Predicting Success for Olympic Track Athletes with a Multi-Level Model

Nick Browen, Eric Ortiz

==========
Project Description:
http://statweb.calpoly.edu/bchance/stat414/Stat414Project.pdf
==========

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

```
Three to five paragraphs (or fewer) that...
  1. Briefly describe your data, where it came from (source), definitions of important variables,
     and how it was collected
  2. Indicate any modifications made to the data, recoding, or decisions about missing data
  3. Briefly but thoroughly describe the statistical inference methods used to quantify the
     association between your outcome and predictor variables in a multilevel analysis.
     What summary statistics were calculated?
     What statistical tests were performed?
  4. Specify strategies employed when building your models
  5. Do not report results in the Materials and Methods section!
```

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

```
The meat of your report, which should include...
  1. A general description of your data (completed via your exploratory data analysis)
  2. A description of the results from your analyses, including interpretations of parameter
  3. Tables that summarize results and figures that illustrate results. These tables and figures
  4. You should interpret tests, confidence intervals, and coefficients in this section, but you
```

In our final data set, we ended up with:

- 585 total observations (completed track events by a medaling athlete)
- 410 total athletes
  - 177 from the USA
  - 54 from the UK
  - 45 from Jamaica
- 45 total countries

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
  - 109 observations from 100m Dash
  - 90 observations from 200m
  - 85 observations from 400m
  - 75 observations from 800m
  - 76 observations from 1500m
  - 68 observations from 5000m
  - 74 observations from 10000m
- `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
  - 413 male athletes
  - 172 female athletes
- `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"
  - 312 athletes from a low GDP per capita country
  - 273 athletes from a high GDP per capita country

Parameter Estimates

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 18.02s.

`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 16.26s 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 kg/mˆ2 increase in an athlete's BMI is associated with a 4.90s 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 1.13s 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 13.2s 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 14.86s slower 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 1.76s 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 (13.2 + 0.02*dist100)s slower finishing time than a male athlete.

`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 1.41s 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 (14.86 + 0.01*dist100)s faster/slower finishing time compared to an athlete competing for a country with a high GDP per capita.

$\hat{\sigma}^2_{u0} = 151.01$: 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 12.29 seconds.

$\hat{\sigma}^2_{u1} = 0.0009$: 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 0.92 seconds.

$\hat{\sigma}^2 = 293.94$: 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 17.14 seconds.

$cov(\hat{\sigma}^2_{u0}, \hat{\sigma}^2_{u1}) = \hat{\tau}^2_{01} = -0.88 * 12.289 * 17.145 = -185.41$: (interpreting the negative sign on the covariance term): 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 in finishing time per 100m increase of the race.

• Countries with high intercepts don't tend to slow down as much as the distance of the race increases compared to countries with lower intercepts.

Notably, all the parameter estimates discussed above are highly statistially significant, with the intercept having the smallest t-value equal to 3.92 (df = 533) corresponding to a p-value of 0.0001. The parameter estimate with the largest t-value belongs to `dist100` with a t-value of 85.45 (df = 533) interestingly followed by the interaction between `dist100` and `sex` with a t-value equal to 28.23 (df = 533) and then the interaction between `dist100` and `gdpPerCap_` with a t-value equal t0 16.96 (df = 533).

```
Left to do:
- Most important: plots and figures!
- tables
- more big picture (but without editorilzing)
- figure out how to format all that interpretation into report format (maybe plots will help integrate a
```

```
mod13 <- lme(data = track,
             fixed = timeSecs ~ dist100 + c_BMI + year1896 + sex + sex*dist100 + gdpPerCap_ + gdpPerCap_*
             random = ~ dist100|country2)
summary(mod13)
```

```
## Linear mixed-effects model fit by REML
##  Data: track
##        AIC      BIC    logLik
##   5142.058 5194.352 -2559.029
##
## Random effects:
##  Formula: ~dist100 | country2
##  Structure: General positive-definite, Log-Cholesky parametrization
##             StdDev       Corr
## (Intercept) 12.288575987 (Intr)
## dist100      0.009230628 -0.888
## Residual    17.144606507
##
## Fixed effects: timeSecs ~ dist100 + c_BMI + year1896 + sex + sex * dist100 +      gdpPerCap_ + gdpPer
##                             Value Std.Error  DF  t-value p-value
## (Intercept)            18.023258  4.595843 533  3.92164   1e-04
## dist100                 0.162607  0.001903 533 85.44675   0e+00
## c_BMI                   4.903095  0.976202 533  5.02262   0e+00
## year1896               -0.283504  0.036703 533 -7.72431   0e+00
## sexW                   13.207413  2.308153 533  5.72207   0e+00
## gdpPerCap_low         -14.861155  2.592148 533 -5.73314   0e+00
## dist100:sexW            0.017582  0.000623 533 28.23421   0e+00
## dist100:gdpPerCap_low   0.014111  0.000832 533 16.96308   0e+00
##  Correlation:
##                       (Intr) dst100 c_BMI  yr1896 sexW   gdpPC_ d100:W
```

```
## dist100               -0.504
## c_BMI                 -0.066  0.134
## year1896              -0.759  0.040 -0.067
## sexW                  -0.050  0.101  0.400 -0.242
## gdpPerCap_low         -0.676  0.152  0.060  0.585  0.052
## dist100:sexW           0.102 -0.097  0.015 -0.016 -0.410 -0.076
## dist100:gdpPerCap_low  0.160 -0.264 -0.125 -0.015 -0.107 -0.400  0.043
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -2.93492975 -0.49713620 -0.02064896  0.40879775  6.47067979
##
## Number of Observations: 585
## Number of Groups: 45
```

```
lmer_mod13 <- lmer(data = track, timeSecs ~ dist100 + c_BMI + year1896 + sex + sex*dist100 + gdpPerCap_
```

```
export_summs(summ(lmer_mod13, r.squared = F), error_format = "[{conf.low}, {conf.high}]", scale = T)
```

	Model 1
(Intercept)	21.676 ***
	[10.705, 32.647]
dist100	0.163 ***
	[0.158, 0.167]
c_BMI	4.966 ***
	[3.054, 6.878]
year1896	-0.315 ***
	[-0.389, -0.241]
sexW	13.538 ***
	[8.991, 18.085]
gdpPerCap_low	-17.112 ***
	[-22.424, -11.800]
dist100:sexW	0.018 ***
	[0.016, 0.019]
dist100:gdpPerCap_low	0.014 ***
	[0.013, 0.016]
N	585
N (country2)	45
AIC	5151.248
BIC	5203.708
R2 (fixed)	
R2 (total)	

*** p < 0.001; ** p < 0.01; * p < 0.05.

```
summ(lmer_mod13, r.squared = F)
```

Observations	585
Dependent variable	timeSecs
Type	Mixed effects linear regression

AIC	5151.248
BIC	5203.708

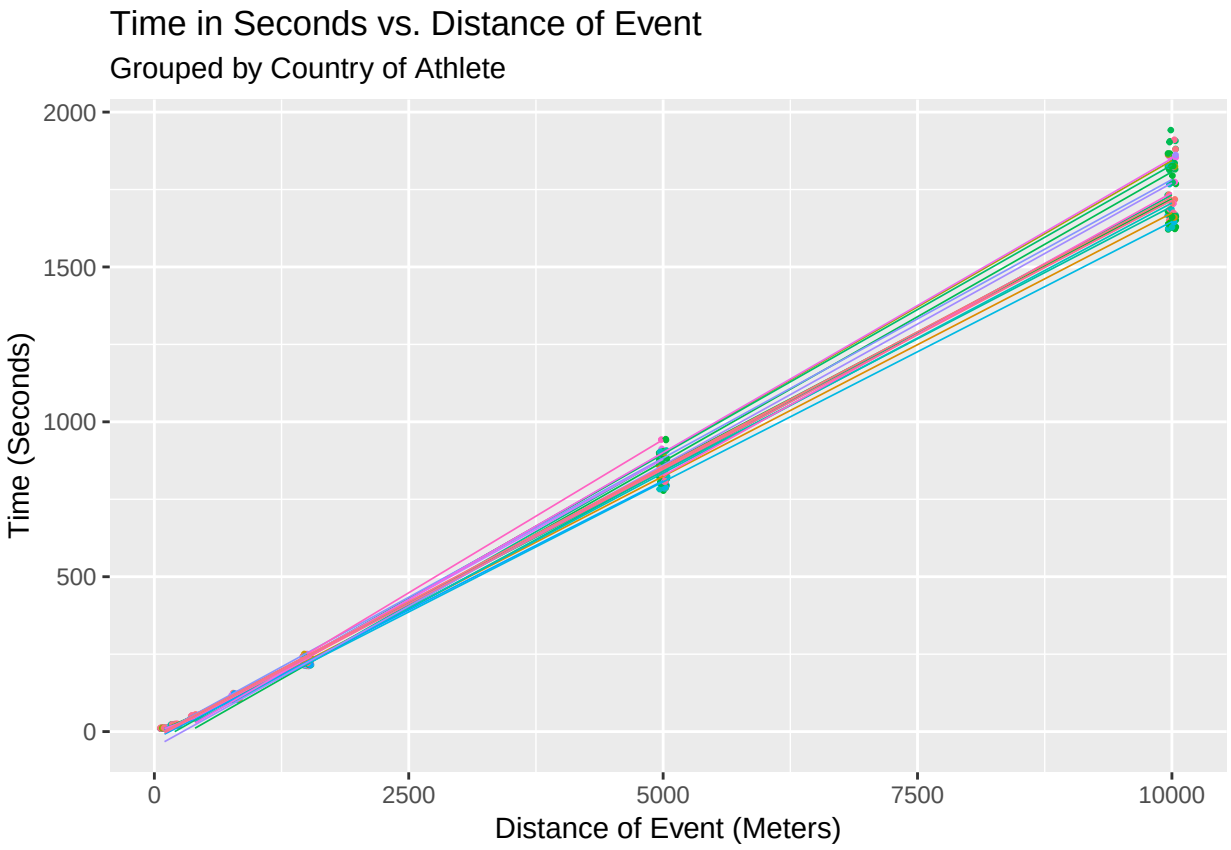
Fixed Effects						
	Est.	S.E.	t val.	d.f.	p	
(Intercept)	21.676	5.752	3.769	102.629	0.000	***
dist100	0.163	0.002	67.495	36.646	0.000	***
c_BMI	4.966	0.984	5.046	547.117	0.000	***
year1896	-0.315	0.039	-7.978	185.752	0.000	***
sexW	13.538	2.350	5.762	519.454	0.000	***
gdpPerCap_low	-17.112	2.800	-6.112	239.511	0.000	***
dist100:sexW	0.018	0.001	28.453	542.848	0.000	***
dist100:gdpPerCap_low	0.014	0.001	17.248	550.511	0.000	***

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

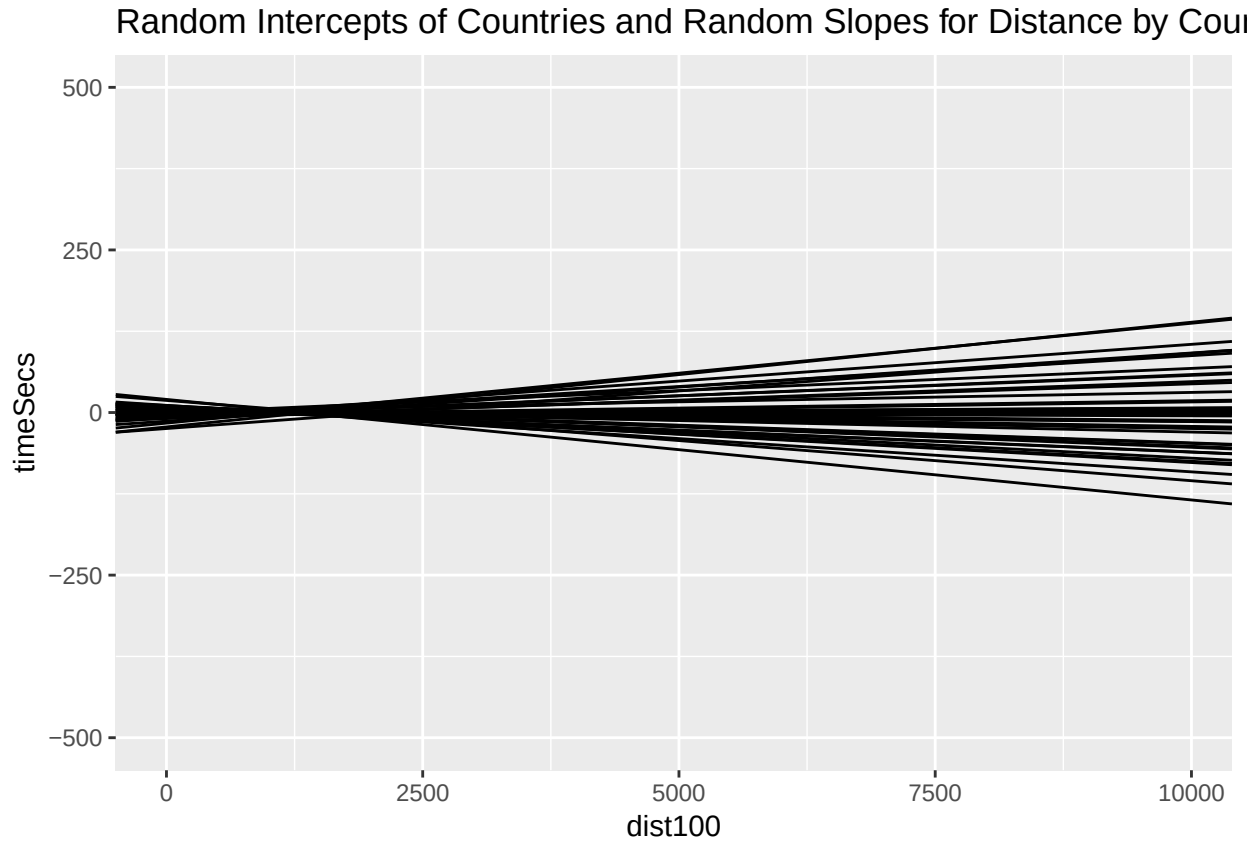
Random Effects		
Group	Parameter	Std. Dev.
country2	(Intercept)	21.466
country2	dist100	0.013
Residual		16.873

Grouping Variables		
Group	# groups	ICC
country2	45	0.618

```
ggplot(track, aes(y=timeSecs, x=dist, color = country2)) +
  geom_point(size = 0.5, position = "Jitter", show.legend = F) +
  geom_smooth(show.legend=F, se=F, size = 0.3, method = "lm") +
  labs(title = "Time in Seconds vs. Distance of Event",
       subtitle = "Grouped by Country of Athlete",
       y = "Time (Seconds)",
       x = "Distance of Event (Meters)")
```




```
ggplot(data = track, aes(x=dist100, y=timeSecs)) + geom_abline(intercept = mod13$coefficients$random$cou
```



Discussion

A few paragraphs that:

1. Begin with an accurate summary statement; describe how the results help answer your research
2. Discuss possible implications of the results in the context of the research question.
3. Make a statement regarding potential confounding variables in your study.
4. Make a statement about the generalizability of your results. Don't give generic statements of
5. Identify any limitations of your study. Discuss the potential impact of such limitations on the conclusions.
6. Identify strengths and weaknesses of your analysis.
7. Make suggestions for future research. Identify important next steps that a researcher could
8. Do not include test statistics or p-values in this section.

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.

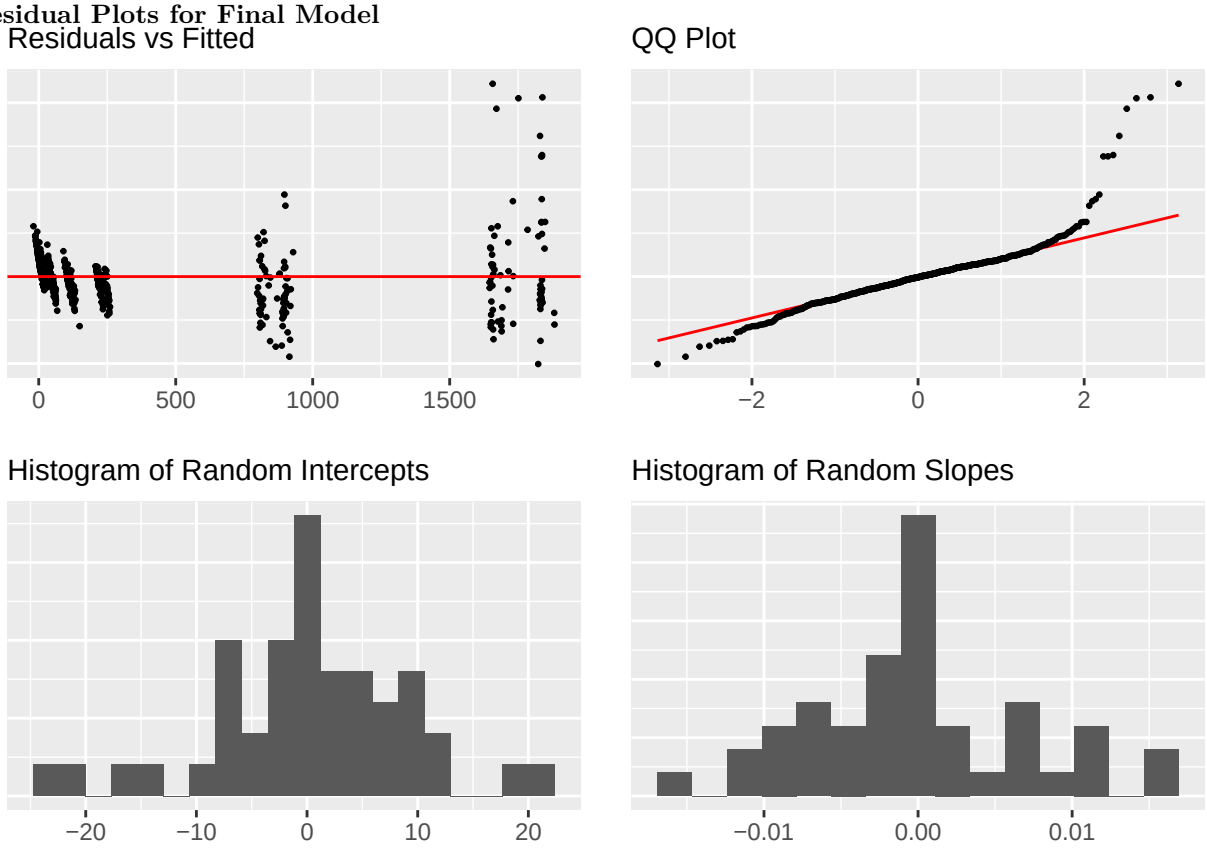
Another case where we had lots of missing values was for countries that have changed names at some point (Russia vs. Soviet Union). In these cases we tried to compile all these observations under the same country name. For instance, we would change all Soviet Union athletes to represent Russia instead.

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, BIC is 8809, and logLik is -4395.

```
mod0 <- lme(data = track, fixed = timeSecs ~ 1, random = ~1|country2)
summary(mod0)
```

```
## Linear mixed-effects model fit by REML
##  Data: track
##        AIC      BIC    logLik
```

```
##   8795.954 8809.064 -4394.977
##
## Random effects:
##  Formula: ~1 | country2
##         (Intercept) Residual
## StdDev:    383.6371 419.2617
##
## Fixed effects: timeSecs ~ 1
##                Value Std.Error  DF  t-value p-value
## (Intercept) 423.1011  67.53933 540 6.264514       0
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -2.66537290 -0.43540375 -0.16202238  0.03379804  4.26957739
##
## Number of Observations: 585
## Number of Groups: 45
```

```
lmer_mod0 <- lmer(data = track, timeSecs ~ 1 + (1|country2))

fit0 <- summ(lmer_mod0)
fit0
```

Observations	585
Dependent variable	timeSecs
Type	Mixed effects linear regression

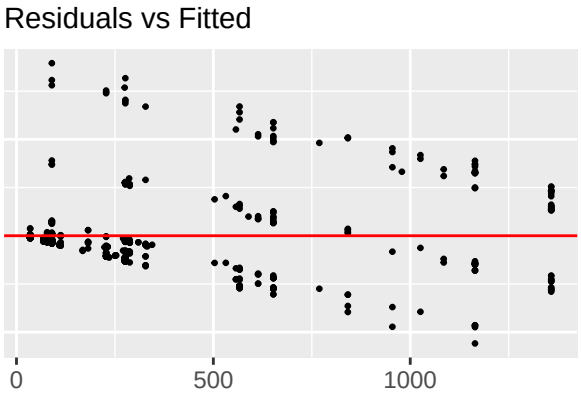
AIC	8795.954
BIC	8809.069
Pseudo-R^2 (fixed effects)	0.000
Pseudo-R^2 (total)	0.456

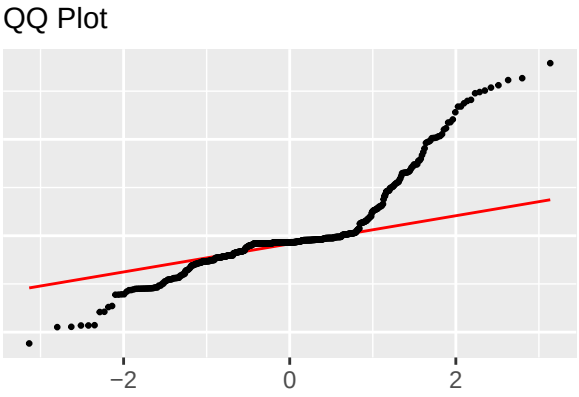
Fixed Effects						
	Est.	S.E.	t val.	d.f.	p	
(Intercept)	423.101	67.750	6.245	40.633	0.000	***
p values calculated using Kenward-Roger standard errors and d.f.						

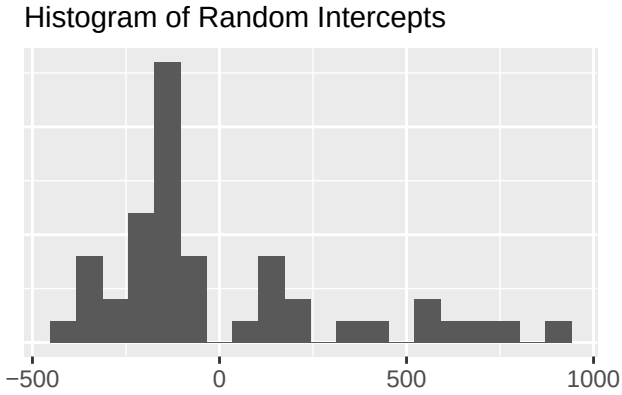
Random Effects		
Group	Parameter	Std. Dev.
country2	(Intercept)	383.637
Residual		419.262

```
#fit0 %>% export_summs(error_format = "[{conf.low}, {conf.high}]", scale = T)
```

Grouping Variables		
Group	# groups	ICC
country2	45	0.456







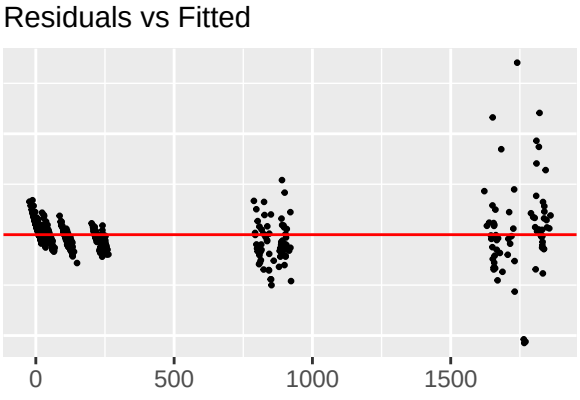
Intermediate Model 1

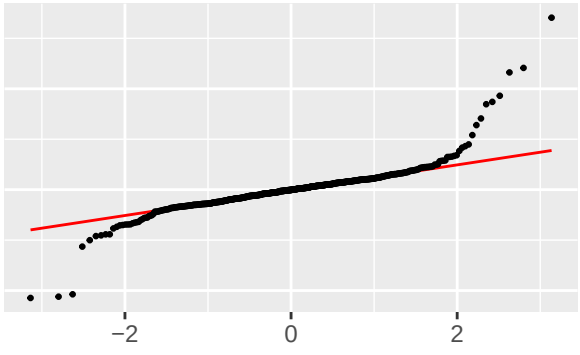
This model includes the variables distance, BMI (centered), year (centered), sex, GDP (centered, in billions), and interactions between BMI and distance, sex and distance, and GDP and distance.

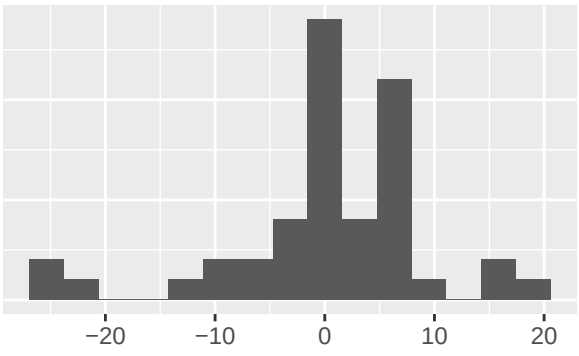
The AIC for this model is 5345, BIC is 5402, and logLik is -2660, which is a vast improvement over the null model. Also, our residual plots show that the assumptions are closer to being met with this model than the null model. The residual plots closely resemble those of the final model, although the histograms for random slopes and random intercepts are slightly more skewed.

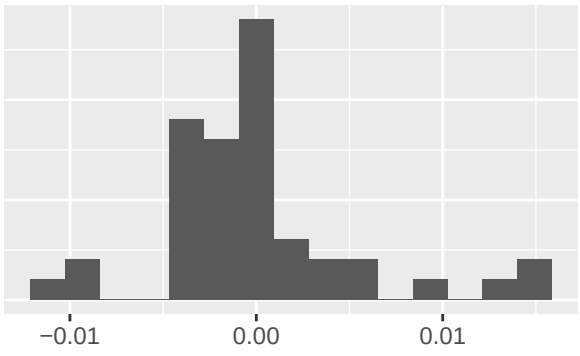
	Model 1
(Intercept)	24.325 ***
	(6.989)
dist100	0.172 ***
	(0.002)
c_BMI	3.937 **
	(1.324)
year1896	-0.452 ***
	(0.047)
sexW	15.048 ***
	(2.916)
c_gdpbillion	8.572 ***
	(1.609)
dist100:c_BMI	0.002 ***
	(0.000)
dist100:sexW	0.019 ***
	(0.001)
dist100:c_gdpbillion	-0.002 *
	(0.001)
N	585
N (country2)	45
AIC	5372.394
BIC	5429.225
R2 (fixed)	0.992
R2 (total)	0.999

*** p < 0.001; ** p < 0.01; * p < 0.05.









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.

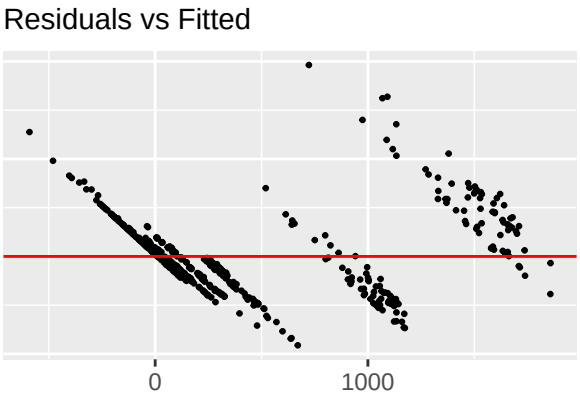
We can't use AIC, BIC, or logLik to assess the fit of this model relative to the previous models because of the log transformation, but judging from the residual plots this is not an improvement over Intermediate Model 1.

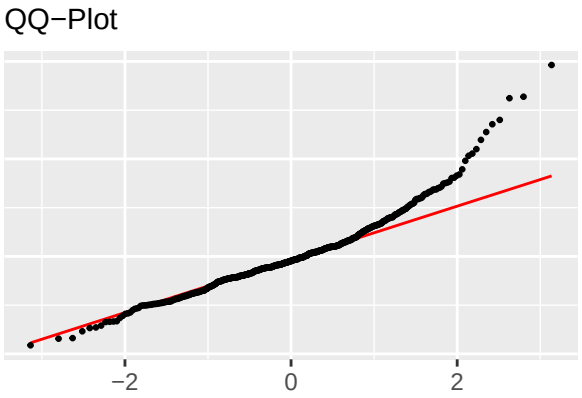
```
## Linear mixed-effects model fit by REML
##  Data: track
##       AIC      BIC    logLik
##   7917.89 7974.519 -3945.945
##
## Random effects:
##  Formula: ~logdist100 | country2
##  Structure: General positive-definite, Log-Cholesky parametrization
##             StdDev   Corr
## (Intercept) 624.1580 (Intr)
## logdist100  206.4461 -0.96
## Residual    190.5047
##
## Fixed effects: timeSecs ~ logdist100 + c_BMI + c_BMI * logdist100 + year1896 +      sex + sex * logdi
##                              Value Std.Error  DF   t-value p-value
## (Intercept)             -615.4558 119.94941 532 -5.130961  0.0000
## logdist100               404.4846  38.28273 532 10.565720  0.0000
## c_BMI                    106.6865  17.86603 532  5.971470  0.0000
## year1896                  -1.2779   0.48534 532 -2.632989  0.0087
## sexW                     191.8102  38.67325 532  4.959766  0.0000
## c_gdpbillion             -32.5653  18.85752 532 -1.726912  0.0848
## logdist100:c_BMI         -48.6902   7.90371 532 -6.160426  0.0000
## logdist100:sexW          -54.6727  16.01726 532 -3.413358  0.0007
## logdist100:c_gdpbillion   48.2808  10.36711 532  4.657115  0.0000
##  Correlation:
##                          (Intr) lgd100 c_BMI  yr1896 sexW   c_gdpb l100:_B
## logdist100               -0.891
## c_BMI                     0.001  0.054
## year1896                 -0.366  0.036 -0.186
## sexW                     -0.033  0.106  0.564 -0.232
## c_gdpbillion              0.266 -0.103 -0.103 -0.548 -0.098
## logdist100:c_BMI         -0.081  0.023 -0.756  0.237 -0.442 -0.024
## logdist100:sexW           0.036 -0.095 -0.416  0.061 -0.750  0.119  0.498
## logdist100:c_gdpbillion -0.036  0.111  0.108 -0.024  0.121 -0.589 -0.005
##                          l100:W
## logdist100
## c_BMI
## year1896
## sexW
## c_gdpbillion
## logdist100:c_BMI
## logdist100:sexW
```

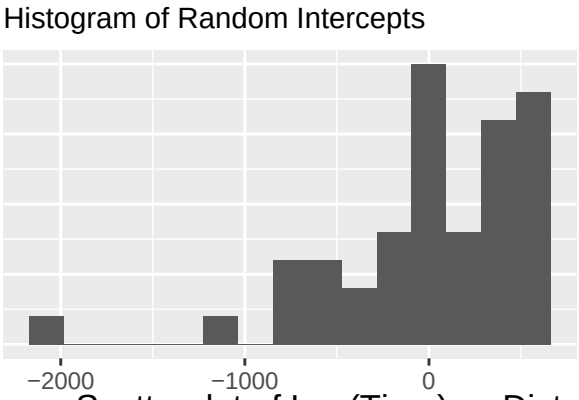
```
## logdist100:c_gdpbillion -0.107
##
## Standardized Within-Group Residuals:
##        Min         Q1        Med         Q3        Max
## -2.3925755 -0.5634903 -0.1256128  0.4028387  5.1532987
##
## Number of Observations: 585
## Number of Groups: 45
```

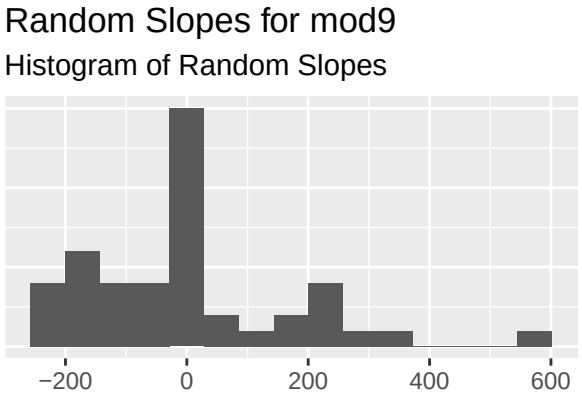
	Model 1
(Intercept)	43.357 ***
	(7.774)
logdist100	-18.398 ***
	(3.137)
c_BMI	-0.872
	(1.411)
dist100	0.181 ***
	(0.001)
year1896	-0.359 ***
	(0.048)
sexW	7.148 *
	(3.095)
c_gdpbillion	10.006 ***
	(2.009)
c_BMI:dist100	0.003 ***
	(0.000)
dist100:sexW	0.020 ***
	(0.001)
logdist100:c_gdpbillion	-3.968 ***
	(1.155)
N	585
N (country2)	45
AIC	5348.212
BIC	5409.414
R2 (fixed)	
R2 (total)	

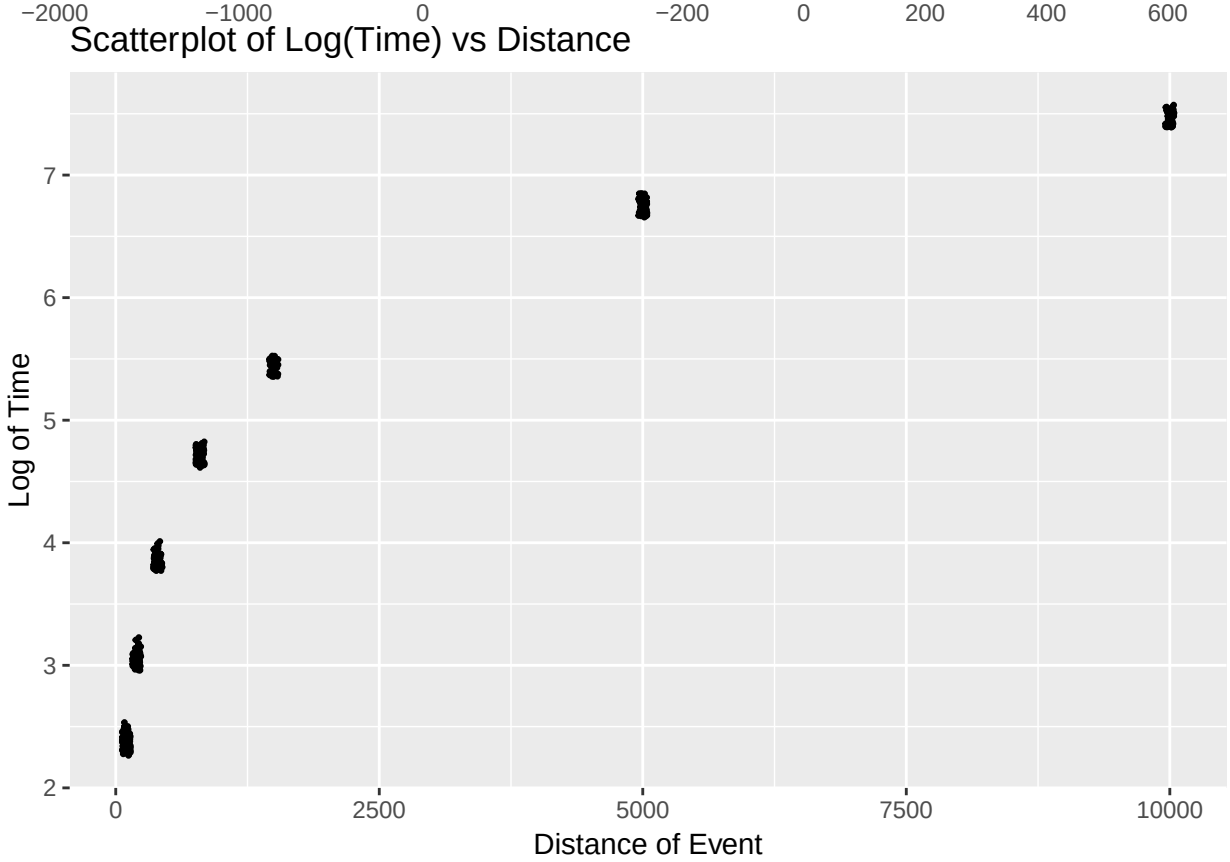
*** p < 0.001; ** p < 0.01; * p < 0.05.











Intermediate Model 3

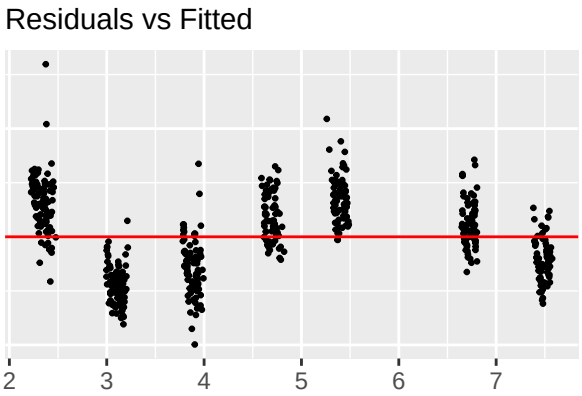
With this model, we attempted to improve upon Intermediate Model 1 by including both a log transformation on Time and a log transformation on distance.

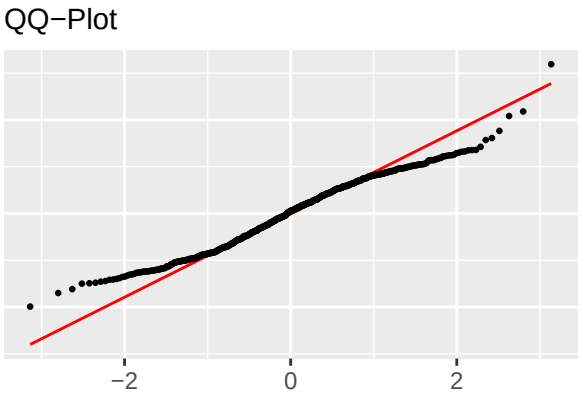
Looking at the scatterplot of Log(Time) vs Log(Distance), it seems as though the association between these two variables is linear with this new transformation. However, when looking at the residual plots, we can see that linearity is in fact violated now.

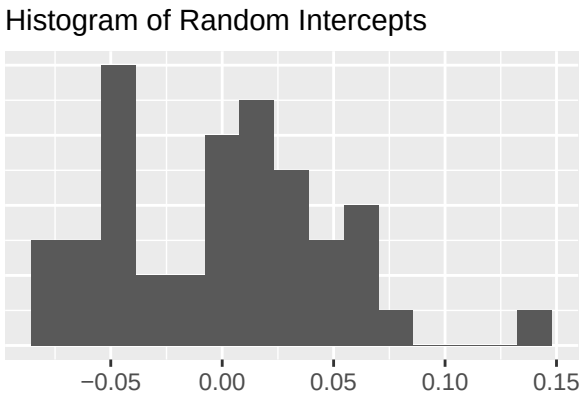
We can't use AIC, BIC, or logLik to assess the fit of this model relative to the previous models because of the log transformation, but judging from the residual plots this is not an improvement over Intermediate Model 1.

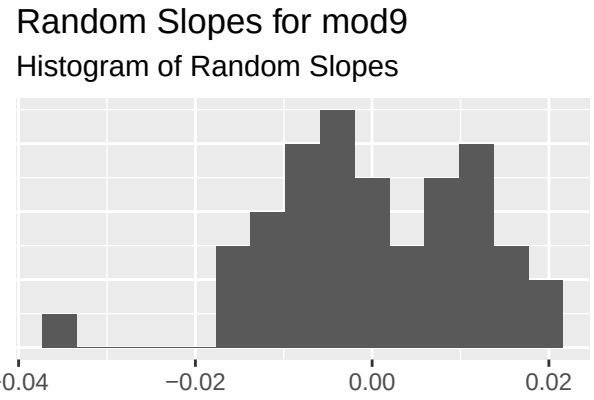
```
## Linear mixed-effects model fit by REML
##  Data: track
##         AIC       BIC   logLik
##   -1933.247 -1876.618 979.6236
##
## Random effects:
##  Formula: ~logdist100 | country2
##  Structure: General positive-definite, Log-Cholesky parametrization
##             StdDev     Corr
## (Intercept) 0.06341843 (Intr)
## logdist100  0.01547263 -0.97
## Residual    0.03921032
##
## Fixed effects: logtimeSecs ~ logdist100 + c_BMI + c_BMI * logdist100 + year1896 +      sex + sex * l
##                               Value   Std.Error  DF   t-value p-value
## (Intercept)              2.4616189 0.015352791 532 160.33690  0.0000
## logdist100               1.1120519 0.003570038 532 311.49579  0.0000
## c_BMI                   -0.0032275 0.003595938 532  -0.89755  0.3698
## year1896                -0.0015264 0.000092422 532 -16.51593  0.0000
## sexW                     0.1176975 0.007770546 532  15.14662  0.0000
## c_gdpbillion             0.0081033 0.003771490 532   2.14855  0.0321
## logdist100:c_BMI        -0.0018225 0.001543737 532  -1.18055  0.2383
## logdist100:sexW         -0.0003421 0.003137697 532  -0.10903  0.9132
## logdist100:c_gdpbillion -0.0037983 0.001876714 532  -2.02392  0.0435
##  Correlation:
##                          (Intr) lgd100 c_BMI  yr1896 sexW   c_gdpb l100:_B
## logdist100               -0.826
## c_BMI                    -0.013  0.113
## year1896                 -0.547  0.103 -0.191
## sexW                     -0.064  0.211  0.556 -0.229
## c_gdpbillion              0.403 -0.212 -0.106 -0.551 -0.110
## logdist100:c_BMI         -0.106  0.028 -0.759  0.268 -0.442 -0.028
## logdist100:sexW           0.060 -0.167 -0.415  0.054 -0.755  0.142  0.508
## logdist100:c_gdpbillion -0.098  0.204  0.119  0.064  0.139 -0.628 -0.024
##                          l100:W
## logdist100
## c_BMI
## year1896
## sexW
## c_gdpbillion
## logdist100:c_BMI
## logdist100:sexW
## logdist100:c_gdpbillion -0.157
##
## Standardized Within-Group Residuals:
```

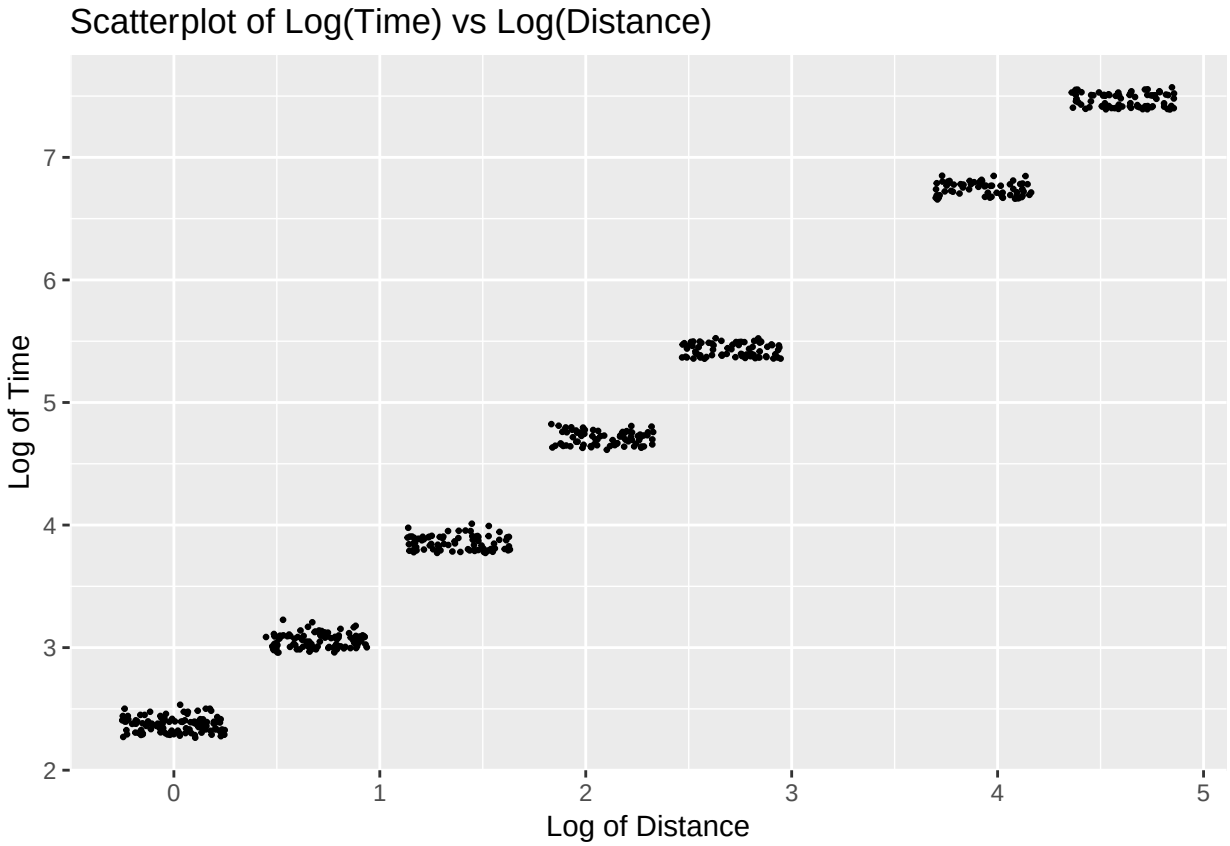
```
##         Min          Q1         Med          Q3         Max
## -2.53932749 -0.77973367  0.06844736  0.75072186  4.06989613
##
## Number of Observations: 585
## Number of Groups: 45
```











Intermediate Model 4

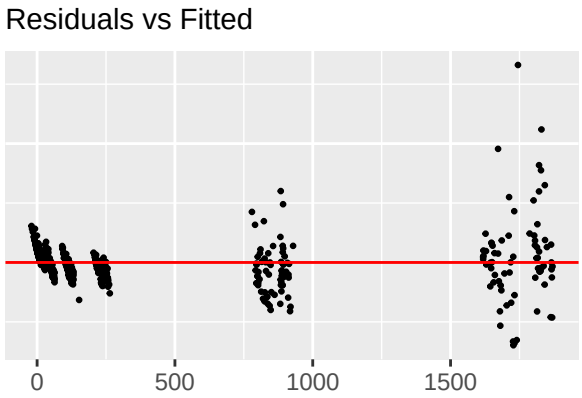
In this model, we attempted to improve on Intermediate Model 1 by factoring GDP into countries that have either `small`, `medium`, or `large` GDP with cutoffs at the 25th and 75th percentile of GDP. There are no transformations of variables in this model, so it is reasonable to compare this to Intermediate Model 1 with AIC, BIC, and logLik. For this model, the AIC is 5257, BIC is 5322, and logLik is -2614. Compared to Intermediate Model 1, this is an improvement!

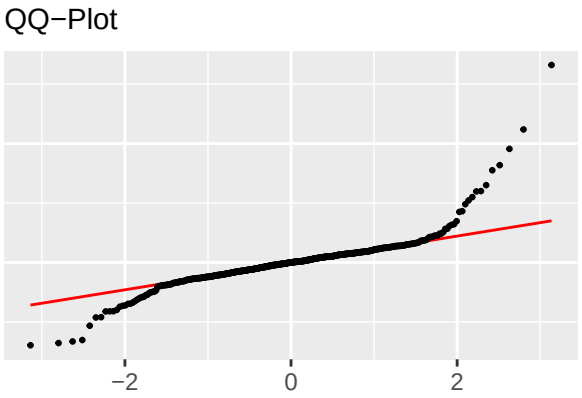
Looking now at the residual plots, it seems as if the assumptions of linearity and equal variance are in a similar as in Intermediate Model 1. The histograms for random slopes and random intercepts both appear to be good, so we can move forward with this model now.

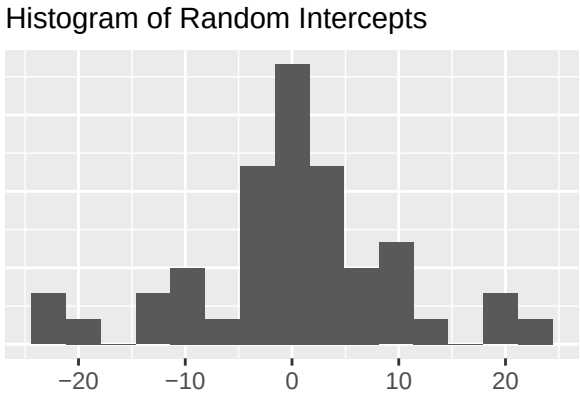
```
mod12 <- lme(data = track,
             fixed = timeSecs ~ dist100 + c_BMI + c_BMI*dist100 + year1896 + sex + sex*dist100 + gdp_ + g
             random = ~ dist100|country2)
summary(mod12)
```

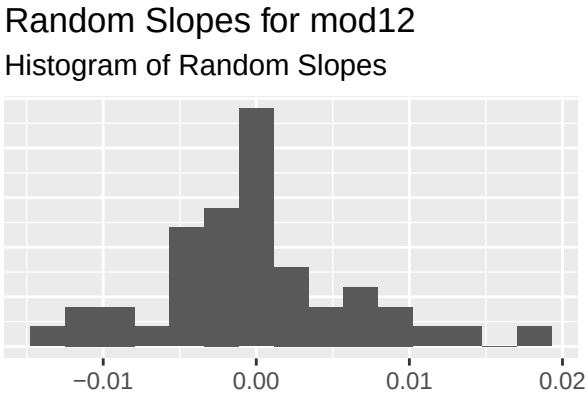
```
## Linear mixed-effects model fit by REML
##  Data: track
##        AIC      BIC    logLik
##   5257.156 5322.445 -2613.578
##
## Random effects:
##  Formula: ~dist100 | country2
##  Structure: General positive-definite, Log-Cholesky parametrization
##             StdDev       Corr
## (Intercept) 13.160951168 (Intr)
## dist100      0.008494513 -0.91
```

```
## Residual    18.716137131
##
## Fixed effects: timeSecs ~ dist100 + c_BMI + c_BMI * dist100 + year1896 + sex +      sex * dist100 + g
##                         Value Std.Error  DF  t-value p-value
## (Intercept)         25.720275  6.213561 530  4.13938  0.0000
## dist100              0.163850  0.002276 530 71.99498  0.0000
## c_BMI                4.520484  1.197282 530  3.77562  0.0002
## year1896            -0.301654  0.039981 530 -7.54496  0.0000
## sexW                12.380543  2.599109 530  4.76338  0.0000
## gdp_medium         -13.278367  3.397030 530 -3.90882  0.0001
## gdp_small          -22.268230  5.257091 530 -4.23585  0.0000
## dist100:c_BMI        0.001344  0.000437 530  3.07627  0.0022
## dist100:sexW         0.020006  0.000812 530 24.62632  0.0000
## dist100:gdp_medium   0.006014  0.001600 530  3.75800  0.0002
## dist100:gdp_small    0.012840  0.001783 530  7.20026  0.0000
##  Correlation:
##                     (Intr) dst100 c_BMI  yr1896 sexW   gdp_md gdp_sm
## dist100             -0.500
## c_BMI                0.034  0.014
## year1896            -0.756  0.095 -0.146
## sexW                -0.035  0.067  0.444 -0.270
## gdp_medium          -0.774  0.297  0.000  0.576  0.022
## gdp_small           -0.680  0.315 -0.119  0.424  0.059  0.643
## dist100:c_BMI       -0.135  0.113 -0.448  0.197 -0.262  0.089  0.095
## dist100:sexW         0.007 -0.067 -0.224  0.119 -0.481 -0.002 -0.027
## dist100:gdp_medium   0.290 -0.644  0.044 -0.115  0.000 -0.439 -0.311
## dist100:gdp_small    0.286 -0.647  0.071 -0.059 -0.033 -0.357 -0.480
##                     d100:_B d100:W dst100:gdp_m
## dist100
## c_BMI
## year1896
## sexW
## gdp_medium
## gdp_small
## dist100:c_BMI
## dist100:sexW         0.532
## dist100:gdp_medium -0.098   0.023
## dist100:gdp_small  -0.126   0.057  0.904
##
## Standardized Within-Group Residuals:
##          Min           Q1          Med           Q3          Max
## -3.722971411 -0.428965276  0.001255295  0.385588216  8.876134672
##
## Number of Observations: 585
## Number of Groups: 45
```









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 5142, BIC to 5194, and logLik increased to -2559. 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