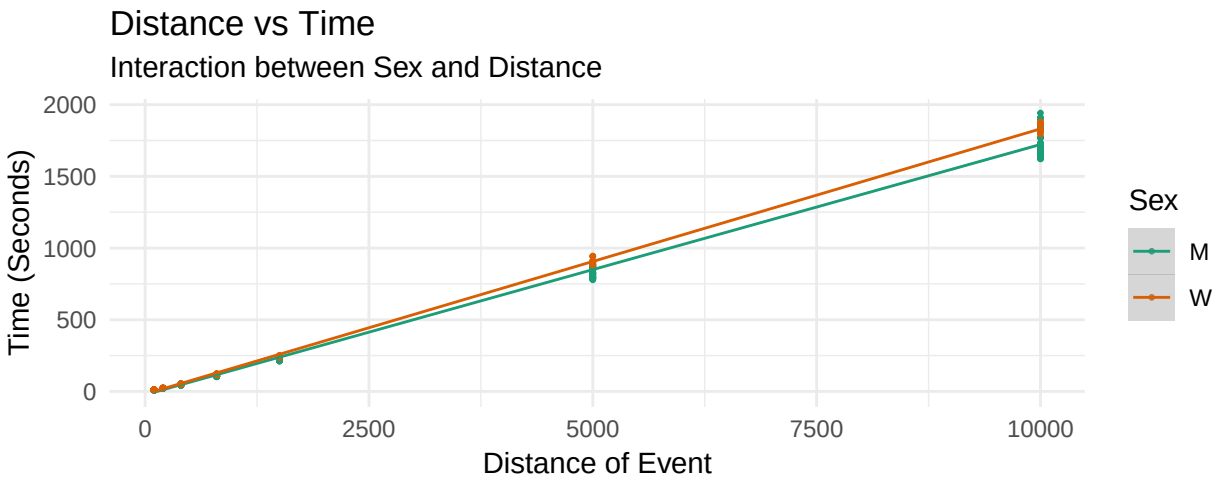
Fixed Effects						
	Est.	S.E.	t val.	d.f.	p	
(Intercept)	21.68	5.75	3.77	102.63	0.00	***
dist100	0.16	0.00	67.49	36.65	0.00	***
c_BMI	4.97	0.98	5.05	547.12	0.00	***
year1896	-0.32	0.04	-7.98	185.75	0.00	***
sexW	13.54	2.35	5.76	519.45	0.00	***
gdpPerCap_low	-17.11	2.80	-6.11	239.51	0.00	***
dist100:sexW	0.02	0.00	28.45	542.85	0.00	***
dist100:gdpPerCap_low	0.01	0.00	17.25	550.51	0.00	***

p values calculated using Kenward-Roger standard errors and d.f.

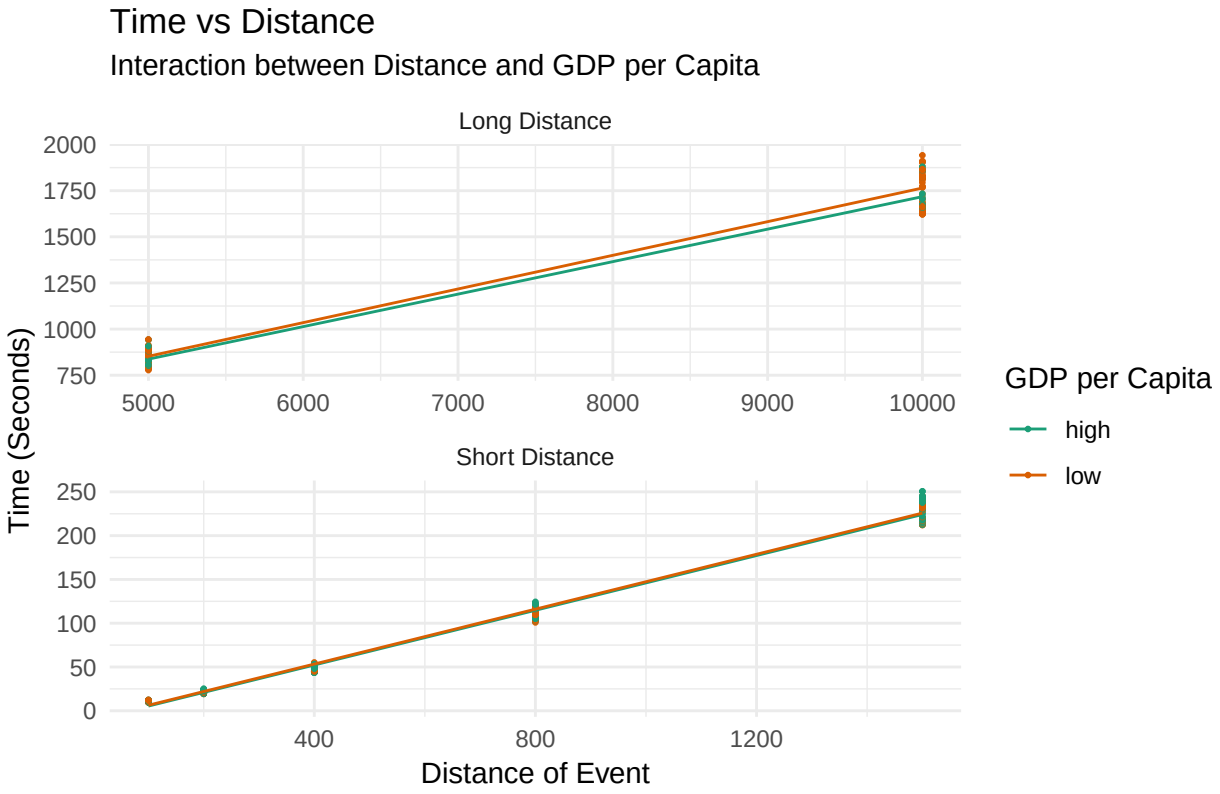
Random Effects		
Group	Parameter	Var.
country2	(Intercept)	460.78
country2	dist100	0.00
Residual		284.69

Grouping Variables		
Group	# groups	ICC
country2	45	0.62

- *Intercept*: The predicted finishing time for the 100-meter race in the year 1896 for a male athlete with an average BMI in a country with a high GDP per capita is 21.68 seconds.
- `dist100`: After adjusting for the year of the race and the BMI of the athlete, each 100-meter increase in the distance of a race is associated with a 16.26 second increase in a male athlete's finishing time for athletes competing for a country with a high GDP per capita.
- `c_BMI`: For an athlete competing for an average country, each $1\frac{kg}{m^2}$ increase in an athlete's BMI is associated with a 4.97 second slower finishing time after adjusting for the distance of the race, the year of the race, sex of the athlete, and GDP per capita of the country the athlete is competing for.
- `year1896`: After adjusting for the distance of the race, the BMI and sex of the athlete, and GDP per capita of the country the athlete is competing for, every 4 years (every Summer Olympic Games) is associated with a 1.26 second decrease in the finishing times of races.
- `sexW`: After adjusting for the year of the race, the BMI of the athlete, and GDP per capita of the country the athlete is competing for, a female athlete is predicted to have a 13.54 second slower finishing time than a male athlete for the 100-meter race.
- `gdpPerCap_low`: After adjusting for the year of the race, the sex, and BMI of the athlete, an athlete competing for a country with a low GDP per capita is predicted to have a 17.11 second faster finishing time than an athlete competing for a country with a high GDP per capita for the 100-meter race.
- `dist100:sexW`: After adjusting for the year of the race and the BMI of the athlete female athletes' associated rate of increase in their finishing times per 100m increase of the race is 1.76 second higher than male athletes.
  - After adjusting for the year of the race, the BMI of the athlete, and GDP per capita of the country the athlete is competing for, a female athlete is predicted to have a (13.54 + 0.02 $\times$ dist100) second slower finishing time than a male athlete. The plot below shows that the sex of the athlete has an effect on the association between distance of the race and finishing time.

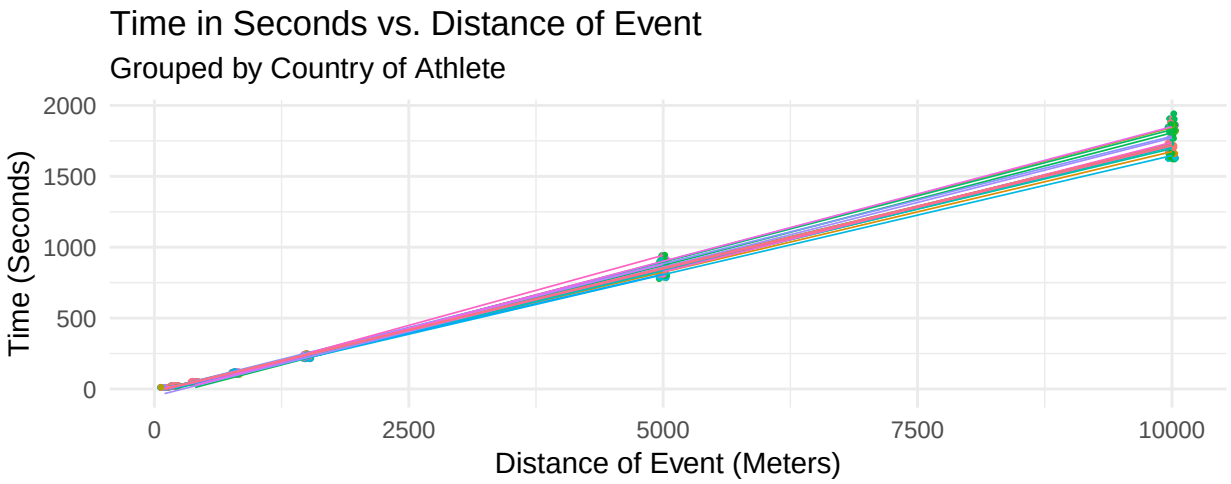


- `dist100:gdpPerCap_low`: After adjusting for the year of the race, the sex, and BMI of the athlete, an athlete competing for a country with a low GDP per capita is predicted to have a 1.43 second higher rate of increase in their finishing times per 100m than an athlete competing for a country with a high GDP per capita.
  - After adjusting for the year of the race, the sex, and BMI of the athlete, an athlete competing for a country with a low GDP per capita is predicted to have a $(17.11 + 0.01 \times \text{dist100})$ second faster/slower finishing time compared to an athlete competing for a country with a high GDP per capita.
  - Looking at the plot below, we can see that this interaction is very subtle. We dive more into the interpretation and implications of this in the discussion section.



- $\hat{\sigma}^2_{u0} = 460.78$: After adjusting for the BMI of an athlete, the standard deviation of countries' predicted finishing time for the 100-meter race in the year 1896 for a male athlete competing for a country with a high GDP per capita is 21.47 seconds.
- $\hat{\sigma}^2_{u1} = 0.0002$: After adjusting for the year of the race, the BMI and sex of the athlete, and GDP per capita of the country the athlete is competing for, the standard deviation of countries' rate of increase in finishing time per 100m is 1.28 seconds. As evidenced by the plot below, there are differing slopes for distance vs time depending on the country the athlete is representing.

Although the estimate for the variance for the random slopes of distance (σ^2_{u1}) is very small ($\hat{\sigma}^2_{u1} = 0.0002$), performing a log-likelihood ratio test to compare our final model to the same model without the random slopes produces $\chi^2 = 376.96$ with $df = 2$, resulting in an extremely small p-value. In fact, including random slopes for distance reduces the within country variance by 53.44% ($\frac{284.69-611.51}{611.51} * 100$). Because there are a large range of values for distance, this small variance of just 0.0002 does make a contribution to our model in predicting finishing times.



- $\hat{\sigma}^2 = 284.69$: After adjusting for the year and distance of the race, the BMI and sex of the athlete, and GDP per capita of the country the athlete is competing for, the standard deviation of athletes' finishing time within a country is 16.87 seconds.
- $cov(u_{0j}, u_{1j}) = \hat{\tau}_{01} = -0.95 * 21.47 * 0.01276 = -0.26$: After adjusting for the BMI of an athlete, countries that have higher predicted finishing times for the 100-meter race in the year 1896 for a male athlete competing for a country with a high GDP per capita tend to have a lower rate of increase (i.e. don't slow down as fast) in finishing time per 100m increase of the race. This relationship can be seen in the graph below.

Notably, all the parameter estimates discussed above are highly statistically significant, with the intercept having the smallest t-value equal to 3.77 (df = 102) corresponding to a p-value of 0.0001. The parameter estimate with the largest t-value belongs to `dist100` with a t-value of 67.49 (df = 36) interestingly followed by the interaction between `dist100` and `sex` with a smaller p-value and t-value equal to 28.45 (df = 542) and then the interaction between `dist100` and `gdpPerCap_` with a t-value equal t0 17.25 (df = 550), note that `lmer` package reports p-values using using Kenward-Roger standard errors and d.f. through the `pbkrtest` package.

Discussion

In our attempt to predict an Olympic track athlete's finishing time, we found that when grouping the athletes by their country there are a number of significant predictors at both the athlete level and country level. As one would expect, at the athlete level we found that distance, BMI, year, and sex are all highly significant. Surprisingly, age was not a significant predictor of finishing time in the presence of the other explanatory variables. This could due to the constraints of our data as we only have information on the top three finishers for any particular event and year. Then, at the country level we found that GDP per capita was a significant predictor of finishing time in the presence of the other explanatory variables although the effect is different from what we expected. This is not what we expected because we thought the resources of a wealthier country and larger population would result in a deeper athlete pool and more specialized training which could be associated with faster times for all races. However, due to the significant interaction between distance and GDP per capita, we found that for the short distance events of 100m to 800m, athletes representing a low GDP per capita country tend to have faster finishing times on average compared to athletes representing a high GDP per capita country. For events longer than 800m, athletes representing high GDP per Capita countries tend to have faster finishing times on average. It seems that a country's GDP per capita in a given year is associated with whether that country tends to perform better in short distance events or long distance events overall for that year. In a paper that addresses a similar topic, Xun Bian concludes that high GDP per capita countries tend to specialize less on particular events but do well over a broad range of events. Although this is in answer to a different research question (a country's number of medals earned), it is interesting to compare the effect of GDP per capita on finishing time across different track events with their findings. A key difference between our studies is that Xun Bian analyzed all Olympic events whereas we studied just track events.

Also mentioned in Xun Bian's study, the economic structure of the country (capitalist or communist) played a significant role in a country's medal count. As we did not distinguish countries by their economic structure, this could be a confounding variable in the effect of GDP per capita and the variation between countries. Furthermore, the number of athletes a country sends to the Olympics as a part of their Olympic team varies greatly between countries and could affect their performance overall. As touched on earlier, the data we used only includes the top three athletes from each event. The results of this study can only be generalized to Olympic medaling athletes in (non-hurdle, non-relay) running track events.

The multilevel model that we employed to address our research question allowed us to investigate both athlete level and country level characteristics and study the variability in finishing times at each of these levels. We found after adjusting for our explanatory variables, 61.81% of the variation in athlete's finishing times was due to variation between countries. Many of the suggested extensions of this study involve attempting to account for that variability. It would be beneficial to use data that include every athlete's finishing time, not only the medaling athletes. This would provide much more information about each country, as many countries in our current data set only have less than athletes representing them. Moreover, an addition of country level variables would be useful in accounting for more country-level variation such as metrics of quality of life. Another extension would be including an auto regressive error structure to account for the use of years as an explanatory variable.

Annotated Appendix

Handling Missing Data

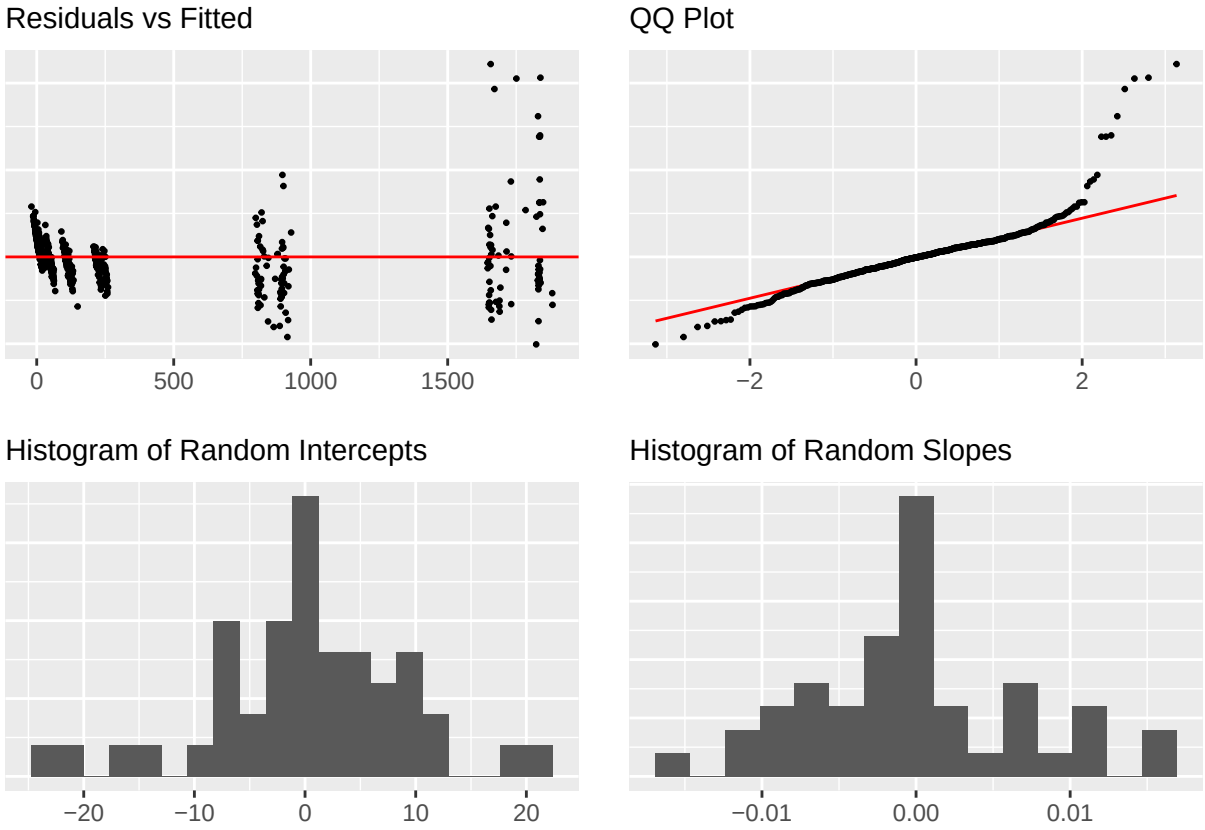
Since we did our own merging in order to compile data at both the athlete level and country level, we did run in to quite a few missing values. In particular, our missing values were almost all pertaining to either incomplete data from Gapminder on country's GDP and population, especially in earlier years. We generally chose to omit these cases since we would not have any country level information on these athletes.

Another case where we had lots of missing values was for countries that have changed names at some point (Russia vs. Soviet Union). In these cases we tried to compile all these observations under the same country name. For instance, we would change all Soviet Union athletes to represent Russia instead.

Handling Outliers

Because this data consisted of only the athletes that earned medals, all their track times were fairly similar and we did not come into any many outliers with respect to Time. The one place that did have outliers was GDP and Population, particularly when looking at China and India. We also noticed that these two variables were highly multicollinear. To take care of these two problems, we included a variable in our model, GDP per Capita. We chose to treat this as categorical and recoded the values to be either `low` or `high`, depending on whether the GDP per capita was above or below $10,000.

Residual Plots for Final Model



Looking at the Residuals vs Fitted plot, we can see that we can see that there is a fan shape, which indicates there is not equal variance of residuals. This is intuitive because longer events take more time and will have more spread in the results than shorter events. However, the residuals do appear to follow a roughly linear trend. Looking at the QQ Plot, we can see that the points follow the diagonal closely, so we can assume the data come from a Normal distribution. Looking at the histogram for Random Slopes and Random Intercepts, we can see that both appear to come from an approximate Normal distribution and it is fair to treat both intercepts and slopes as random.

Intermediate Models

Null Model

Looking at the residual plots for the null model below, we can see that the residuals do not follow a linear trend about the zero line and they are not normally distributed. However, we can see that random intercepts is a reasonable assumption to make. This model can be treated as a baseline to compare more complex models to. Note that for this model, the AIC is 8796 and BIC is 8809.

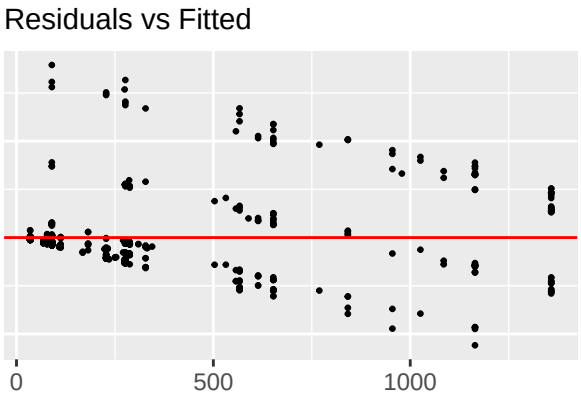
Observations	585
Dependent variable	timeSecs
Type	Mixed effects linear regression

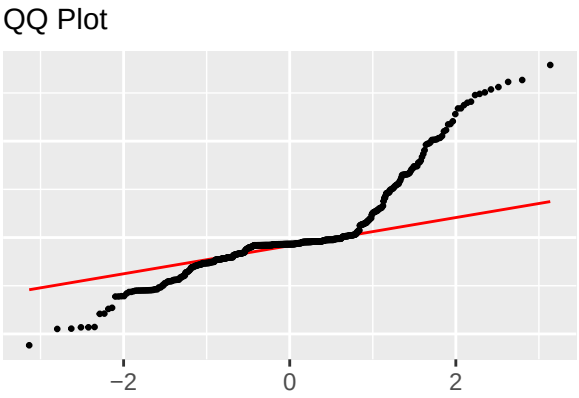
AIC	8795.95
BIC	8809.07

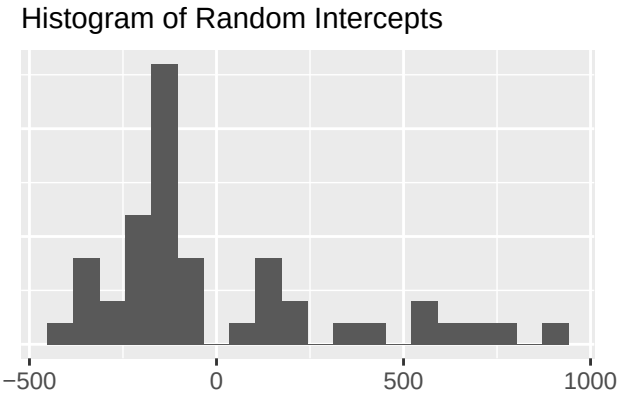
Fixed Effects						
	Est.	S.E.	t val.	d.f.	p	
(Intercept)	423.10	67.75	6.25	40.63	0.00	***
p values calculated using Kenward-Roger standard errors and d.f.						

Random Effects		
Group	Parameter	Var.
country2	(Intercept)	147177.40
Residual		175780.40

Grouping Variables		
Group	# groups	ICC
country2	45	0.46







Intermediate Model 1

This model includes the variables distance, BMI (centered), year (centered), sex, GDP (centered, in billions), and interactions between BMI and distance, sex and distance, and GDP and distance.

The AIC for this model is 5372 and BIC is 5429, which is a vast improvement over the null model. Also, our residual plots show that the assumptions are closer to being met with this model than the null model. The residual plots closely resemble those of the final model, although the histograms for random slopes and random intercepts are slightly more skewed.

Observations	585
Dependent variable	timeSecs
Type	Mixed effects linear regression

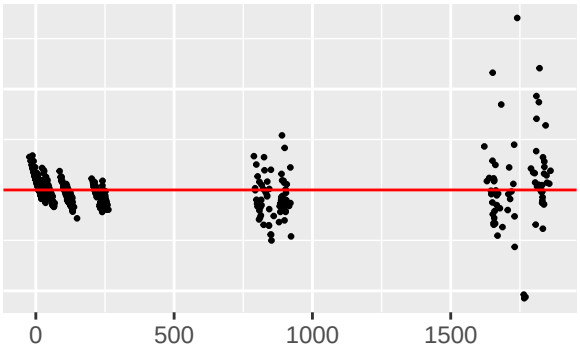
AIC	5372.39
BIC	5429.22

Fixed Effects						
	Est.	S.E.	t val.	d.f.	p	
(Intercept)	24.32	6.99	3.48	83.56	0.00	***
dist100	0.17	0.00	70.79	40.25	0.00	***
c_BMI	3.94	1.32	2.97	549.05	0.00	**
year1896	-0.45	0.05	-9.58	213.78	0.00	***
sexW	15.05	2.92	5.16	513.53	0.00	***
c_gdpbillion	8.57	1.61	5.33	409.74	0.00	***
dist100:c_BMI	0.00	0.00	3.83	539.51	0.00	***
dist100:sexW	0.02	0.00	21.75	540.78	0.00	***
dist100:c_gdpbillion	-0.00	0.00	-2.24	268.57	0.01	*
p values calculated using Kenward-Roger standard errors and d.f.						

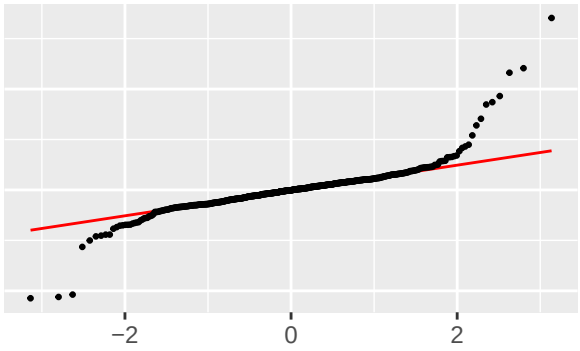
Random Effects		
Group	Parameter	Var.
country2	(Intercept)	1090.24
country2	dist100	0.00
Residual		407.00

Grouping Variables		
Group	# groups	ICC
country2	45	0.73

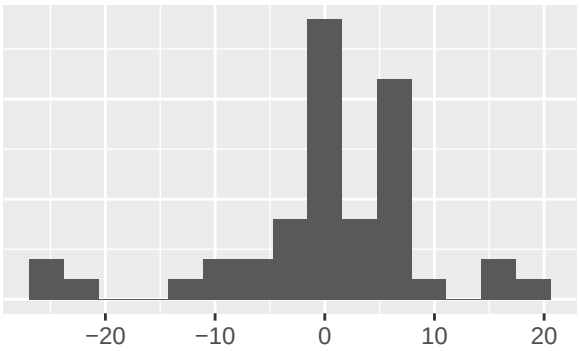
Residuals vs Fitted



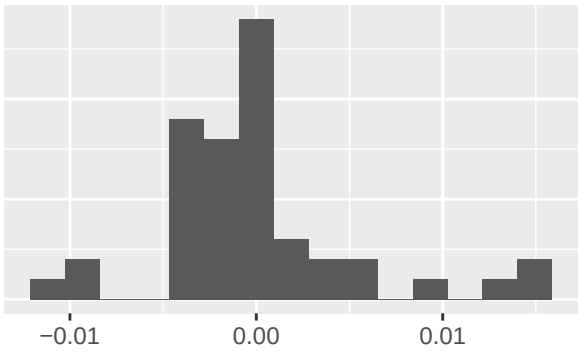
QQ−Plot



Histogram of Random Intercepts



Histogram of Random Slopes



Intermediate Model 2

In an attempt to improve Intermediate Model 1, we did a log transformation on distance because the distance of events does not change in a linear fashion, but roughly exponentially (100m, 200m, 400m, 800m, 1500m, 5000m, etc). However, upon looking at the residual plots, we could see that this transformation introduced many violations of assumptions, namely linearity which we were trying to address. This also changed the distribution of random slopes and random intercepts so that they no longer resemble a normal distribution, but became heavily skewed. It is easily observable from the scatterplot of Time vs Distance below that the association between the two is not linear.

We can't use AIC or BIC to assess the fit of this model relative to the previous models because of the log transformation, but judging from the residual plots this is not an improvement over Intermediate Model 1.

Observations	585
Dependent variable	timeSecs
Type	Mixed effects linear regression

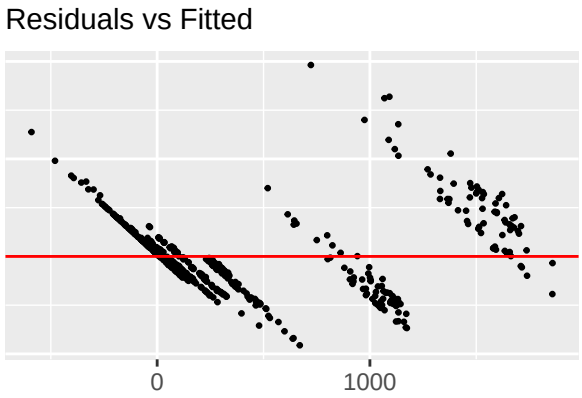
AIC	5348.21
BIC	5409.41

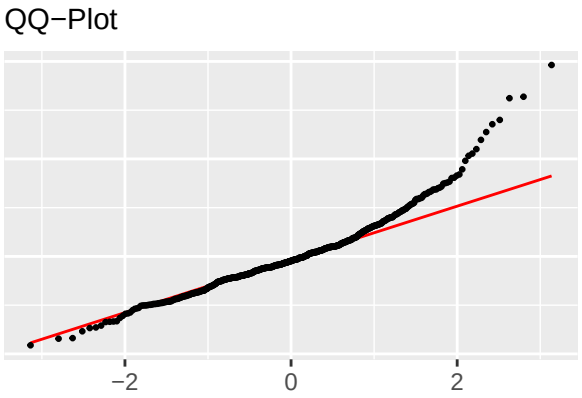
Fixed Effects						
	Est.	S.E.	t val.	d.f.	p	
(Intercept)	43.36	7.77	5.58	47.07	0.00	***
logdist100	-18.40	3.14	-5.87	57.32	0.00	***
c_BMI	-0.87	1.41	-0.62	540.02	0.27	
dist100	0.18	0.00	231.97	401.38	0.00	***
year1896	-0.36	0.05	-7.51	102.41	0.00	***
sexW	7.15	3.10	2.31	508.07	0.01	*
c_gdpbillion	10.01	2.01	4.98	505.82	0.00	***
c_BMI:dist100	0.00	0.00	6.08	537.28	0.00	***
dist100:sexW	0.02	0.00	22.26	563.59	0.00	***
logdist100:c_gdpbillion	-3.97	1.15	-3.44	277.69	0.00	***

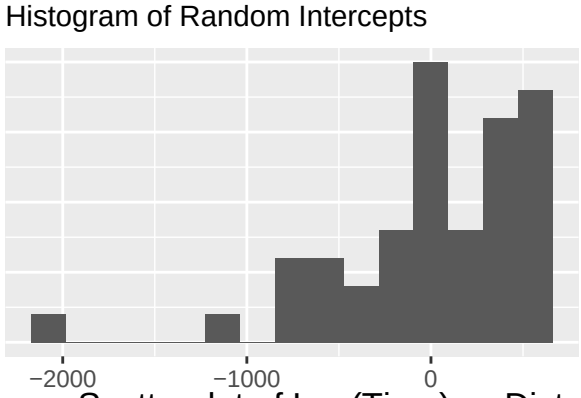
p values calculated using Kenward-Roger standard errors and d.f.

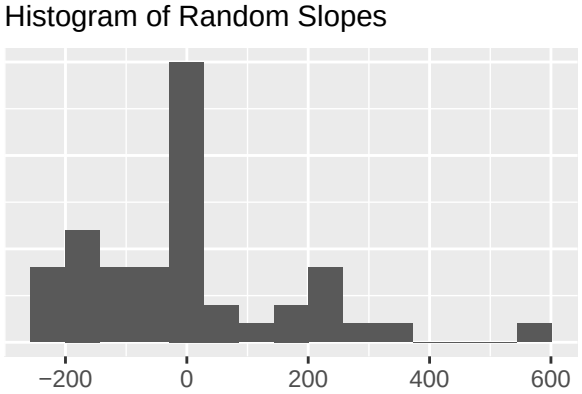
Random Effects		
Group	Parameter	Var.
country2	(Intercept)	965.92
country2	logdist100	188.04
Residual		442.29

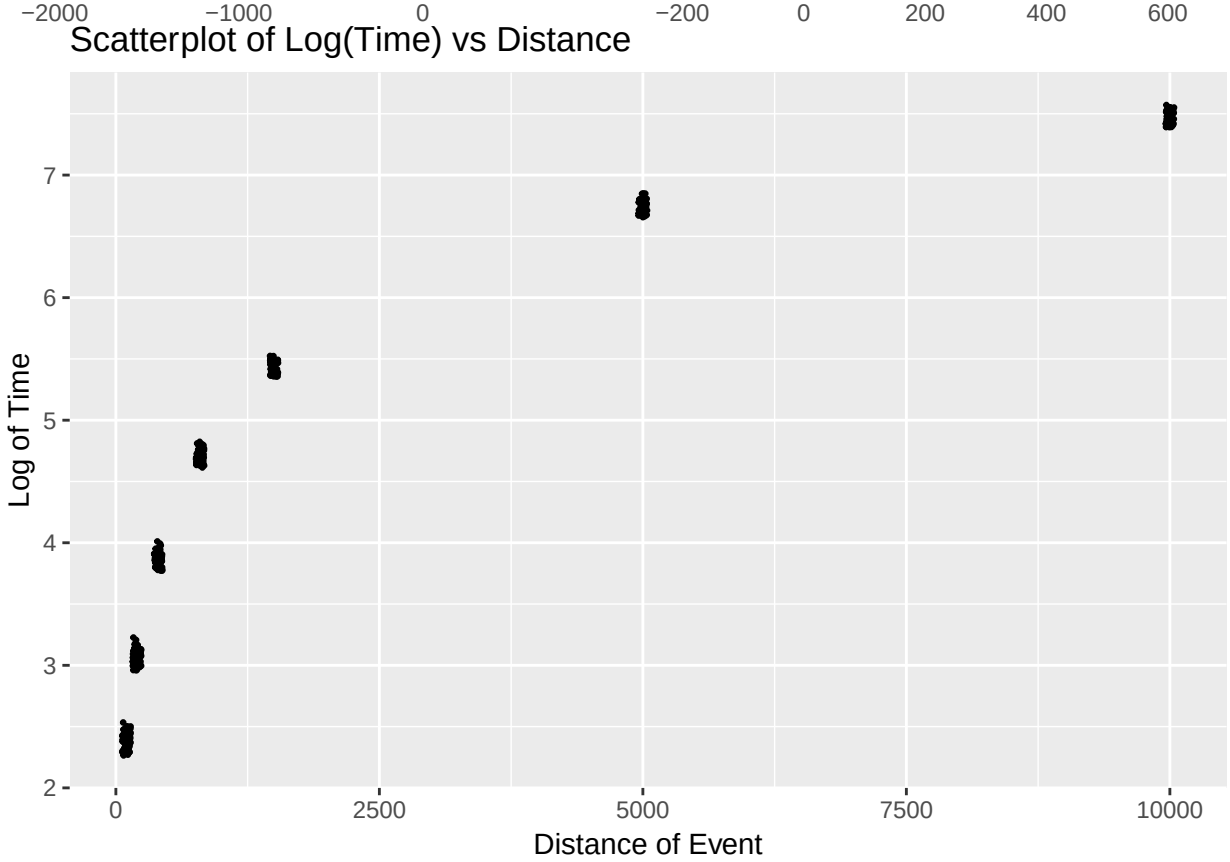
Grouping Variables		
Group	# groups	ICC
country2	45	0.69











Intermediate Model 3

With this model, we attempted to improve upon Intermediate Model 1 by including both a log transformation on Time and a log transformation on distance.

Looking at the scatterplot of Log(Time) vs Log(Distance), it seems as though the association between these two variables is linear with this new transformation. However, when looking at the residual plots, we can see that linearity is in fact violated now.

We can't use AIC or BIC to assess the fit of this model relative to the previous models because of the log transformation, but judging from the residual plots this is not an improvement over Intermediate Model 1.

Observations	585
Dependent variable	logtimeSecs
Type	Mixed effects linear regression

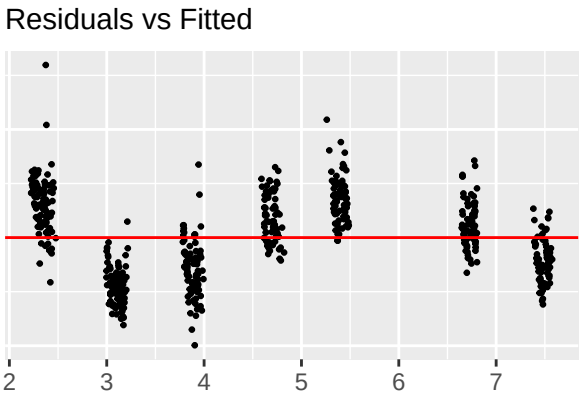
AIC	-1943.10
BIC	-1881.89

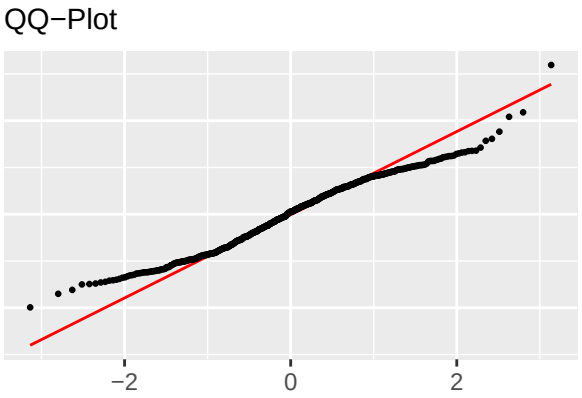
Fixed Effects						
	Est.	S.E.	t val.	d.f.	p	
(Intercept)	2.40	0.01	217.58	53.58	0.00	***
logdist100	1.14	0.00	280.53	67.02	0.00	***
c_BMI	-0.01	0.00	-2.73	564.42	0.00	**
dist100	-0.00	0.00	-6.39	348.34	0.00	***
year1896	-0.00	0.00	-15.50	368.28	0.00	***
sexW	0.11	0.01	15.30	524.55	0.00	***
c_gdpbillion	0.00	0.00	0.89	393.32	0.19	
c_BMI:dist100	0.00	0.00	1.76	519.25	0.04	*
logdist100:sexW	0.00	0.00	0.82	551.94	0.21	
logdist100:c_gdpbillion	-0.00	0.00	-0.64	321.17	0.26	

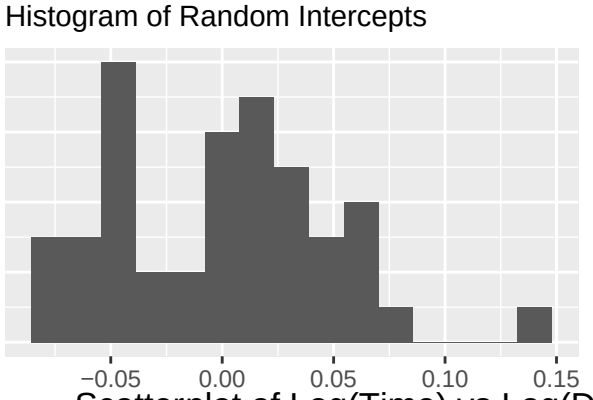
p values calculated using Kenward-Roger standard errors and d.f.

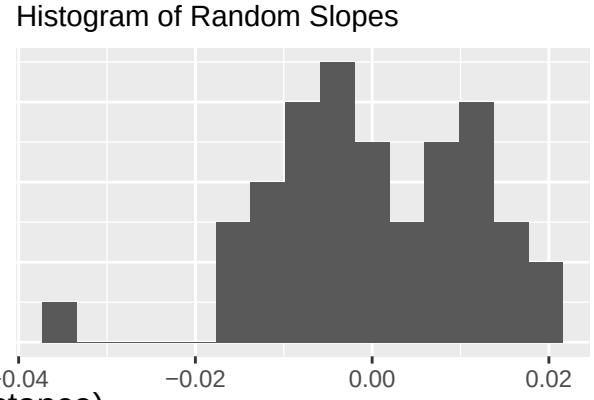
Random Effects		
Group	Parameter	Var.
country2	(Intercept)	0.00
country2	logdist100	0.00
Residual		0.00

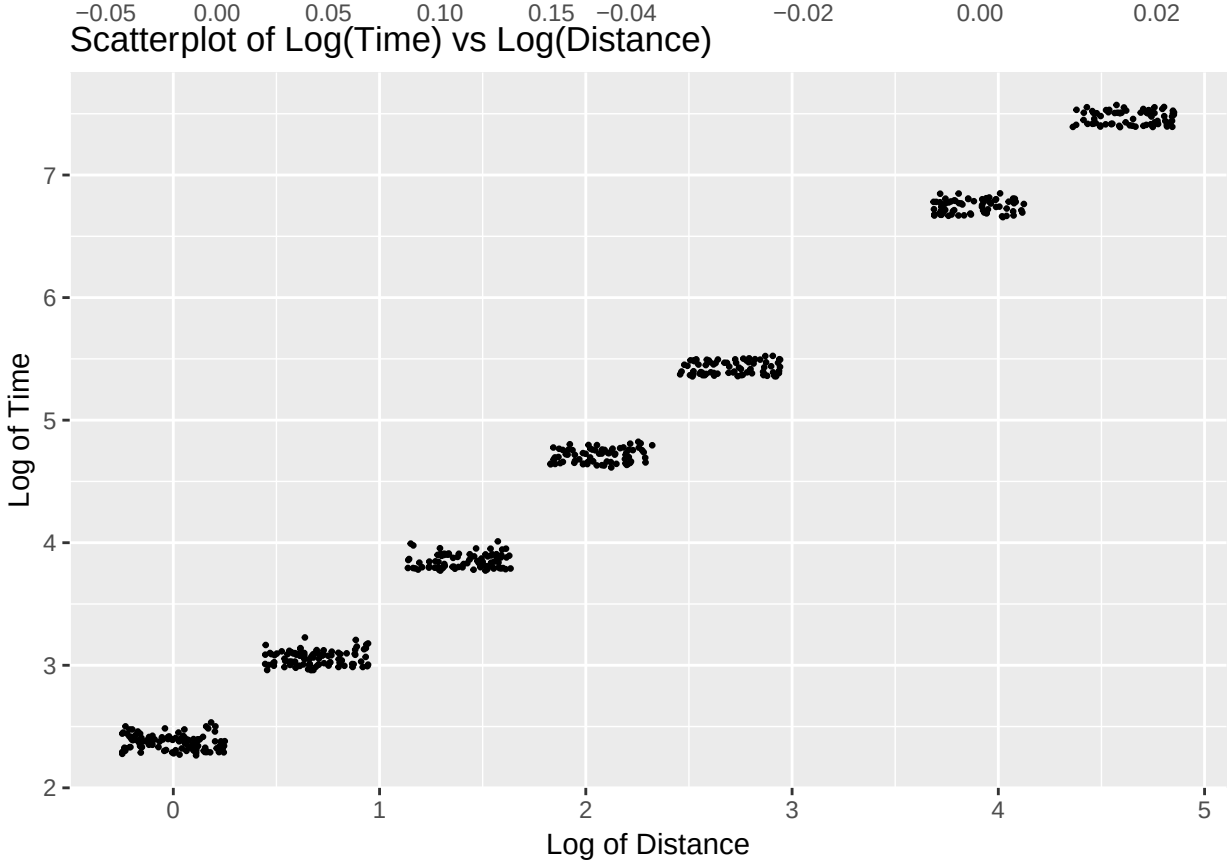
Grouping Variables		
Group	# groups	ICC
country2	45	0.24











Intermediate Model 4

In this model, we attempted to improve on Intermediate Model 1 by factoring GDP into countries that have either `small`, `medium`, or `large` GDP with cutoffs at the 25th and 75th percentile of GDP. There are no transformations of variables in this model, so it is reasonable to compare this to Intermediate Model 1 with AIC and BIC. For this model, the AIC is 5260 and the BIC is 5325. Compared to Intermediate Model 1, this is an improvement!

Looking now at the residual plots, it seems as if the assumptions of linearity and equal variance are in a similar as in Intermediate Model 1. The histograms for random slopes and random intercepts both appear to be good, so we can move forward with this model now.

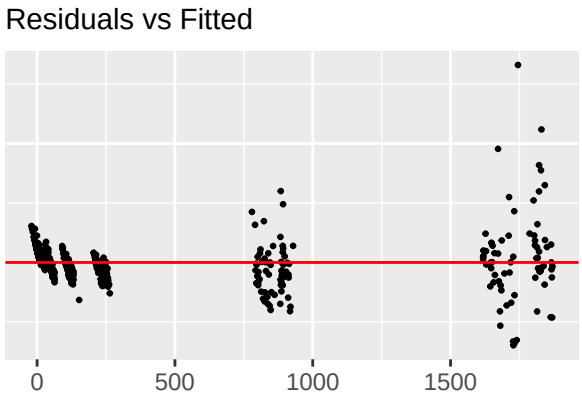
Observations	585
Dependent variable	timeSecs
Type	Mixed effects linear regression

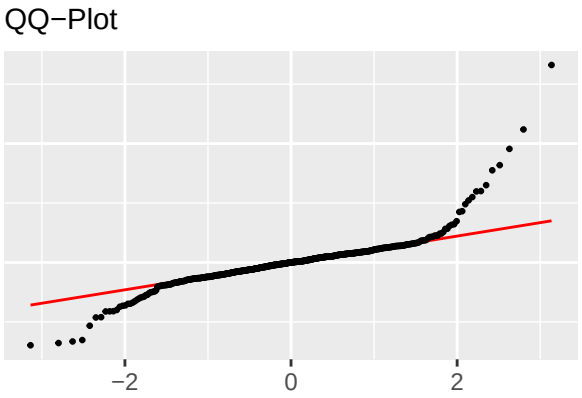
AIC	5259.89
BIC	5325.46

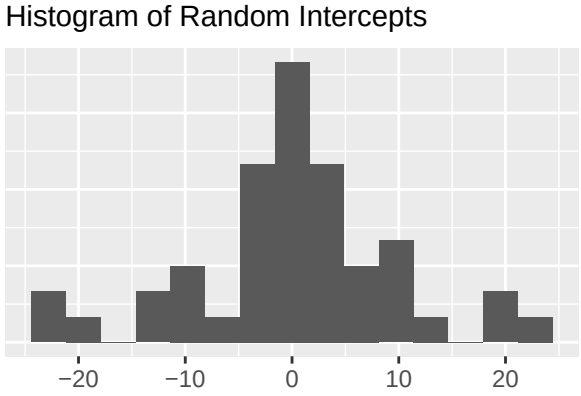
Fixed Effects						
	Est.	S.E.	t val.	d.f.	p	
(Intercept)	29.67	7.11	4.17	102.13	0.00	***
dist100	0.16	0.00	66.29	73.44	0.00	***
c_BMI	4.55	1.21	3.75	521.78	0.00	***
year1896	-0.32	0.04	-7.60	166.70	0.00	***
sexW	12.62	2.64	4.77	505.03	0.00	***
gdp_medium	-14.94	3.58	-4.18	293.56	0.00	***
gdp_small	-26.45	5.85	-4.52	150.21	0.00	***
dist100:c_BMI	0.00	0.00	3.10	553.16	0.00	**
dist100:sexW	0.02	0.00	24.69	545.02	0.00	***
dist100:gdp_medium	0.01	0.00	3.84	507.41	0.00	***
dist100:gdp_small	0.01	0.00	7.34	554.24	0.00	***
p values calculated using Kenward-Roger standard errors and d.f.						

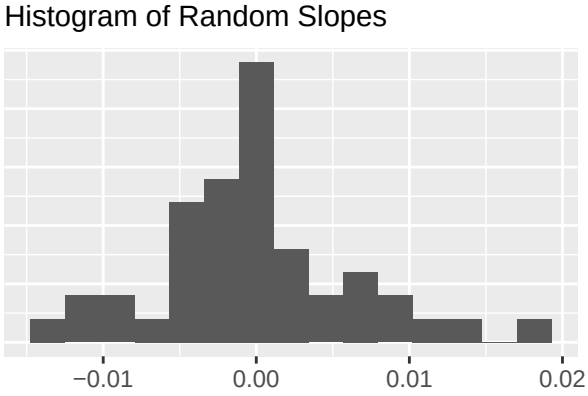
Random Effects		
Group	Parameter	Var.
country2	(Intercept)	329.44
country2	dist100	0.00
Residual		343.71

Grouping Variables		
Group	# groups	ICC
country2	45	0.49









Getting to the Final Model

The difference between Intermediate Model 4 and our final model was that we chose to use a measure of GDP per Capita rather than just straight GDP as one of the predictor variables. This fixed our earlier issue with the multicollinearity between GDP and Population, as well as contributed more information to our model. Changing this variable decreased the AIC to 5151 and the BIC to 5203. The residual plots did not change too much between models, so we decided to make this our final model as seen in the body of the report.

Citations

1. Radicchi F (2012) Universality, Limits and Predictability of Gold-Medal Performances at the Olympic Games.
https://doi.org/10.1371/journal.pone.0040335

2. Bian, X. 2005. Predicting Olympic Medal Counts: The Effects of Economic Development on Olympic Performance.
https://pdfs.semanticscholar.org/7293/1ab692bcab9e724b0e5ed4adb53b7ff8097f.pdf

3. Country Level Data such as GDP and Population.
https://www.gapminder.org/data/

4. Athlete Level Data for Finishing Time.
https://www.kaggle.com/jayrav13/olympic-track-field-results

5. Athlete Level Data for Height, Age, Sex, and Country.
https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results