Olympic Relay Predictions

Steven Ward

2025-12-16

Advised By: Brian Macdonald, Yale Statistics and Data Science

Abstract

Every four years the Summer Olympics come around and the world is reminded that swimming is an entertaining and highly competitive sport. Each race coming down to hundredths of seconds, exciting fly over starts in relays, and the highest amount of medals won is what encapsulates the viewers. In a sport that is so competitive and is won by fingertips, it is important to be precise. This report aims to predict swimmers' Olympic times based on how they have swam since the previous Olympics and use that to build the fastest possible relays. In order to make these predictions, we used data given to us from USA Swimming, which contains USA swimmers' times from various meets from 2020 to the 2024 Paris Olympics. This data contains the swimmer's name, gender, age, time (in seconds), event, meet, lane, and date. There are 960 observations split over 6 events, the 100 of every stroke along with the 50 and 200 freestyle (the races with relays). In order to predict Olympic times, we used a linear mixed-effects model. The predictors include whether or not the swimmer is tapered, age, gender, along with random effects for each swimmer and each meet. A model was made for each event as the strokes differ enough where one model would be insufficient. The predictions are evaluated by measuring error, bias, and the rank of the swimmers. The results show that the prediction models built in this project yield accurate results and can be used to help Team USA choose who to put on relays. This project builds a user friendly RShiny App. This app allows the user to choose from a few different analyses that can predict times and put together the best relay combinations. The app and code can be found at this link: https://github.com/sward48/Swim-Case-Study.

Introduction

The Olympics are a world renowned event in which countries from across the world send their best athletes to compete in a variety of different sports. Although coun-

tries do not directly gain any prizes from the Olympics, there are benefits to winning medals. These benefits are the intangibles such as national pride and bragging rights. Although this may seem like a small victory for a country, many of them provide financial incentive to their athletes to win medals. For a full breakdown of which countries give their athletes the highest incentives, please see the Wikipedia article cited. (https://en.wikipedia.org/wiki/Incentives_for_Olympic_medalists_by_country)

Since winning at the Olympics has a nation wide effect, countries should prioritize winning as many medals as possible. The sports that give the most medals during the summer Olympics are Swimming, Gymnastics, and Track & Field. This is due to the large variety of events. This means that countries should not only push training in these sports, but also strategy. Although largely individual, track and swimming have relay events. In track, the strategy for relays is simpler. Every athlete has one style of running. This allows track relays to put their best 4 members on the team. There are also dynamic starts in track that may vary from runner to runner, but the effect should be negligible compared to the rest of the race. Swimming has a more difficult challenge. Each swim race takes more of a toll on the entire body, so resting is more of a factor. This is simply because a 100 meter race in track is an all out effort for about 10-13 seconds, and a 100 meter race in swimming is an all out effort for about 50-60 seconds. In conjunction with that, in the medley relays, there are 4 completely different strokes being swam in one race. This makes choosing the relay team more difficult. There may be situations such as in 2021, where Caeleb Dressel was the fastest 100 freestyle and 100 butterfly (he won gold in both this year). Knowing that he is the best suited for both legs of the relay, where do you put him? This becomes more complex when you consider the mixed medley relay, a medley relay that consists of 2 women and 2 men. This is the crux of this project. How can we predict the best relay formation at the Olympics?

The motivation from this project came from a frustrating relay in the 2021 Olympics. The 4x100 Mixed Medley relay for team USA got 5th. Team USA got gold in the men's 100 freestyle, bronze in men's 100 backstroke, 4th in the women's 100 fly, and gold in the women's 100 breaststroke. How could a team composed of 4 top 4 finishers get 5th when put together? To someone who does not understand the inner workings of swim, this would seem near impossible for four other countries to best Team USA in this relay. This is where strategy beats putting in your highest placed swimmers. Looking closer into that final, Team USA was the only team to have a women swimming the breaststroke leg. So despite putting a gold medal winner in the relay, this was likely the wrong choice for Team USA. Although, Lydia Jacoby was the fastest women's breaststroker in the world, there were likely better alternatives for the line up. This is where strategy can be so important in swimming. Although a quick look at the swimmers and their placements would probably favor the chosen team, if an analysis was run, the team probably would have looked different.

This project aims to build an analysis tool that helps determine relay teams for Team USA. Not only will this tool help to build the best relay teams possible, but it will also allow for customization. This will include leaving out a swimmer from a relay in order to rest. This will be important in the case that a swimmer has many events in one day and likely will under

perform in one of their events if they swim all of them. This is done by leveraging data to build models that predict a swimmer's performance at the Olympics. After the models have been evaluated, we run an algorithm that compares every iteration of relay to find which one is predicted to be the best. This functionality will be encapsulated in an RShiny App which will allow for easy use.

The data was obtained from USA Swimming. Thanks to a connection from the advisor, we were able to get in contact and request a data set. This dataset consists of year, swimmer name, gender, age, event, date, meet, lane, time (minutes and seconds), and time (seconds). The data set has 960 rows, each containing a swim race. It spans from 2020 to the 2024 Paris Olympics. The races contained in this data set are from professional swim meets, some taper (rested) meets, some in season meets. The data is split across the 100s of each stroke along with the 50 and 200 freestyle.

Section 2 includes more information about the data and its attributes. In Section 3, we build and discuss several different analyses, including prediction models, relay combinations, and age curves. Section 4 contains the evaluation of all the analyses run in the project. We conclude with Section 5 where we discuss conclusions and next steps.

Data Exploration and Visualization

As mentioned in the Introduction, the data set consists of many interesting variables. The variables in the rawdata are year, swimmer name, gender, age, event, date, meet, lane, time (minutes and seconds), and time (seconds). Through some basic cleaning and manipulation, we expanded the number of columns. The feature engineered columns are relay_indicator, taper, is_olympics, pb_time_sec, and avg_prev_2_time_sec. The first 3 are binary variables that indicate whether a race was a relay, was a taper meet, or was at the Olympics. The last two are based on each swimmer. One column is their previous best time and one is their previous two times averaged. These features of the original data set will be helpful for modeling.

To start, we will look at some overarching summary statistics.

```
       V1            V2                V3                V4
 Min.   :2021   Length:960         Length:960         Min.   :14.0
 1st Qu.:2022   Class :character   Class :character   1st Qu.:20.0
 Median :2023   Mode  :character   Mode  :character   Median :22.0
 Mean   :2023                                        Mean   :22.5
 3rd Qu.:2024                                        3rd Qu.:26.0
 Max.   :2024                                        Max.   :31.0
      V5                V6                V7                V8
 Length:960         Length:960         Length:960         Length:960
 Class :character   Class :character   Class :character   Class :character
```

```
Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

```
     V9                 V10
Length:960         Min.   : 21.07
Class :character   1st Qu.: 50.18
Mode  :character   Median : 55.40
                   Mean   : 64.46
                   3rd Qu.: 67.40
                   Max.   :122.56
```

The above summary statistics are from the raw data. As you can see, the variables do not have names and there is little information to be gleamed from the table. Because of this, we will be investigating the cleaned version of the data. This version has names, more variables, and will provide more insights to this project.

```
  year         name            gender       age           stroke
2021:228   Length:960         F:467   Min.   :14.0   Length:960
2022:136   Class :character   M:493   1st Qu.:20.0   Class :character
2023:238   Mode  :character           Median :22.0   Mode  :character
2024:358                              Mean   :22.5
                                      3rd Qu.:26.0
                                      Max.   :31.0
```

```
  distance             date                meet                lane
Length:960         Min.   :2020-09-06   Length:960         NULL   :364
Class :character   1st Qu.:2022-03-02   Class :character   4      :308
Mode  :character   Median :2023-05-17   Mode  :character   5      :132
                   Mean   :2023-01-24                      3      : 52
                   3rd Qu.:2024-03-06                      6      : 32
                   Max.   :2024-07-27                      2      : 24
                                                           (Other): 48
time_rounded          time_sec      relay_indicator      taper
Length:960         Min.   : 21.07   Min.   :0.0000   Min.   :0.0000
Class :character   1st Qu.: 50.18   1st Qu.:0.0000   1st Qu.:0.0000
Mode  :character   Median : 55.40   Median :0.0000   Median :0.0000
                   Mean   : 64.46   Mean   :0.1729   Mean   :0.3156
                   3rd Qu.: 67.40   3rd Qu.:0.0000   3rd Qu.:1.0000
                   Max.   :122.56   Max.   :1.0000   Max.   :1.0000
```

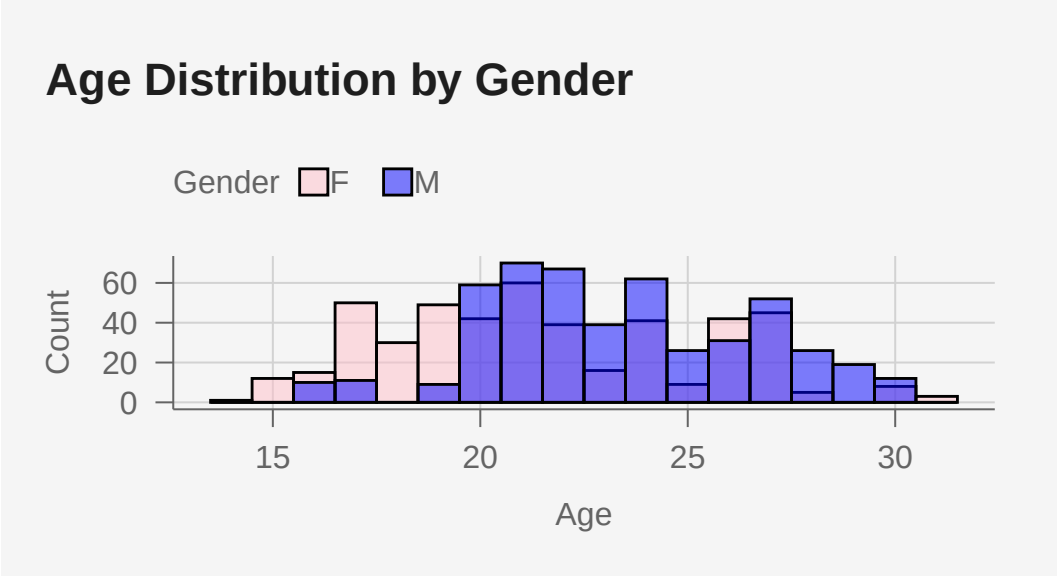
```
 is_olympics      pb_time_sec     avg_prev_2_time_sec   date_of_pb
```

```
Min.   :0.0000   Min.   : 21.07   Min.   : 21.45      Min.   :2020-09-06
1st Qu.:0.0000   1st Qu.: 49.40   1st Qu.: 50.80      1st Qu.:2021-01-14
Median :0.0000   Median : 54.67   Median : 54.99      Median :2021-07-24
Mean   :0.1281   Mean   : 62.80   Mean   : 63.69      Mean   :2022-01-13
3rd Qu.:0.0000   3rd Qu.: 65.32   3rd Qu.: 66.64      3rd Qu.:2023-01-11
Max.   :1.0000   Max.   :120.30   Max.   :120.95      Max.   :2024-07-27
                 NA's   :115      NA's   :190
days_since_pb
Min.   :   0.0
1st Qu.:  49.0
Median : 193.0
Mean   : 376.4
3rd Qu.: 679.0
Max.   :1420.0
```

Right from the summary statistics above we can gain all sorts of insights about the data. Some important features are the gender ratio, the span of dates, and any of the binary variables. We see roughly equal amounts of each gender and see races from September 2020 to July 2024. The other binary variables can be described through their means. relay_indicator, taper, and is_olympics all have means lower than 0.33. This means a majority of the data set is individual races, in season meets, and not the Olympics.

Another aspect of summary statistics are missing data. The cleaned version of the data set only has missing values in lane, pb_time_sec, and avg_prev_2_time_sec. The lane information was the only missing data from the raw data set. This information was not given to us. The other two are feature engineered variables from the raw data. The previous best time is missing for 115 rows because there were 115 distinct first swims, which as the first swim, can not have a previous best. This is different from distinct swimmers in the data set, for which there are 75. If Jack Alexy swam both the 100 free and the 50 free, those are two distinct first swims. The average of a swimmers previous two best times has 190 missing rows because there are 190 distinct swims that do not have 2 identical swims (same swimmer and event) before it. This again has to do with the confinements of the data set. For other variables, the summary statistics are not as telling. For this, we will add some plots to better represent these variables.

Part of the analysis is age curves. Because of this, it will be important to check the age distribution.



As seen in the plot above, the age distribution is different between the genders. Females tend to be younger in this data set. From ages 20 to 24 there are more male swimmers but still a decent amount of female swimmers. An early observation from this is that females may peak earlier than males and this is why we are seeing younger ages for females than males. From this age histogram, we can see it will be beneficial to have different age curves for males and females.

There are 960 observations in this data set spread across 6 events. To ensure that all the events can be reasonably modeled, we will check the distribution.

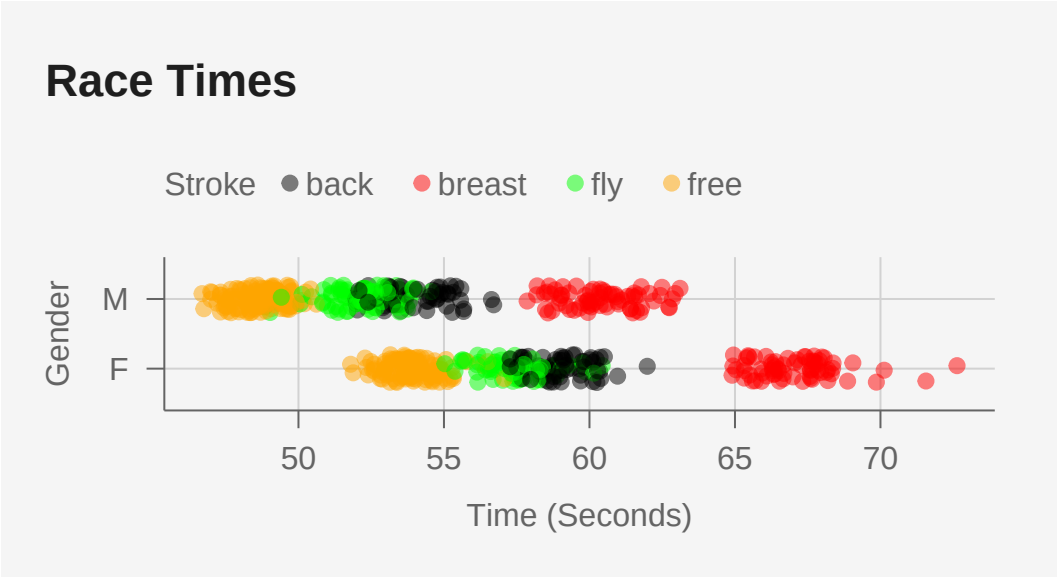
Table 1: Total Swims by Event

Event	Number of Swims
100 back	128
100 breast	139
100 fly	134
100 free	233
200 free	211
50 free	115

As seen in the table above, each event has over 100 races, which means there are plenty of data points for training the models. We will have to subtract out the 2024 Olympics races to use as a validation set, but as we saw from the binary variable is_olympics (which includes the 2021 Olympics as well), makes up a very small portion of the data set. This means each model will be able to have meaningful results. The 100 and 200 free have the most events,

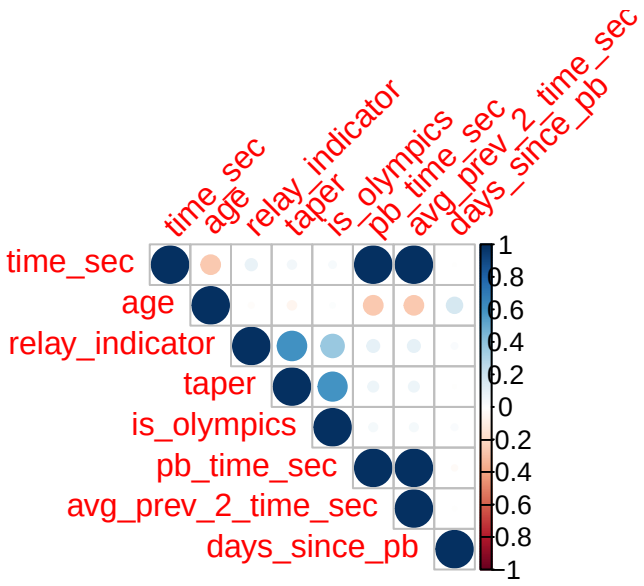
but this makes sense. Free has its own relays where 4 people swim freestyle in the same race. Free is the most swam race, so it makes sense that the data reflects this.

Another interesting plot is a scatter plot of events.



This plot shows why each stroke and distance needs its own model. Each stroke has many differences from the others. Beyond the end time, the technique and strategy of each race is unique. Because of this, we would expect that each race will need its own model. The only two races that seem similar time wise are back and fly. And ignoring all other factors, fly races have the ability to be changed by a relay start. This alone is a reason to have separate models.

The last bit of EDA we will display is a corrplot.



This corrplot shows that not many of our numeric variables are strongly correlated with the race time. Other than the variables that are directly linked to times, the highest magnitude of correlation is with age and even that seems to be far under 0.5. This is part of the reason why swim is so hard to model and predict.

Modeling and Analyses

There were 3 analyses run in the project: linear mixed effects modeled predictions, relay building based on the prediction models, and age curves. We will start by discussing the linear mixed effects models.

Linear Mixed Effects Models

As mentioned before, swimming is a sport that is very athlete dependent. Because of this, a simple linear model will not capture the difference in each athlete. Having a linear mixed effects model allows each swimmer to have their own intercept.

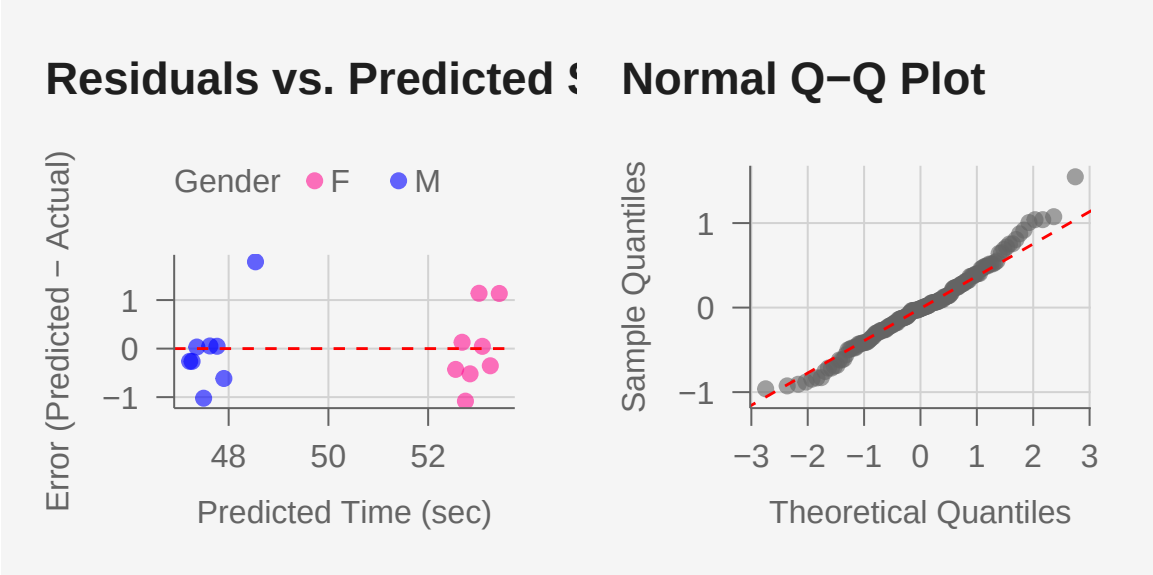
There are 4 assumptions for a linear mixed effects model.

- Assumptions
  - Linearity - the variables used scale linearly with time.
  - Normality of Residuals - the residuals (actual - predicted) are normally distributed around 0.

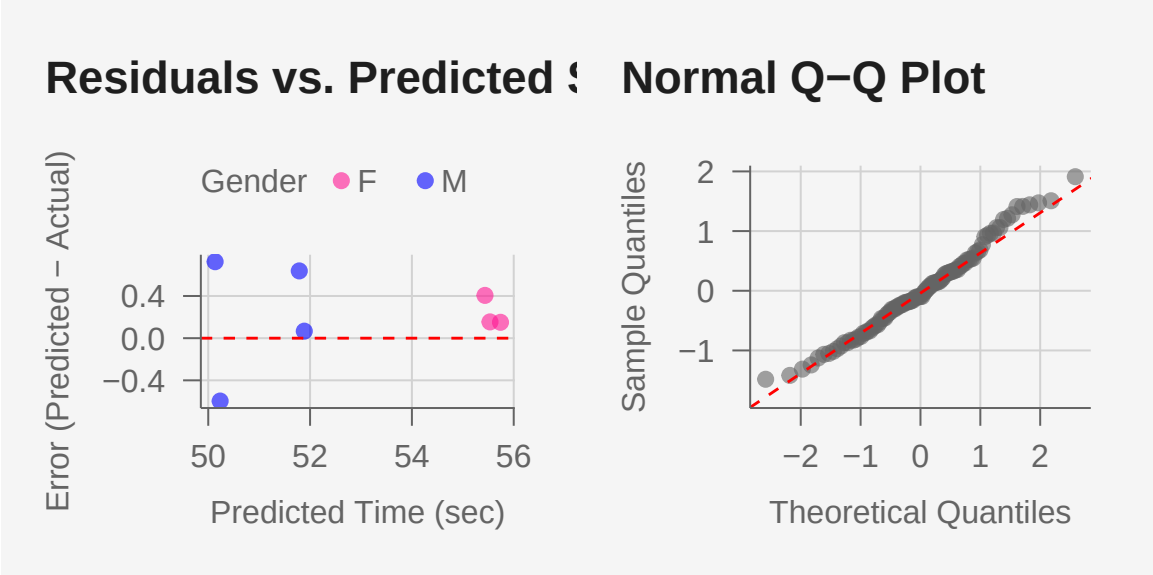
- Homoscedasticity - the spread of errors is consistent across the model.
- Normality of Random Effects - the random swimmer effects should be normal. ie assume most swimmers are average, some are faster, and some are slower.

The first and third assumptions can be checked by plotting. For the linearity and homoscedasticity assumption we can look at the residuals vs predicted plots below. In order to uphold the assumption, we are looking for an even spread around the 0 point. If the predictions are linear, we will see an even scattering above and below the 0 line. If the predictions are homoscedastic, we will see the vertical spread of the dots at the faster times is the same as the vertical spread of the slower times. Note, if we did not build a separate model for each distance and stroke the homoscedasticity assumption would surely fail as we are able to be much more precise in a 50 free than we are in a 100 breast. The normality of residuals will be tested with a QQ plot.

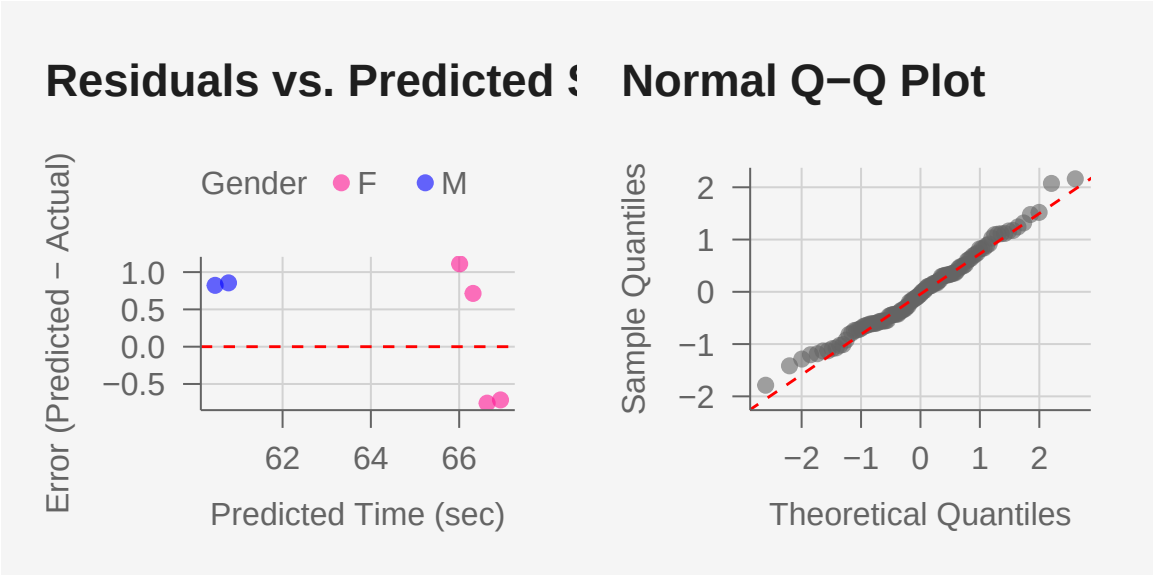
100 Free



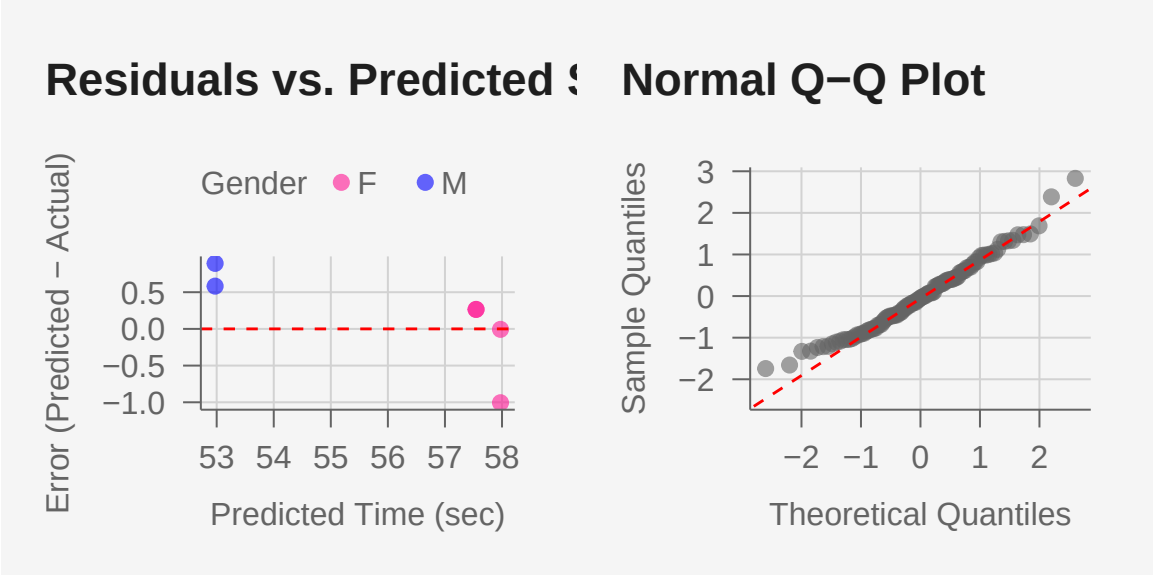
100 Fly



100 Breast



100 Back



Going by stroke (leaving out 50 and 200 free), we see that the 100 free model holds these two assumptions. The error vs predicted plot is scattered around the 0 mark and the mens' and womens' times have similar vertical spread. The other events have less validation points, so it is harder to tell.

The 100 fly follows linearity because the points are scattered around the 0 without any discernible pattern. Homoscedasticity does not hold up as well because we can see that the spread of the womens' times are less than that of the mens' times.

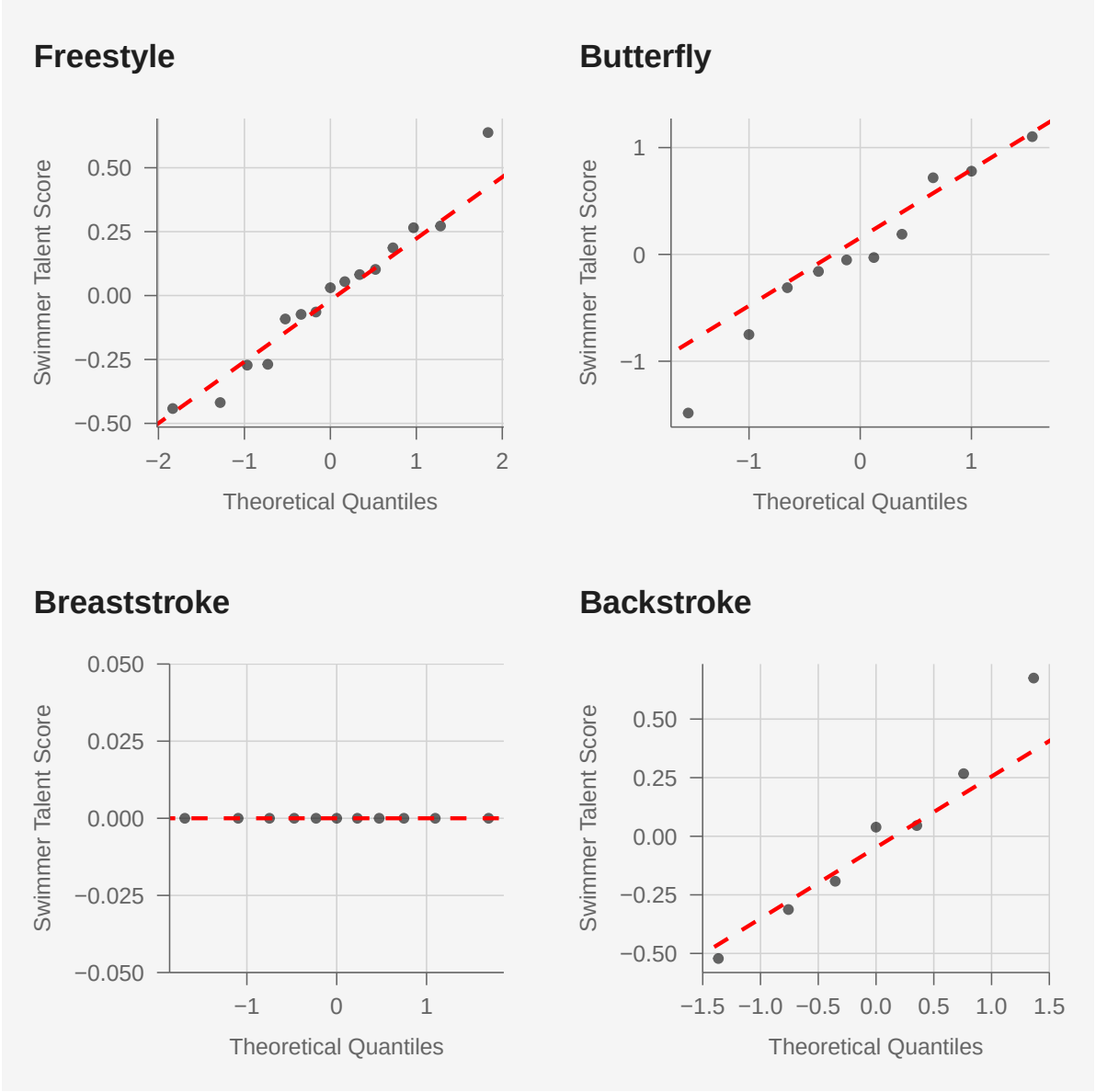
The 100 breast and 100 back follow linearity because the points are scattered around the 0 without any discernible pattern. Homoscedasticity does not hold up as well for either because we can see that the spread of the mens' times are less than that of the womens' times.

Although the strokes other than free do not exactly follow homoscedasticity, we do not view this as a problem due to the small sample size. In each case of a group having less spread than the other, there were very few validation points. It is hard to tell if this is violating homoscedasticity or a product of having too few points.

As seen in the QQ plots above, the residuals appear to be normal for all the strokes. For free and fly, the right tail wavers a little and for breast and back the left tail wavers a bit. Overall these QQ plots are encouraging that the model does hold the normality of residuals assumption.

The final assumption is the normality of random effects. This means that we would expect most swimmers to be average and have a 'talent score' of 0. Talent score refers to the individual effect of each swimmer (deviation of their intercept from the average). Negative talent scores

means the swimmer is faster than average and positive scores means they are slower than average.



Based on the QQ plots, we can see that the random effects are almost normally distributed. There does seem to be some snaking around the line for free, fly, and back. This slight fluctuation around the line is expected as it means that our distribution is skewed. This is because for all swimmers we would expect to see a bell curve of performance. However, only dealing with professionals, and mostly professionals that go to the Olympics, we are dealing

with the best of the best. This means in the bell curve of all swimmers, we are seeing essentially only the very far out tail. This means that we are inherently looking at skewed data. Because of this, we expect predictions to be a little slow. We expect the top swimmers to be pulled down toward the mean. This is especially true when validating only using Olympic races. We will be validating models based on the best swimmers at the fastest meet. Although we expect predictions to be a little slow, this project is still meaningful. The QQ plots only snake slightly, so we do not expect our predictions to be very slow, but we would expect slight bias towards a slower prediction. Since we know the predictions will be slow, we can still look at the order of the predictions and when optimizing relays, we should still find meaningful insights as to which relays will be the fastest.

As for the model itself, the observations are single races. The columns are described as before: year, name, gender, age, stroke, distance, date, meet, lane, relay_indicator, taper, is_olympics, pb_time_sec, avg_prev_2_time_sec, date_of_pb, and days_since_pb. Not all of these are used in every model, but these are the options. The outcome is time_sec. This is the time of the race. The training is the entire data set except the 2024 Olympics. The validation set is all the races during the 2024 Olympics.

100 free

The linear mixed effects model for free is:

$$\begin{aligned}\text{Time}_{ijk} = \beta_0 &+ \beta_1(\text{Taper}) + \beta_2(\text{Relay}) + \beta_3(\text{Gender}) \\ &+ f(\text{Age}) + \beta_4(\text{AvgPrev2}) + \beta_5(\text{PB}) \\ &+ u_{j(\text{name})} + v_{k(\text{meet})} + \epsilon_{ijk}\end{aligned}$$

$f(\text{Age})$ is a natural spline with 3 degrees of freedom.

$u_j \sim N(0, \sigma^2_{\text{name}})$ (Swimmer random intercept).

$v_k \sim N(0, \sigma^2_{\text{meet}})$ (Meet random intercept).

The coefficients of this model are telling and match what we would expect. First, the intercept (47.41) is for a female swimmer who is not tapered and not in a relay. Because of this we see negative coefficients for taper (-0.87) and relay_indicator (-0.41). This makes sense as we would expect swimmers to be faster when tapered and faster when in a relay. We also see a large negative coefficient for gender (-4.49), which also makes sense as men are faster than women. All these coefficients are significant as seen from the t values ($|t|>2$ means significant). We see neither the avg_prev_2_time_sec and pb_time_sec are insignificant. The age curve is also not significant for the 100 free. Based on these coefficients, we see that the model relies heavily on the random effect on name to capture a swimmer's true time.

In order to see how this model performs we look to the standard deviation of the residuals and the random effects. The standard deviation of the residuals (0.50) indicates that on average

the predictions deviate half a second from the actual Olympic times. This is might not seem ideal as a predictive model, but overall a half second is enough to distinguish an Olympic winner in the 100 free from other swimmers in the same final. The standard deviation of the meet and name random effect are 0.413 and 0.364 respectively. This means that the random effect on meet has a larger change and is essential for modeling. The random effect on names is also necessary but has a tighter range. The last measure of fit is the residual distribution. The median being at -0.022 means that there is very little bias in this model. This indicates that the model is not consistently over or under predicting times. The maximum residual (3.09) is relatively large compared to the minimum residual (-1.91) which means that there is a least one case, and maybe a few, where a swimmer under performs compared to their prediction. The ends of the residuals essentially check to see if there are outliers.

100 fly

$$\begin{aligned}\text{Time}_{ijk} = \beta_0 &+ \beta_1(\text{Taper}) + \beta_2(\text{Relay}) + \beta_3(\text{Gender}) \\ &+ f(\text{Age}) + \beta_4(\text{AvgPrev2}) \\ &+ u_{j(\text{name})} + v_{k(\text{meet})} + \epsilon_{ijk}\end{aligned}$$

$f(\text{Age})$ is a natural spline with 3 degrees of freedom.

$u_j \sim N(0, \sigma^2_{\text{name}})$ (Swimmer random intercept).

$v_k \sim N(0, \sigma^2_{\text{meet}})$ (Meet random intercept).

We see a lot of the same from free to fly. Again the intercept (53.27) is for a female swimmer who is not tapered and not in a relay. Every binary variable's coefficient is still the direction we would expect. Negative for relay_indicator (-0.09), taper (-1.40), and genderM (-3.66). For 100 fly, the first part of the age curve is significant with a coefficient of -3.89. This means that swimmers that race 100 fly peak earlier. The other parts of the age curve, avg_prev_2_time_sec, and taper are all not significant.

In order to evaluate the model fit, we will again look at the standard deviation of the residuals and the random effects. The standard deviation of the residuals is 0.78. This is a bit higher than what we saw for the 100 free. This indicates that the difference in predicted vs actual time is about 0.78 seconds. This again is able to differentiate gold medal winning swimmers from swimmers in the same race. We would also expect this to be higher as fly is a more technical stroke which causes more variance. The standard deviation of name is 0.91. This is higher than what we saw for freestyle, but makes sense. Again since fly is such a demanding stroke, the race depends more on the swimmer than free. This means the model relies heavily on this random effect. The model summary gives a warning at the bottom that says boundary (singular) fit. This means that one of the random effects did not really have an effect at all. The standard deviation for meet is 9.687e-08, which is essentially 0. This means that the random effect on meet had no impact on the model. This supports the theory that since fly is

so demanding, it depends less on the atmosphere and more on the swimmer. Lastly, looking at the distribution of the residuals, we see the median is -0.12. Because of this there might be a slight bias towards predicting faster times. The maximum (2.46) and minimum (-1.91) residuals are almost evenly spread around the median, though there is a slight increase in the maximum indicating slow predicted outliers. Through this, we trust the model's predictive ability.

100 breast

$$\begin{aligned}\text{Time}_{ijk} = \beta_0 &+ \beta_1(\text{Taper}) + \beta_2(\text{Relay}) + \beta_3(\text{Gender}) \\ &+ f(\text{Age}) + \beta_4(\text{AvgPrev2}) + \beta_5(\text{PB}) \\ &+ u_{j(\text{name})} + v_{k(\text{meet})} + \epsilon_{ijk}\end{aligned}$$

$f(\text{Age})$ is a natural spline with 3 degrees of freedom.

$u_j \sim N(0, \sigma^2_{\text{name}})$ (Swimmer random intercept).

$v_k \sim N(0, \sigma^2_{\text{meet}})$ (Meet random intercept).

We see a lot of the same from fly to breast. Again the intercept (53.56) is for a female swimmer who is not tapered and not in a relay. Every binary variable's coefficient is still the direction we would expect. Negative for relay_indicator (-0.30), taper (-0.53), and genderM (-4.89). For 100 breast, the second part of the age curve is significant with a coefficient of -3.09. This means that swimmers that race 100 breast peak in their mid twenties. The other parts of the age curve, avg_prev_2_time_sec, relay_indicator, and taper are all not significant. The only significant predictors are gender, the middle of the age curve, and pb_time_sec (0.46).

In order to evaluate the model fit, we will again look at the standard deviation of the residuals and the random effects. The standard deviation of the residuals is 0.87. This is a bit higher than what we saw for the 100 fly. This indicates that the difference in predicted vs actual time is about 0.87 seconds. This again is able to differentiate gold medal winning swimmers from swimmers in the same race. We would also expect this to be higher as breast is the slowest stroke and therefore has the most variance and is the hardest to predict. The standard deviation of meet is 0.71. This indicates that the atmosphere of a meet is important to modeling the 100 breast. The standard deviation for name is 0. This shows that the swimmer racing has no effect on the model. We see again the boundary (singular) fit warning. Breast, being the slowest stroke, makes sense as the lowest name random effect. Although, the result of name having no influence at all is surprising. Lastly, looking at the distribution of the residuals, we see the median is -0.054. This shows nearly no bias. The maximum (2.49) and minimum (-2.06) residuals are nearly evenly spread around the median. There is still a higher maximum magnitude than minimum magnitude which indicates some slow predictions again. Through this, we trust the model's predictive ability, but expect the predictions to be the worst out of all of the models.

100 back

$$\begin{aligned}\text{Time}_{ijk} = \beta_0 &+ \beta_1(\text{Taper}) + \beta_2(\text{Gender}) \\ &+ f(\text{Age}) + \beta_3(\text{AvgPrev2}) \\ &+ u_{j(\text{name})} + v_{k(\text{meet})} + \epsilon_{ijk}\end{aligned}$$

$f(\text{Age})$ is a natural spline with 3 degrees of freedom.

$u_j \sim N(0, \sigma^2_{\text{name}})$ (Swimmer random intercept).

$v_k \sim N(0, \sigma^2_{\text{meet}})$ (Meet random intercept).

We see a lot of the same in back. This time, the intercept (53.27) is for a female swimmer who is not tapered. There is no relay_indicator for back since it is the lead off in the medley relays. Every binary variable's coefficient is still the direction we would expect. Negative for taper (-1.27), and genderM (-4.54). The age curve, avg_prev_2_time_sec, and taper are all not significant. The only significant predictors are gender and taper.

In order to evaluate the model fit, we will again look at the standard deviation of the residuals and the random effects. The standard deviation of the residuals is 0.90. This is a bit higher than what we saw for the other models. This indicates that the difference in predicted vs actual time is about 0.90 seconds. This again is able to differentiate gold medal winning swimmers from swimmers in the same race. This goes against what we expected, as back is very similar to free. The standard deviation for name is 0.53. This indicates that there is a strong effect in name and that swimmers have varying intercepts. Again, there is a boundary (singular) fit warning. This time it is for meet, same as it was for fly. This means there is no effect on the 100 back from the atmosphere of the meet. A potential reason for this is the start. No matter how high the blocks are, the start in backstroke remains constant. This can change for starts in the other strokes depending on the pool. The distribution of the residuals is very similar to free. The median is -0.053, which shows nearly no bias. However, we do see a larger maximum residual (3.12) than minimum residual (-1.92) which shows that some swimmers are under performing.

The entire output of the model summaries are in the Appendix.

Since all but one of the assumptions were met (we acknowledge that we do not have a random sample of swimmers) and since the models' predictive ability can be trusted, we would say a linear mixed effects model is appropriate for this kind of data. These models will be evaluated in the next section. As for the results of the model, the interpretability depends on how deep the user goes. If the user takes the predictions at face value, then the results are very easy to understand. The output of the predictions is a time that can be directly compared to a swimmer's time at the Olympics. If the user dives further and begins to look at the coefficients and residuals, then the above explanation will be necessary.

Relay Combinations

The relay building analysis follows directly from the predictive modeling. We use the models to build out every possible combination of relay. The possible swimmers are those in the data set that have swam the specific event needed for a leg in the relay. For example, in the men's 4x100 Medley relay, the options for fly are any male swimmers that have a documented 100 fly race in the data set. This is not limited to those who swam at the Olympics.

For each free relay, we limit the combinations of possible swimmers. Since there was no leg factor in the data set, we do not believe that a team of A,B,C,D is any different than A,D,B,C. In other words, since the data set did not have any information on what the order of the relay is, we were unable to capture the effect of a second leg versus the anchor leg. Because of this, the only difference we were able to make was the first leg from the others since we do have a relay_indicator variable. Thus, for free relays, A,B,C,D = A,C,D,B. Therefore, when combining the results, we only allow for one iteration of the same three relay start swimmers. This is different for the medley relays since each leg is a different stroke. For this, we allowed every combination.

For this, we assume that the first leg is a flat start and the rest are relay starts. This is reflected in how we predict the times. We also assume that the swimmers are tapered. This is again met and reflected in how we predict the times. The model assumes all relay starts are created equal. This is not true, but will be considered true for this model. In other words, we will assume that all relay starts are safe, but not slow.

This is not a standard analysis where there are new models or observations or predictors. This simply puts relays together for Team USA to easily have predictions as to what will be the fastest combination of swimmers to maximize the probability of getting a medal.

Age Curves

The last analysis is the age curve. We had to use a slightly different model for the age curves so that we would be able to capture the different aging process for men and women.

The assumptions are

- Assumptions:
  - All swimmers are Average - we ignore the random effects because we want to follow general age trends not the age trend for each swimmer.
  - Swimmers are tapered - we want to model peak performance, so the swimmers must be tapered.
  - Smooth trends - we are using natural splines which assumes a smooth trajectory of aging.
  - Men and Women peak at different ages - we assume that men and women are not the same with respect to the age they peak at.

We meet all but one of these assumptions. First, the way we set up the model, we set re.form = NA, as part of the predict function. This means the predict function is ignoring the random effects and is making age curves for the average swimmer. However, this data set only consists of professional swimmers. None of these swimmers are average by any means. Because of this, the average in this data set will not reflect the normal human physiology. The results may be skewed.

Swimmers are tapered is also met in the predict function. The new data set we are testing on only contains tapered swimmers so that we are able to get peak performance at each age.

Smooth trends are biologically true. We know that as swimmers reach their 20s, they are getting faster bit by bit and after their peak, they begin to fall off. There are not usually huge jumps in performance. This is less true for younger swimmers. Often times, young teenagers will see massive jumps in their performance. After maybe a year or two of jumps, ability steadily inclines to the peak.

Men and women biologically peak at different times. We model this by using an interaction term of gender and age. Based on Rüst et al. 2012, in "Women achieve peak freestyle swim speed at earlier ages than men" we hypothesize that women will peak at a younger age than men.

The observations and outcomes are the same as before. The observations are races, the outcome is time. The predictors are slightly different, but a combination of the same variables from above. The predictors in this case are age, gender, taper, meet, and swimmer.

$$\begin{aligned}\text{Time}_{ijk} = \beta_0 &+ \beta_1(\text{Taper}) + \beta_2(\text{Gender}) \\ &+ \underbrace{\beta_3 f_1(\text{Age}) + \beta_4 f_2(\text{Age}) + \beta_5 f_3(\text{Age})}_{\text{Baseline Age Curve}} \\ &+ \underbrace{\beta_6[f_1(\text{Age}) \times \text{Gender}] + \beta_7[f_2(\text{Age}) \times \text{Gender}] + \beta_8[f_3(\text{Age}) \times \text{Gender}]}_{\text{Gender-Specific Shape Adjustment}} \\ &+ u_{j(\text{name})} + v_{k(\text{meet})} + \epsilon_{ijk}\end{aligned}$$

Where:

i, j, k: Observation i for swimmer j at meet k.

$f_d(\text{Age})$: The natural cubic spline basis functions ($df = 3$). This creates the "curved" line.

The Interaction ($\times$Gender): These terms allow the bend of the curve to change depending on gender. For example, it allows women to peak at age 22 while men peak at age 25.

$u_j \sim N(0, \sigma^2_{\text{name}})$: Random effect for the swimmer (Talent).

$v_k \sim N(0, \sigma^2_{\text{meet}})$: Random effect for the meet (Conditions).

The above model is fit and used to make predictions about how an average tapered swimmer will do at each age. This creates a curve of ages (which are broken up using a natural spline with df=3) for each gender and each stroke. The results are extremely easy to understand. The x axis of the plot is age and the y axis is predicted time. The lower the curve is, the faster the predicted time is. At the lowest point is the age that a swimmer of the chosen gender and event will peak.

Visualization and Results

Prediction Models

To evaluate the results, we will look at 2 plots and 5 metrics based on the validation set of Olympic races. The first plot is a predicted time by error plot. We already saw this one when checking assumptions, but it can be telling for results too. For better results, we hope to see errors around 0. The second plot is the actual Olympic time by the model's predicted time. On this plot, there is a y=x line. We hope to see the points closely scattered around this line to indicate good predictions. Beyond visual results, we have a few numeric metrics:

- Numeric Metrics:
  - Mean Average Error (MAE) - The average amount of seconds our validation predictions were off by.
  - Mean Squared Error (MSE) - The average of the squared errors (mathematical penalty, no direct relation to swim).
  - Bias - The skew of the model. Bias of 0 means no skew, Positive bias means swimmers are predicted too slow, Negative bias means swimmers are predicted too fast.
  - Root Mean Squared Error (RMSE) - Similar to MAE, except there is a penalty for large outliers.
  - Pearson Correlation - Measures the ranking ability of the model. A correlation of close to 1 means the model is good at predicting the rank of the swimmers (the predictive model might not be perfect, but it gets the order of the swimmers correct). A correlation of close to 0 means the swimmers are all jumbled in no particular order. A correlation of close to -1 means the model got the order of swimmers backwards. This is very important. We are expecting the simulations to be slightly slow, but if the order is correct, this project can still be useful in choosing relay teams.

100 free

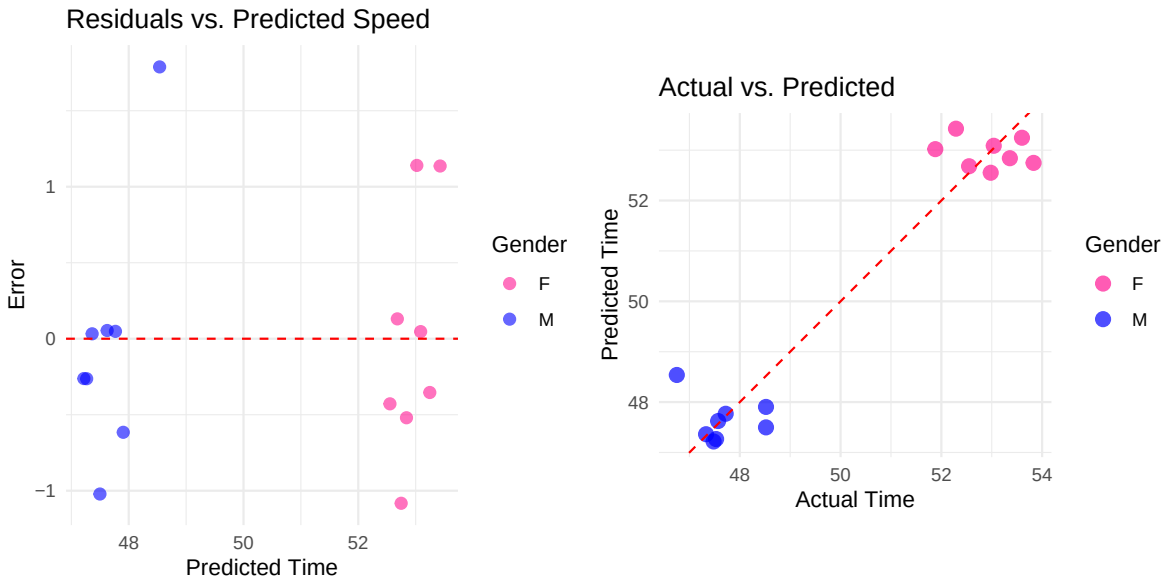


Table 2: Overall Model Performance

Metric	Value
MAE (Mean Absolute Error)	0.558
MSE (Mean Squared Error)	0.571
Bias (Avg Error Direction)	-0.011
RMSE (Root Mean Squared Error)	0.755
Pearson Correlation	0.961

As seen from the residual plot, most of the validation points hover around 0. There are a few outliers, most notably on the men's side there is an outlier of nearly 2 seconds. This is Hunter Armstrong, which we will discuss him later. Overall, this plot is promising that there are good predictions. This is furthered by the actual by predicted plot. We see most of the points gathered around the line where the predictions would be a perfect match. From this plot we do not expect any bias.

To ensure what we see on the plot is true, we check the metrics. First we look at MAE, the mean average error (average amount of seconds we expect to be off by). The MAE is 0.558, which shows that our predictions are off by about half a second. Looking back to the plot, we see that our predictions are likely a bit better than half a second off. In the 100 free there are some outliers. Hunter Armstrong, who went to the Olympics to swim the 100 back, ended up having the fastest time in this entire data set by about half a second. Because he has such an

incredible swim at the Olympics, no amount of data would have been able to predict this. The MAE would be much less without his near 2 second different in predicted and actual times. This outlier effect is also displayed in the RMSE. The RMSE is sensitive to outliers. If the RMSE is larger than the MAE, it means we have outliers in our model. With an RMSE of 0.755, we see that we do indeed have outliers. There is almost no Bias in this model (-0.011). This is a good sign, but we need to be careful with how we interpret this because of the outliers. There appear to be a few outliers on each side of the residual plot, which shows that there truly might not be any bias in this model. The MSE has no direct ties to swim results, but can still be interesting to look at. This is a mathematical penalty. For a race that is about 50 seconds long (average of mens and womens), an MSE of 0.571 is a small penalty. Looking to the Pearson Correlation, we see that the model does predict the correct order of swimmers for the most part. A correlation value of 0.961 means that the model is predicting the order of the swimmers nearly perfectly. This is promising for the choosing relay combinations.

100 fly

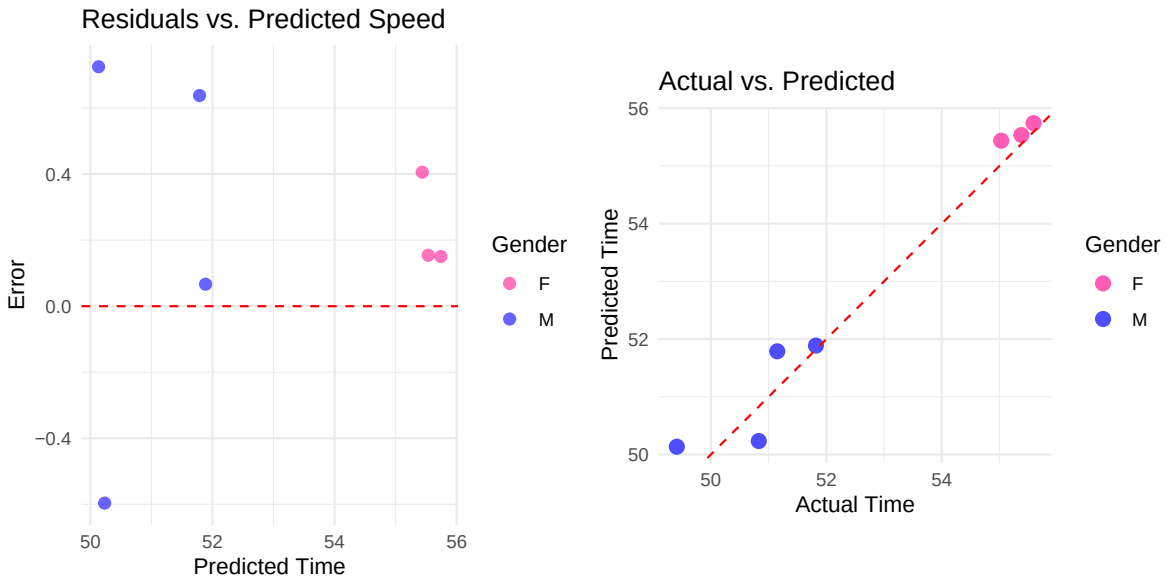


Table 3: Overall Model Performance

Metric	Value
MAE (Mean Absolute Error)	0.391
MSE (Mean Squared Error)	0.215
Bias (Avg Error Direction)	0.220
RMSE (Root Mean Squared Error)	0.463

Metric	Value
Pearson Correlation	0.985

Looking first to the residual plot, we see a much tighter interval than for the 100 free. We also see most of the data points fall above the line, which indicates a small positive bias. The actual by predicted plot tells the same story. Here is it clearer that the predicted times are closer to the line than they were for free, especially for the women.

As expected from the plots, the MAE is lower for the fly, 0.391, than it was for the free. This means we can predict the 100 fly with an error of 0.39 seconds on average. Also as expected, we see some bias, 0.220. This confirms what was inferred from the residual plot. A bias of 0.220 matches what we expected from the assumptions. The assumption that the random effects would be normal was not strictly upheld. Because of this, we expected the predictions to be slow, which is reflected by a positive bias. The RMSE is 0.463, which is higher than the MAE, but not that much higher that we would expect outliers for 100 fly. The MSE is 0.215, which again, as a penalty for a race that is roughly 53 seconds long, is very low. Lastly, we see that the Pearson Correlation is strong again with an R value of 0.985. This again means the model nearly perfectly orders the predicted times.

100 breast

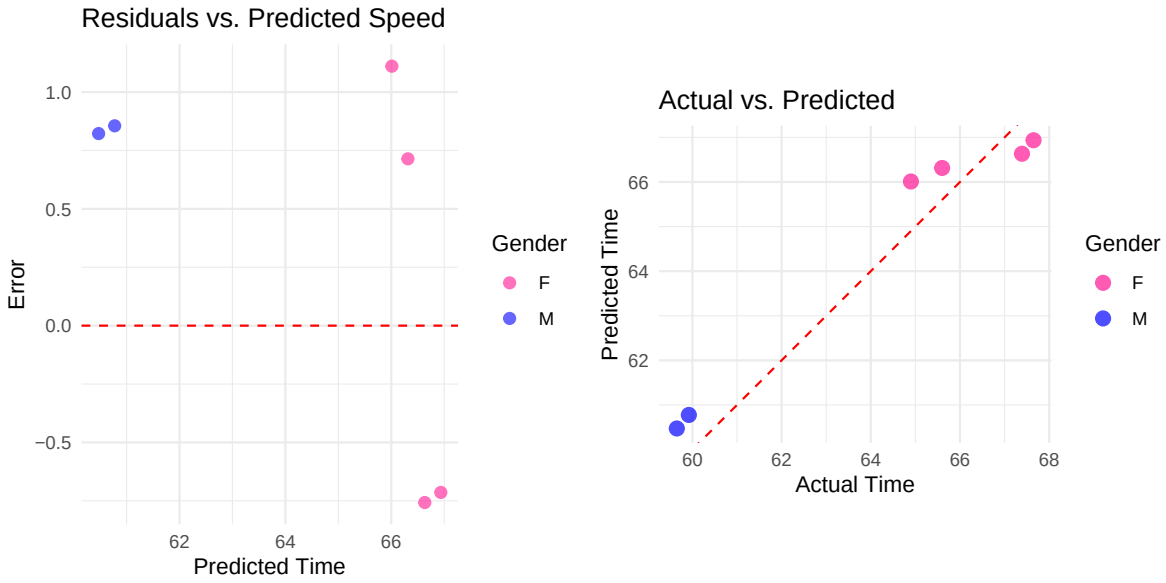


Table 4: Overall Model Performance

Metric	Value
MAE (Mean Absolute Error)	0.829
MSE (Mean Squared Error)	0.706
Bias (Avg Error Direction)	0.339
RMSE (Root Mean Squared Error)	0.840
Pearson Correlation	0.980

Based on the model section above, we knew breast would be the hardest to predict. Looking at the plots, this seems to be true. The residual plot shows again that there will likely be a positive bias since most of the points lie well above the predicted line. This plot also does not have any points near the 0 line, which indicates that the predictions may not be very accurate. The same story is told by the actual by predicted plot. We see the points are around the line, but not necessarily near the line.

The metrics confirm what we expected from the model and what we saw from the plots. The MAE is 0.829, the highest by a decent margin. This means that for breast, we expect an error of 0.829 seconds for our predictions. The RMSE is slightly larger than the MAE with a value of 0.840, which means we do not expect outliers. This is not surprising because the predictions are already off. If we saw outliers, then the model might be useless. The bias is also the highest we see, with a value of 0.339. This is again somewhat expected because of the broken assumption. The MSE of 0.706, is also the highest. This shows mathematically that breast has the worst model. Despite all of the metrics of the breast model being worse than the other models, the Pearson Correlation is still high. The R value is 0.980. Even though the breast model does not predict times very well, it still gets the order of the swimmers correct.

100 back

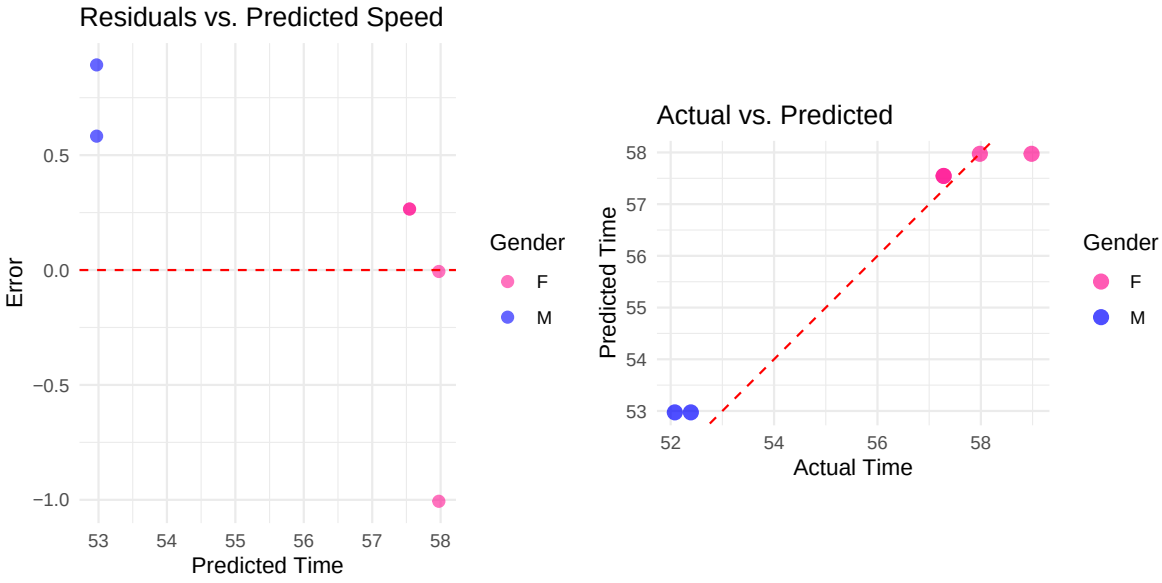


Table 5: Overall Model Performance

Metric	Value
MAE (Mean Absolute Error)	0.503
MSE (Mean Squared Error)	0.382
Bias (Avg Error Direction)	0.166
RMSE (Root Mean Squared Error)	0.618
Pearson Correlation	0.988

The residual plot for back shows a tighter interval than free and breast. Having said that, the plot shows a majority of the points not close to the error equals 0 line. This indicates slight bias and some outliers. The actual by predicted plot shows that the women's predictions are closer to the actual times than then men's predictions.

The metrics confirm this story. The 100 back has an MAE of 0.503, which is similar to the 100 free. It means that we expect to be off by about half a second for back predictions. As expected from the residual plot, we do see some outliers. The RMSE of 0.618 is higher than the MAE, which confirms the presence of outliers. There is some bias in this model. The bias of 0.166 is still low, but is likely caused by the violation of the normality of random effects assumption. The MSE (.382) is again low for a race that is roughly 55 seconds. The R value of 0.988 for the Pearson Correlation shows that the model is once again able to predict the correct order of swimmers.

As seen from the above breakdown, all the models are able to predict within a second of the actual time when tested on the Olympic race validation set (MAE less than 1 for all models). Along with predictions that yield close results and just a few outliers, the correlation of the predicted order and actual order was very high showing that these models are great for predicting the rank of swimmers at the Olympics. Between these two facts, these models help to predict times and can help to build relays that will optimize Team USA's chance of winning at the Olympics.

Relay Combinations

The results of the relay combinations are best seen in the app. We highly recommend the reader goes in and experiments with the app. Going by relay, we will discuss the predicted results compared to the actual results from the Paris Olympics.

Men's 4x100 Free Relay:

Predicted:

Total time = 190.22 seconds

Jack Alexy, Chris Guiliano, Matt King, Zach Apple.

Actual:

Total Time = 189.28

Jack Alexy, Chris Guiliano, Hunter Armstrong, Caeleb Dressel.

This was our best prediction. The predicted relay is less than 1 second off the real relay. As mentioned before Hunter Armstrong had a truly incredible swim in this relay. The fastest relay that featured Hunter in our predictions was the 69th ranked relay. This is because his predicted time is 48.48 and his actual time was 46.75. This split was the fastest on the team by over half a second and was not predicted by anyone, especially not our model. The other stark difference here is Caeleb Dressel and Zach Apple. Apple's predicted time was 47.71 and Caeleb's was 47.79. Essentially a toss up here, but would likely go to Caeleb because he is more experienced as a leader and in the Olympics. This is where having the ability to model the pressure that comes with being an anchor would be helpful and likely lean towards Caeleb as opposed to Apple.

Men's 4x100 Medley Relay:

Predicted:

Total Time = 209.29

Hunter Armstrong, Nic Fink, Caeleb Dressel, Jack Alexy.

Actual:

Total Time = 208.01

Ryan Murphy, Nic Fink, Caeleb Dressel, Hunter Armstrong.

This prediction is again about a second off, but again with extra data we can explain this difference. The main difference here, again comes from Hunter Armstrong. In the predicted model, he is predicted 0.01 seconds faster than Ryan Murphy. For those who follow swim, we know Ryan Murphy has some X factor that makes him swim faster in relays. This is something we did not figure out how to encode into our model. We did try a relay_indicator factor within the random effect of name, but did not get any better results. Given the hundredth of a second difference and a knowledge of Ryan Murphy's ability in relays, we likely would have gone with Ryan Murphy on back. Also, the predicted time for Murphy is slower because again, we can not predict this X factor yet. Another factor here, is that the Men's 4x100 Free relay is swam before the Men's 4x100 Medley relay. This means when choosing the relay team, the coaches had the extra information that Hunter Armstrong went faster than Jack Alexy in the 100 free. Because of this, it makes more sense to plug in Hunter. This again, is information that our model does not have and therefore would have been wrong to predict Hunter for the free leg of this relay. Most of the difference in time comes from Ryan Murphy being predicted much slower than he went in this relay.

Women's 4x100 Free Relay:

Predicted:

Total Time = 211.59

Kate Douglass, Abbey Weitzeil, Gretchen Walsh, Torri Huske.

Actual:

Total Time = 210.20

Kate Douglass, Gretchen Walsh, Torri Huske, Simone Manuel.

Again, we see a little over a second difference. Most of this difference comes from the fact that Simone Manuel is not in the data set. We are not sure why, but she does not show up. Beyond this, Torri Huske is about a second slow. She was another one of the outliers in the free model. Beyond these two facts, the prediction is very accurate.

Women's 4x100 Medley Relay:

Predicted:

Total Time = 231.33

Regan Smith, Lilly King, Gretchen Walsh, Kate Douglass.

Actual:

Total Time = 229.63

Regan Smith, Lilly King, Gretchen Walsh, Torri Huske.

The main difference in this relay is from the breast leg. Lilly King is predicted at 65.92, but swam a 64.90. Beyond this, the model performs very well. The difference in the free leg is understandable. Kate Douglass had a busy load during this Olympics and Torri Huske had an incredible swim in the 4x100 Free relay. This allowed Kate to focus on her other events while not losing any speed on the relay. This is information that our model does not have and would not have been able to predict. We allow the model to filter out certain swimmers for this reason exactly. Given that Kate was predicted to be the fastest, but would no longer be in the relay, we would filter her out and be able to generate predictions without her. Torri Huske again went faster than predicted.

Mixed 4x100 Medley Relay:

Predicted:

Total Time = 219.02

Regan Smith, Nic Fink, Gretchen Walsh, Jack Alexy.

Actual:

Total Time = 217.43

Ryan Murphy, Nic Fink, Gretchen Walsh, Torri Huske.

Considering what we have already discussed, we see a lot of the same trends. We see that Ryan Murphy was again left off from the relay. This is again the difference of Ryan Murphy, an experienced relay swimmer. The model predicted Murphy a whole second slower than he went. With more data on Ryan Murphy's swims, especially relays, we would be able to adjust better for this. The model also predicts Torri Huske almost a second slower than she went. Torri and Hunter were both breakout swimmers at the Paris Olympics. Our models do not predict breakout swimmers, but rather trends of swimmers.

Beyond a few outliers caused by break out swimmers, real time Olympic trends, and the inability to model experience in relays, the model performs nearly perfectly. Because of this, we are happy with the model we built and believe it is useful for Team USA to use as a quick baseline for who should be included in relays. It should not be solely trusted and should mix data driven decisions with real time coaching decisions.

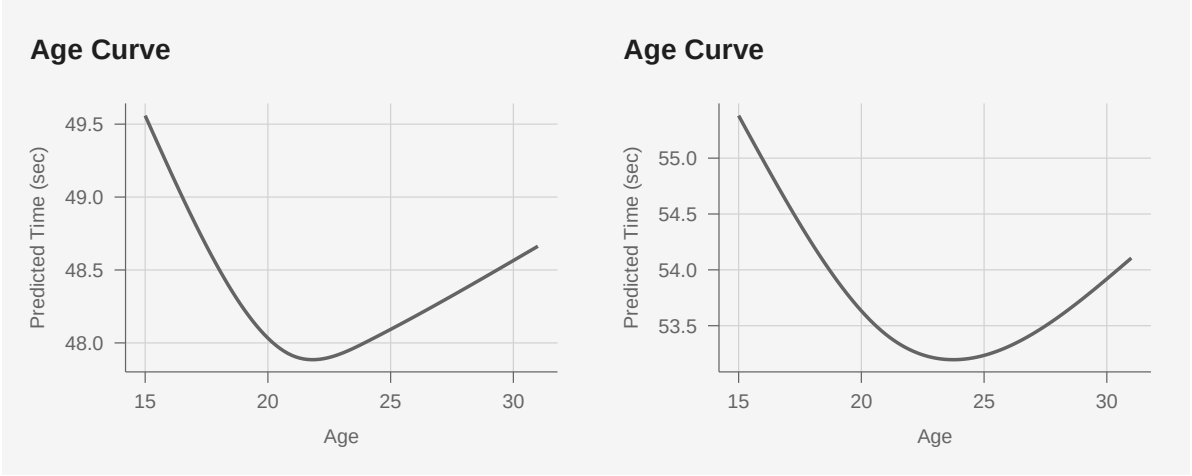
Age Curves

The age curve results may have been skewed from the fact that we do not have a proper representation of the swimming population. We are looking at the best swimming athletes in the country. Because of this we do not see as much separation between men and women as we predicted. This is true for the best athletes in any sport. Looking to the NBA and NFL, we see athletes like LeBron, Kevin Durant, Tom Brady, and Aaron Rodgers that are able to

play well past when their primes should be. The same happens in swim. An example of this is Michael Phelps, who made the Olympics at 16 and at 31. This is a much longer prime than anyone should experience. While not that extreme, most of the swimmers have a similarly extended prime.

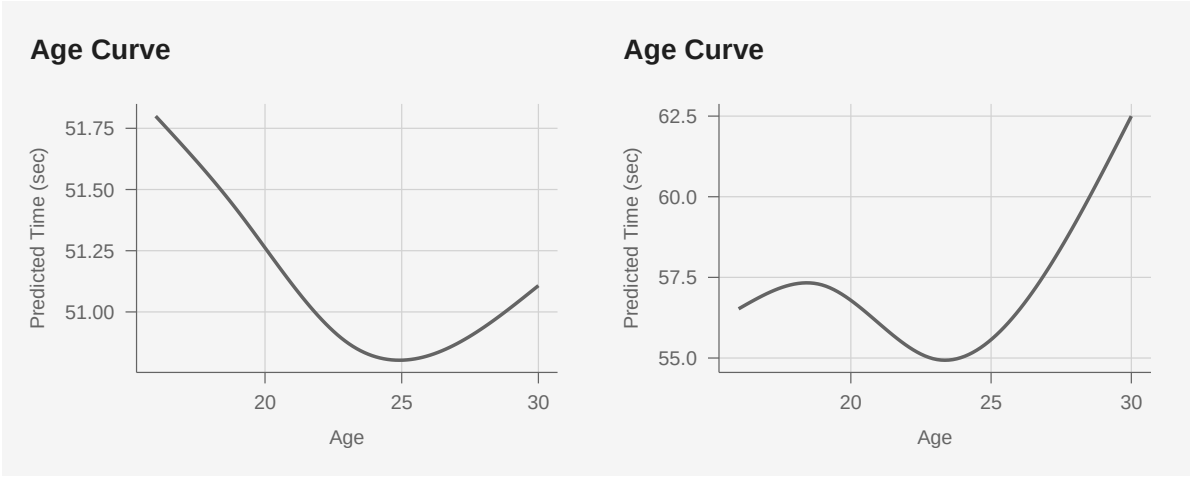
The results for the age curves are as follows. The lowest point in the graph is the age that swimmers peak for a given event. For all the strokes, the men's age curve will be plotted on the left, and the women's age curve will be plotted on the right:

100 free



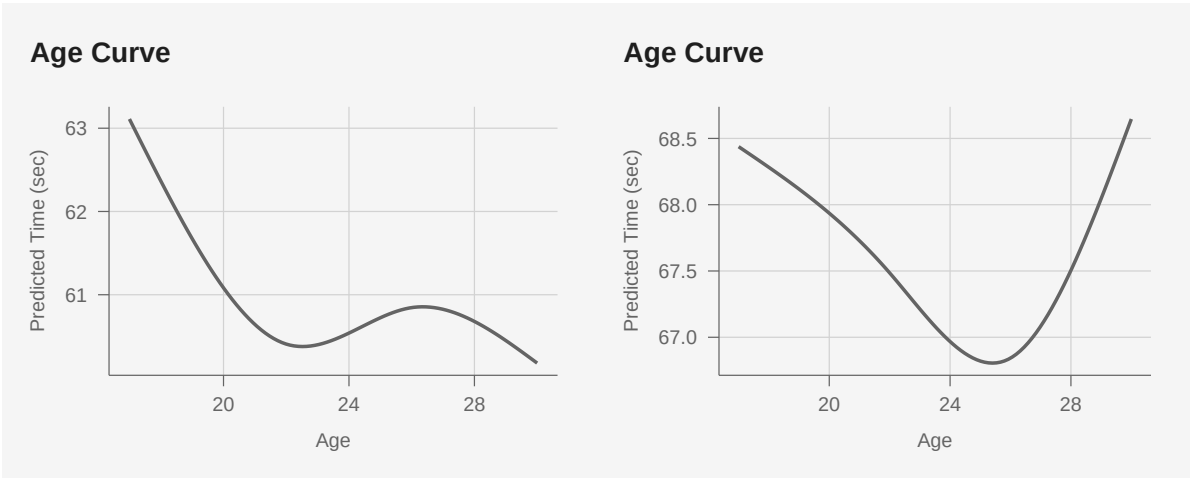
As seen, the left curve (men) seems to peak at a younger age than the right curve (women). This is the opposite of what we predicted. The trend for both is quite similar. Race time decreases from 15 to about 22 for men and about 24 for women, then race time begins to increase again.

100 fly



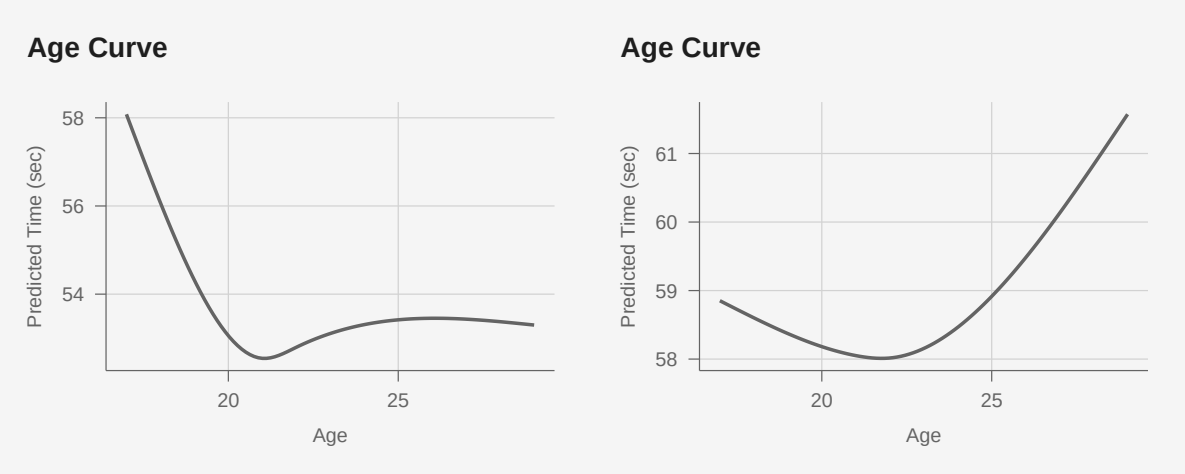
The fly age curves do perform as expected. The women's curve peaks a little bit younger, around 23, and then increases and the men's curve peaks at 25 and then increases. The women's age curve shows that as women get younger the trend would be that they are getting faster. This is a danger when working with splines. Someone in the data set likely had a fast time when they are young that is bringing that tail down. Since the spline only has three groups, that trend will continue downwards. This is why we cut off the graph. Without cutting off the graph, the age curve would predict a 1 year old as a world record holder which is obviously incorrect.

100 breast



The breast age curves are the opposite of what we expect. The men's curve seems to peak around 22 and the women's curve peaks around 25. We see a similar effect of the splines for the men's curve as we did in the women's 100 fly curve. According strictly to the age curve, the older you get, the faster your 100 breast will be. This again is obviously not true, and again shows why limiting the graph is so important.

100 back



This again shows the trend that men peak younger than women, which is against our hypothesis based on the Rüst et al. 2012 paper. According to these curves, men peak around 21 and then level out for the rest of time, which is again due to the nature of natural splines. This is not what we believe truly happens. The key take away here is that through 28, back times for men do not fall off much. For women, the peak is around 22 with a large fall off quickly after.

Overall, the age curves are not what we expected. According to our analysis, men generally peak younger than women. This is of course with a broken assumption that we are dealing with an average swimmer. Working with only professional swimmers, and the fastest of the professional swimmers at that, is likely skewing our age curves.

Conclusions and Future Work

As seen from the results of the relay combinations above, the relay building model predicts very well. The only flaws are those mentioned in the blurbs above. Most of these errors come from factors that occur during the Olympics. This is something that can not be in the model that we are using to predict events prior to the Olympics. Beyond Hunter Armstrong, Torri Huske, Ryan Murphy, and assuming we have everyone in the data set, these models work very well for putting together the fastest relay possible. At very least they nearly match what

Team USA put together as the fastest relay. This model should be used in conjunction with real time analysis of what is happening at the Olympics to determine relays. This could be a reassuring time saver for Team USA when determining relays. Instead of having coaches ponder for hours who should be on the relay, they can use this app. By using this model as a baseline, and having human input to keep up to date with the trends occurring during the Olympics, Team USA can start leveraging data to save time and make more informed choices for their relays.

Some future steps would be to add on extra years of data, find a true relay effect, and add variance to make the relay combinations a true simulation. Adding more years of data to this model will increase its accuracy for predicting events. If we had the data for the years prior to the 2021 Olympics, we would be able to add in an Olympic predictor to the model. Right now, we tried to add a column called is_olympics to the model. This would act as a more refined taper variable. This could create an "Olympic Effect", which could be useful to see how the best athletes perform at the Olympics compared to any other taper meet. Having more data to make a true relay effect would also be a good future step. We tried this, but ultimately our data fell short of a quality standard so we were unable to use it. Right now, each model, except back, has a generic relay effect. It would be more interesting to find the X factor we mentioned before to properly predict a swimmer like Ryan Murphy. This would yield better relay combinations. The last step we will suggest here is to turn the relay combinations into relay simulations. As of right now, our relay combinations do not take into account how consistent a swimmer is. This relay combination just takes the predicted time with no variance added. Adding a variance to each swimmer and then running simulations would add some insight into the combinations. This, along with adding in other countries predicted times, would allow for us to get probabilities of placement at the Olympics for each combination of team. A scenario in which this would be helpful, is say we have a swimmer that has high variance. They could be fantastic but on average are slower than another swimmer. We might want to choose this high variance swimmer as a chance to win gold as opposed to playing it safe and getting a silver. This was not done for this project because we did not have enough data on most given races. In order to find a per swimmer variance per race, we would need more data. We tried this, but got high variances for all swimmers and the results of the relays ended up being scattered.

References

Rüst CA, Knechtle B, Rosemann T. Women achieve peak freestyle swim speed at earlier ages than men. Open Access J Sports Med. 2012 Nov 12;3:189-99. doi: 10.2147/OAJSM.S38174. PMID: 24198602; PMCID: PMC3781914.

Wikipedia contributors. (2025, August 17). Incentives for Olympic medalists by country. In Wikipedia, The Free Encyclopedia. Retrieved 14:33, December 12, 2025, from https://en.wikipedia.org/w/index.php?title=Incentives_for_Olympic_medalists_by_country&oldid=1306342373

Appendix

The model summary for the 100 free:

```
Linear mixed model fit by REML ['lmerMod']
Formula: time_sec ~ taper + relay_indicator + ns(age_num, df = 3) + gender +
    avg_prev_2_time_sec + pb_time_sec + (1 | name) + (1 | meet)
   Data: train_df

REML criterion at convergence: 310.7

Scaled residuals:
     Min       1Q   Median       3Q      Max
-1.91933 -0.53490 -0.02211  0.49408  3.09160

Random effects:
 Groups   Name        Variance Std.Dev.
 meet     (Intercept) 0.1705   0.4130
 name     (Intercept) 0.1327   0.3642
 Residual             0.2506   0.5006
Number of obs: 166, groups:  meet, 43; name, 15

Fixed effects:
                     Estimate Std. Error t value
(Intercept)          47.41472    5.53848   8.561
taper                -0.87007    0.24548  -3.544
relay_indicator      -0.40615    0.16288  -2.493
ns(age_num, df = 3)1 -0.42731    0.55642  -0.768
ns(age_num, df = 3)2 -1.88968    2.11764  -0.892
ns(age_num, df = 3)3  0.12543    0.89544   0.140
genderM              -4.48887    0.55502  -8.088
avg_prev_2_time_sec   0.04529    0.09315   0.486
pb_time_sec           0.10349    0.09266   1.117

Correlation of Fixed Effects:
            (Intr) taper  rly_nd n(_,d=3)1 n(_,d=3)2 n(_,d=3)3 gendrM a__2__
taper       -0.036
relay_ndctr -0.229 -0.298
ns(g_,d=3)1 -0.198  0.039  0.024
ns(g_,d=3)2 -0.359  0.106 -0.011  0.514
ns(g_,d=3)3 -0.373  0.092  0.003  0.179     0.799
genderM     -0.904  0.004  0.239  0.032     0.120     0.210
```

```
avg_prv_2__ -0.534  0.018  0.126  0.066     0.067    -0.040     0.488
pb_time_sec -0.512 -0.004  0.129  0.035     0.148     0.330     0.489 -0.437
```

The model summary for the 100 fly:

```
Linear mixed model fit by REML ['lmerMod']
Formula: time_sec ~ taper + relay_indicator + ns(age_num, df = 3) + gender +
    avg_prev_2_time_sec + (1 | meet) + (1 | name)
   Data: train_df

REML criterion at convergence: 257.7

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.9088 -0.6351 -0.1238  0.5341  2.4611

Random effects:
 Groups   Name        Variance  Std.Dev.
 meet     (Intercept) 9.385e-15 9.687e-08
 name     (Intercept) 8.316e-01 9.119e-01
 Residual             6.031e-01 7.766e-01
Number of obs: 103, groups:  meet, 29; name, 10

Fixed effects:
                     Estimate Std. Error t value
(Intercept)          53.26641    6.91304   7.705
taper                -1.40378    0.24138  -5.816
relay_indicator      -0.09872    0.31491  -0.313
ns(age_num, df = 3)1 -3.89370    1.01890  -3.821
ns(age_num, df = 3)2 -1.25715    0.96256  -1.306
ns(age_num, df = 3)3 -1.16632    0.95337  -1.223
genderM              -3.66310    0.89275  -4.103
avg_prev_2_time_sec   0.08040    0.11977   0.671

Correlation of Fixed Effects:
            (Intr) taper  rly_nd n(_,d=3)1 n(_,d=3)2 n(_,d=3)3 gendrM
taper        0.157
relay_ndctr -0.151 -0.636
ns(g_,d=3)1 -0.303  0.029  0.047
ns(g_,d=3)2 -0.256  0.016  0.020  0.335
ns(g_,d=3)3 -0.270  0.017  0.042  0.362     0.673
genderM     -0.569 -0.121  0.074 -0.180    -0.073    -0.197
```

```
avg_prv_2__ -0.996 -0.165  0.153  0.302     0.220     0.268     0.521
optimizer (nloptwrap) convergence code: 0 (OK)
boundary (singular) fit: see help('isSingular')
```

The model summary for the 100 breast:

```
Linear mixed model fit by REML ['lmerMod']
Formula: time_sec ~ taper + relay_indicator + ns(age_num, df = 3) + gender +
    avg_prev_2_time_sec + pb_time_sec + (1 | name) + (1 | meet)
   Data: train_df

REML criterion at convergence: 309.9

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.06131 -0.65002 -0.05445  0.54956  2.49292

Random effects:
 Groups   Name        Variance Std.Dev.
 meet     (Intercept) 0.5085   0.7131
 name     (Intercept) 0.0000   0.0000
 Residual             0.7523   0.8673
Number of obs: 109, groups:  meet, 29; name, 11

Fixed effects:
                     Estimate Std. Error t value
(Intercept)          53.55574    8.26875   6.477
taper                -0.52774    0.42326  -1.247
relay_indicator      -0.30328    0.34378  -0.882
ns(age_num, df = 3)1 -0.16253    0.39504  -0.411
ns(age_num, df = 3)2 -3.08531    1.21362  -2.542
ns(age_num, df = 3)3  0.01601    0.45300   0.035
genderM              -4.89039    0.67874  -7.205
avg_prev_2_time_sec  -0.23837    0.15129  -1.576
pb_time_sec           0.46405    0.16024   2.896

Correlation of Fixed Effects:
            (Intr) taper  rly_nd n(_,d=3)1 n(_,d=3)2 n(_,d=3)3 gendrM a__2__
taper       -0.110
relay_ndctr -0.162 -0.273
ns(g_,d=3)1  0.229  0.033 -0.064
ns(g_,d=3)2 -0.625  0.071  0.085 -0.112
```

```
ns(g_,d=3)3 -0.522  0.088  0.082 -0.123     0.519
genderM     -0.936  0.081  0.151 -0.151     0.425     0.356
avg_prv_2__ -0.351 -0.004  0.016  0.055     0.219    -0.252     0.398
pb_time_sec -0.429  0.080  0.110 -0.240     0.248     0.639     0.337 -0.695
optimizer (nloptwrap) convergence code: 0 (OK)
boundary (singular) fit: see help('isSingular')
```

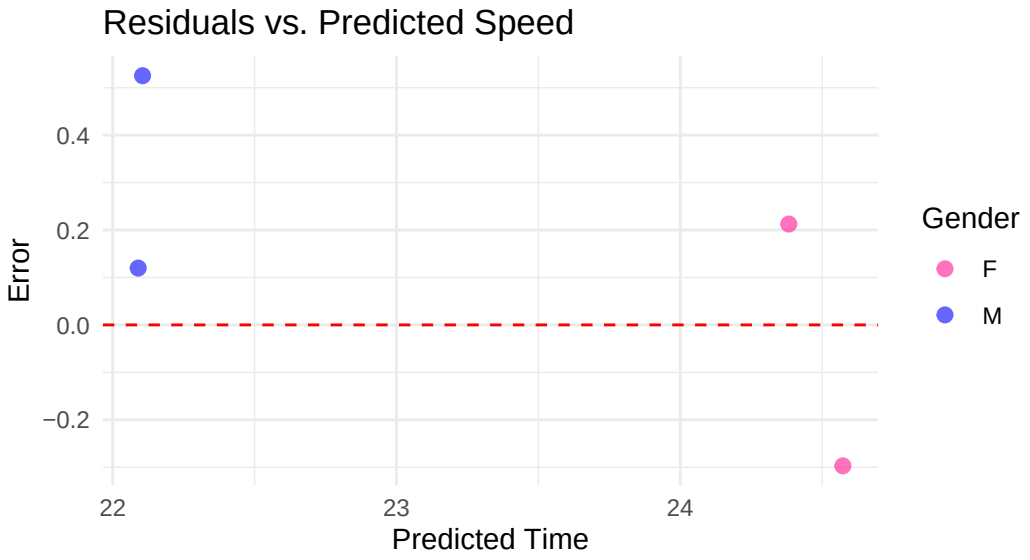
The model summary for the 100 back:

```
Linear mixed model fit by REML ['lmerMod']
Formula:
time_sec ~ taper + ns(age_num, df = 3) + gender + avg_prev_2_time_sec +
    (1 | name) + (1 | meet)
   Data: train_df

REML criterion at convergence: 291.4

Scaled residuals:
     Min       1Q   Median       3Q      Max
-1.92434 -0.75499 -0.05304  0.62256  3.12436

Random effects:
 Groups   Name        Variance Std.Dev.
 meet     (Intercept) 0.0000   0.0000
 name     (Intercept) 0.2795   0.5287
 Residual             0.8200   0.9056
Number of obs: 108, groups:  meet, 29; name, 7

Fixed effects:
                     Estimate Std. Error t value
(Intercept)          51.53326    7.61635   6.766
taper                -1.27241    0.20168  -6.309
ns(age_num, df = 3)1 -0.06741    0.61118  -0.110
ns(age_num, df = 3)2 -0.51789    1.33422  -0.388
ns(age_num, df = 3)3  0.21428    0.70245   0.305
genderM              -4.54069    0.80900  -5.613
avg_prev_2_time_sec   0.13904    0.12631   1.101

Correlation of Fixed Effects:
            (Intr) taper  n(_,d=3)1 n(_,d=3)2 n(_,d=3)3 gendrM
taper        0.038
ns(g_,d=3)1 -0.061  0.198
```

```
ns(g_,d=3)2 -0.282  0.064  0.466
ns(g_,d=3)3 -0.124  0.112  0.559     0.588
genderM     -0.804 -0.072 -0.128     0.025    -0.154
avg_prv_2__ -0.996 -0.051  0.026     0.213     0.095     0.800
optimizer (nloptwrap) convergence code: 0 (OK)
boundary (singular) fit: see help('isSingular')
```

The results for the 50 and 200 free will be posted here. These were not included as the bulk of this project was geared towards the 4x100 relays. However models were tested and evaluated for these events so we think it would be a shame to not post them somewhere. These analyses are all available in the app as well.

50 free



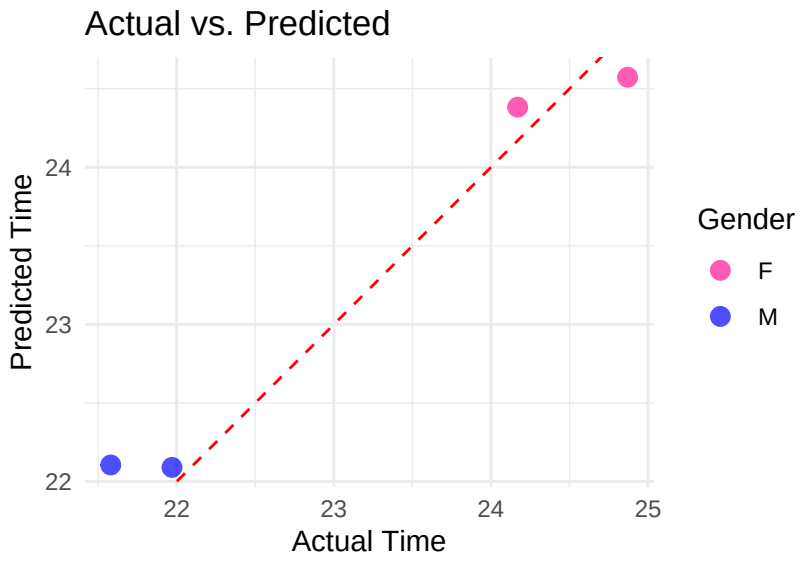
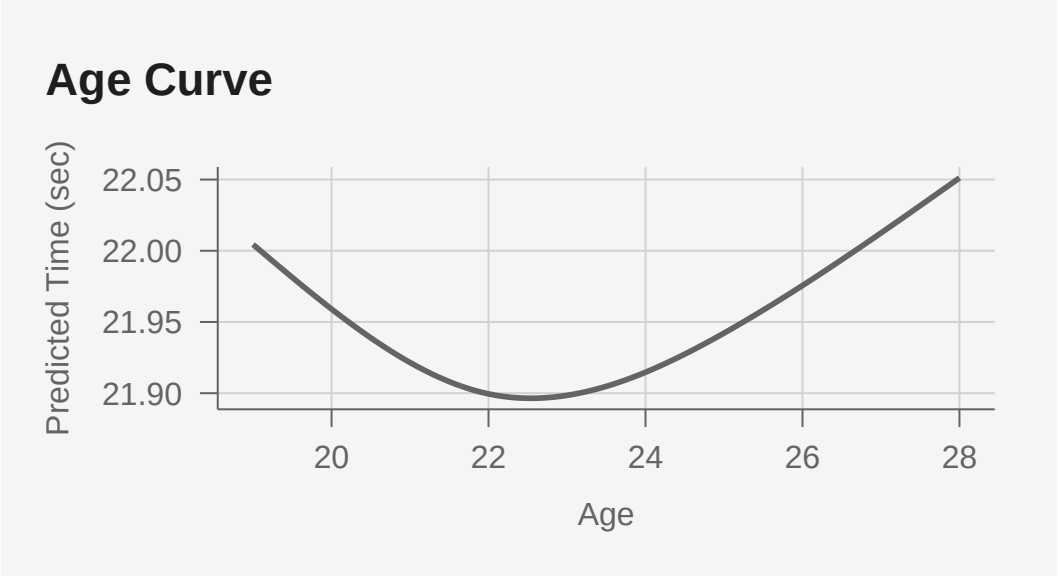
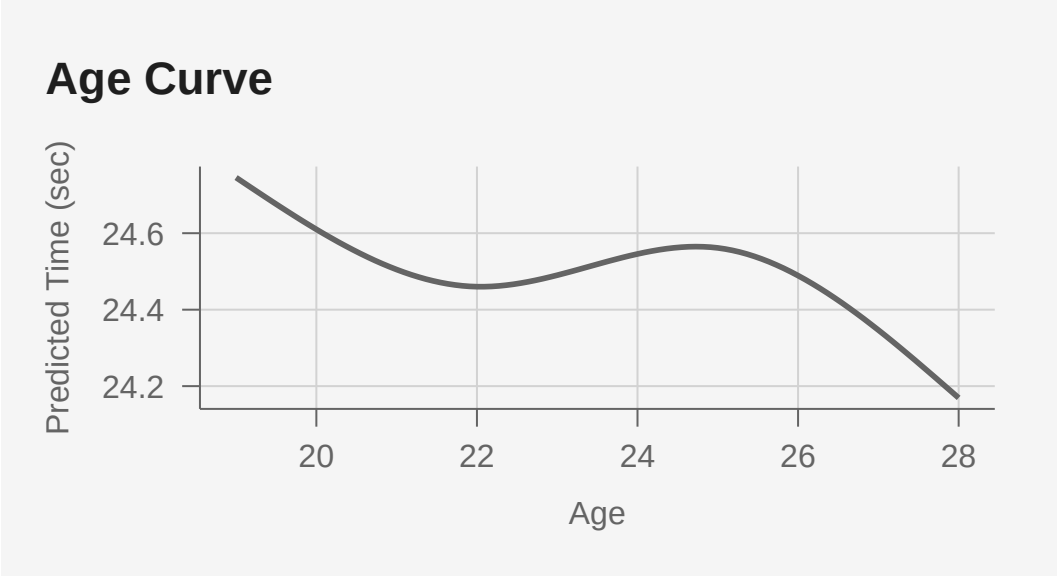


Table 6: Overall Model Performance

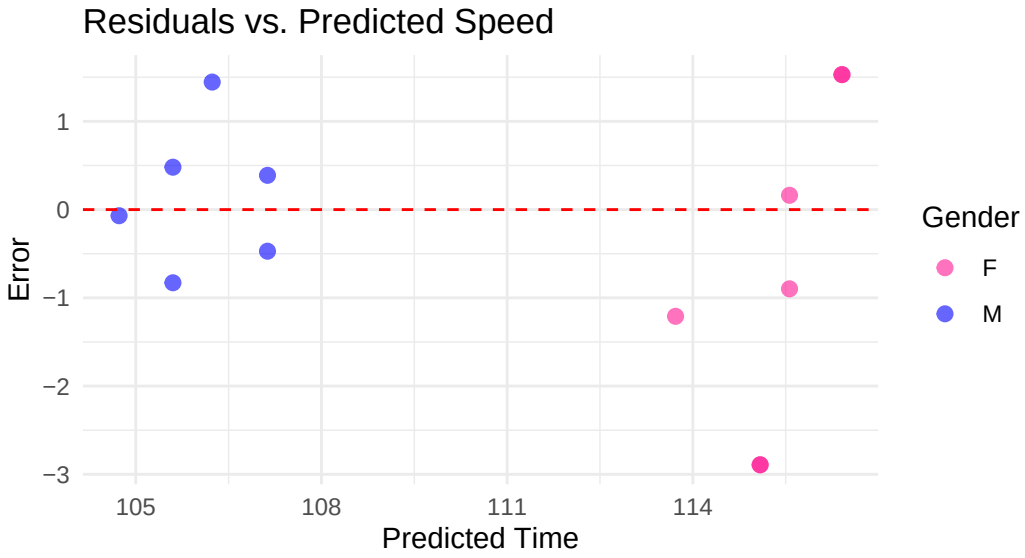
Metric	Value
MAE (Mean Absolute Error)	0.289
MSE (Mean Squared Error)	0.106
Bias (Avg Error Direction)	0.140
RMSE (Root Mean Squared Error)	0.325
Pearson Correlation	0.987

Men age curve still on top


Age Curve
Predicted Time (sec)
22.05
22.00
21.95
21.90
20
22
24
26
28
Age


Age Curve
Predicted Time (sec)
24.6
24.4
24.2
20
22
24
26
28
Age

200 free



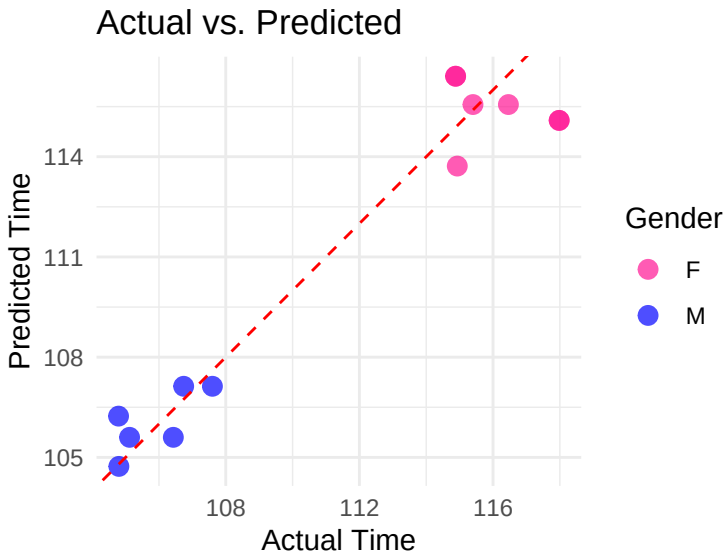


Table 7: Overall Model Performance

Metric	Value
MAE (Mean Absolute Error)	1.138
MSE (Mean Squared Error)	2.084

Metric	Value
Bias (Avg Error Direction)	-0.287
RMSE (Root Mean Squared Error)	1.444
Pearson Correlation	0.964

Men age curve still on top