Predicting swimming race time with a linear model

1 Introduction

The data consist of competitor's times in 446 swimming races in the individual finals of the 2016 Olympics and similar events at the 2016 World Championships. Four characteristics of the races are provided: the distance of the race, the sex, i.e., whether the event was women's or men's, the course, i.e., the length of the pool, and the type of stroke in the event. We investigate how race times depend on the four characteristics of the race under a normal linear model.

2 Data Exploration

The possible observations for the four characteristics are given in Table 1. Additionally, the name of the event was provided in the dataset, although this was only a combination of the distance and stroke, e.g., "200 Backstroke". In in a medley race, all four of the other strokes are swum with an equal distance for each stroke. A competitor's race time is measured in seconds up to the second decimal place.

Table 1: The possible observations for each of the four characteristics (variables) in a race.

Variable	**Possible Observations**
Distance (m)	50, 100, 200, 400
Sex	F, M
Stroke	Backstroke, Breaststroke, Butterfly, Freestyle, Medley
Course	Long (50 m), Short (25 m)

A histogram of the race times reveals four clusters of times. As shown in Figure 1, these four clusters can be entirely explained by the distance variable. Thus, it is apparent that distance explains the largest amount of the variation in race times. Furthermore, the histogram in Figure 2 reveals that within each distance, females tend to have slightly greater times than males.

There are $\binom{4}{2} = 6$ possible pairs of the variables shown in Table 1. Observations exist for each of these pairs, with a few exceptions: the 400 metre distance occurred only for races with the Freestyle or Medley strokes, and there was no 50 metre race with the Medley stroke. The number of observations of each of these two-way combinations reveals potential for bias in the course variable, in that a majority of the 50 and 100 metre distance races used the short course, whereas the 200 and 400 metre distances were evenly split between the short and long course. Thus the overall effect that the short course might have on reducing race times would appear to be greater than it actually is from overall statistics. Likewise, certain strokes were associated with longer or shorter distance races. There is no difference in the events that males and females competed in, except

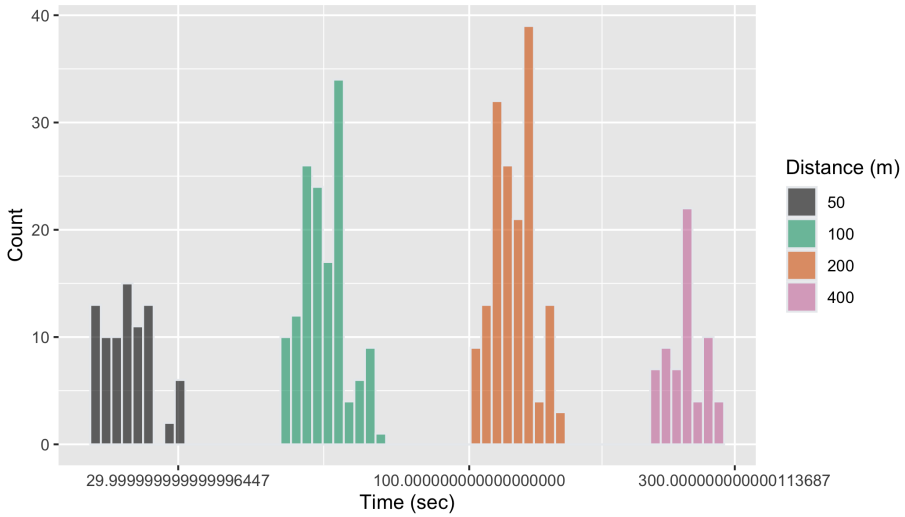

Figure 1: Histogram of times on a log scale, grouped by distance.

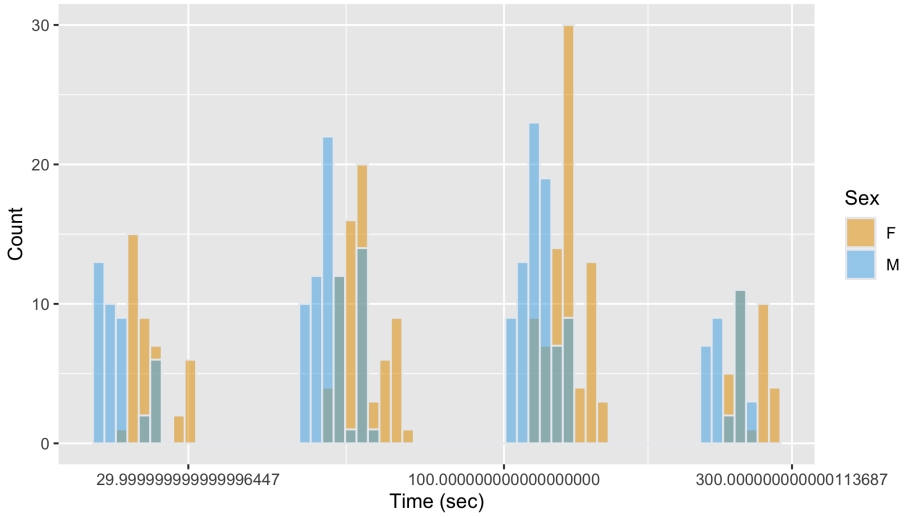

Figure 2: Histogram of times on a log scale, grouped by Sex.

that because there were two more males in the dataset than females, two of the possible three-way intersections of the other four characteristics involved eight males and only seven females.

We would like to observe the overall association that each of the four variables has with time. Thus, boxplots are shown in Figure 3 of the the four variables against time. As was discussed a moment ago, the overall effects that the strokes and the type of course might have may be muddled or exaggerated in these boxplots because of the unequal number of observations these variables have with each possible distance, which has the strongest association with race time. Nonetheless, the overall association of sex and time can reliably be inferred from the boxplots in Figure 3, because of the lack of bias between males and females, described earlier. The median race time is 89.9 seconds for females versus 80.4 seconds for males.

To obtain a clearer idea of the association that stroke, course, and sex have with time, we will consider only the races with the same distance. We will use the distance of 200 metres because it has the greatest number of observations (160), and contains the same number of observations

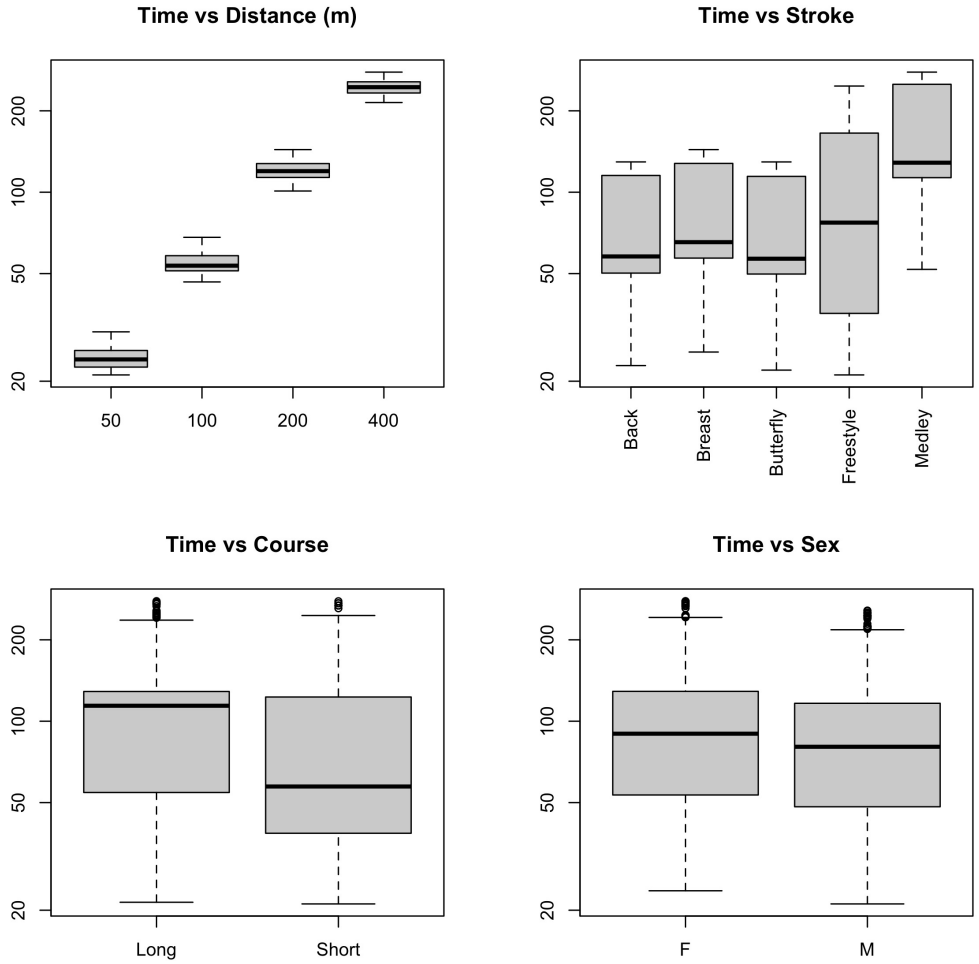

Figure 3: Boxplots of the four characteristics of the swimming race against a log scale of competitors' times.

for each possible sex (80), course (80), and stroke (32). In these refined comparisons, shown in Figure 4, it is clear that Breaststroke is associated with a long race time and Freestyle is associated with a low race time. The distribution of race times for back, breast, and freestyle overlap. The interquartile ranges of the stroke distributions range from 9.9 seconds for freestyle to 14.1 seconds for breaststroke. It is also clear that males have a lower race time than females for the 200 metre distance, with the third quartile of the distribution of the male race times below the first quartile of the distribution of the female race times. The interquartile range is slightly greater for males (8.0 seconds) than for females (7.0 seconds). Furthermore, for the 200 metre distance, there is a marginal benefit to race time by swimming in the short (25 metre) pool rather than the long (50 metre) pool. The difference between the medians of these two distributions is 3.4 seconds, and the interquartile range is slightly greater for the distribution of 200 metre races in the short pool (13.5 seconds versus 12.8 seconds).

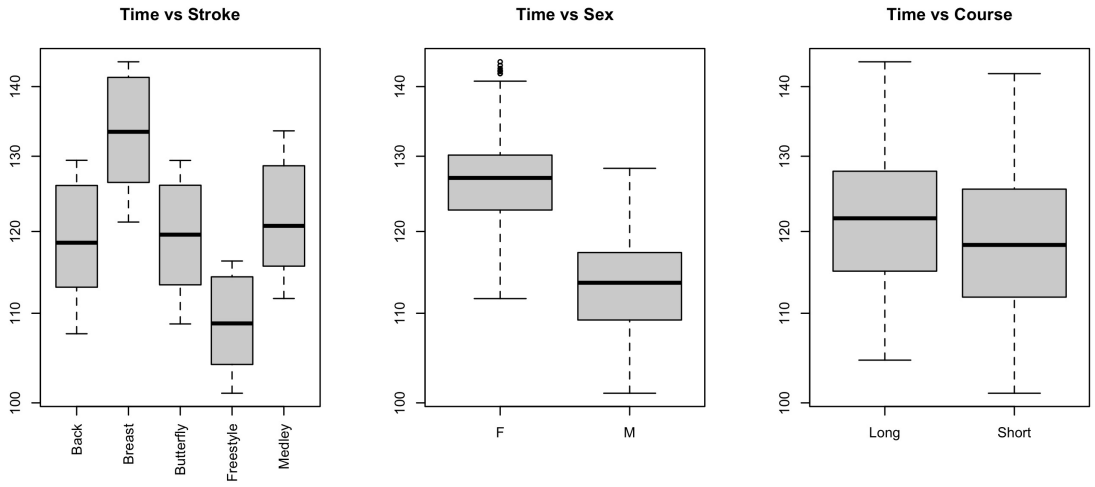

Figure 4: Boxplots of three characteristics against a log scale of competitors' times for 200 metre distance races.

The event in which a swimmer reaches the wall and turns around, occurs roughly twice as often in 25 metre pools than in 50 metre pools: 7 versus 3 times, respectively, in 200 metre distance races. The observed benefit in Figure 4 to race time of a smaller pool is interesting because it implies that the additional time required to reverse directions is more than offset by the opportunity to push off the wall, which provides a boost to a swimmer's momentum.

3 Modeling

3.1 Model selection

We would like to create a linear model for swim time based on the four variables. Our linear model will predict the logarithm of time for a few reasons: we might expect differences between the course, sex, or stroke to have relative, rather than absolute, associations with time. For instance, male races might be associated with a 5% reduction in time, rather than a constant 8-second reduction in time, regardless of the other factors of the race. Furthermore, the association between time and distance might not be exactly linear but might be such that time is proportional to some unknown power of distance: $t \propto d^p$, in which case the log of both sides must be taken to fit a linear model. Lastly, we might expect the random errors to also be relative rather than absolute. Using the logarithm of time is thus supported by common sense. It also turns out that using the logarithm of time is supported by a greater observed goodness of fit.

While we might initially fit a linear model with time proportional to some power of distance, a lower residual standard error is obtained when distance is treated as a factor (0.015 versus 0.017), with all four variables, no interactions, and all other variables treated as factors. Thus, our initial model has the following form:

$$y = \log(time)$$

$$y = \alpha + \sum_{i=2}^{4} \delta_i d_i + \beta m + \sum_{j=2}^{5} \theta_j s_j + \gamma r + \epsilon$$

The factors are ordered alphabetically, as they appear in Table 1, such that the variables are represented by m, r, d_i, and s_j, which are indicators of a male swimmer, a short course, distance i,

and stroke j, respectively. Additionally, ϵ is the random normal error, such that each observation $k = 1, ..., n$ has random error $\epsilon_1, \epsilon_2, ..., \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. We will estimate σ with the residual standard error s.

In this initial model, each coefficient is highly significantly different than zero ($p < 0.001$), and by ANOVA, each variable is also significantly predictive ($p < 0.001$). We might then suppose there are informative interactions between some of the variables that can obtain even greater predictions. Indeed, including all possible two-way interactions obtains a lower s (0.015 to 0.011) and greater R^2 (0.99960 to 0.99976) than having no interactions. However, some of the coefficients are not significantly different than zero. Most prominently, all of the coefficients for the interaction between distance and course are not significantly different than zero, and this interaction is the least predictive by ANOVA, even though it is significant ($p < 0.01$). This significance depends on the order of the variables in the ANOVA, here the variables were ordered as shown in the above equation and in Table 1. We thus exclude the interaction between course and distance such that the following two-way interaction terms are added to the initial model:

$$\phi mr + m\sum_{i=2}^{4}\rho_i d_i + m\sum_{j=2}^{5}\psi_j s_j + r\sum_{j=2}^{5}\omega_j s_j + \sum_{i=2}^{4}\sum_{j=2}^{5}\tau_{ij} d_i s_j$$

Four of the τ_{ij} coefficients cannot be estimated because of the lack of observations for certain intersections of distance and course, described earlier. It may be worth remarking that for the model with all three-way interactions, none of the coefficients are significantly different than zero, and none of the three-way interactions are predictive by ANOVA.

3.2 Model assessment and outlier analysis

The linear model assumed that the residuals were normally distributed and unassociated with the fitted values. These assumptions are confirmed in Figure 5. The Q-Q plot of the studentised residuals shows approximate normality for large n, and a plot of the studentised residuals against the fitted values shows no sign of correlation or non-constant variance. The Q-Q has borderline issues in the lower quantiles. It appears that the top race times were better than would be expected under a normal distribution of errors. Perhaps, the top swimmers 'break' the normal curve through exceptional performance.

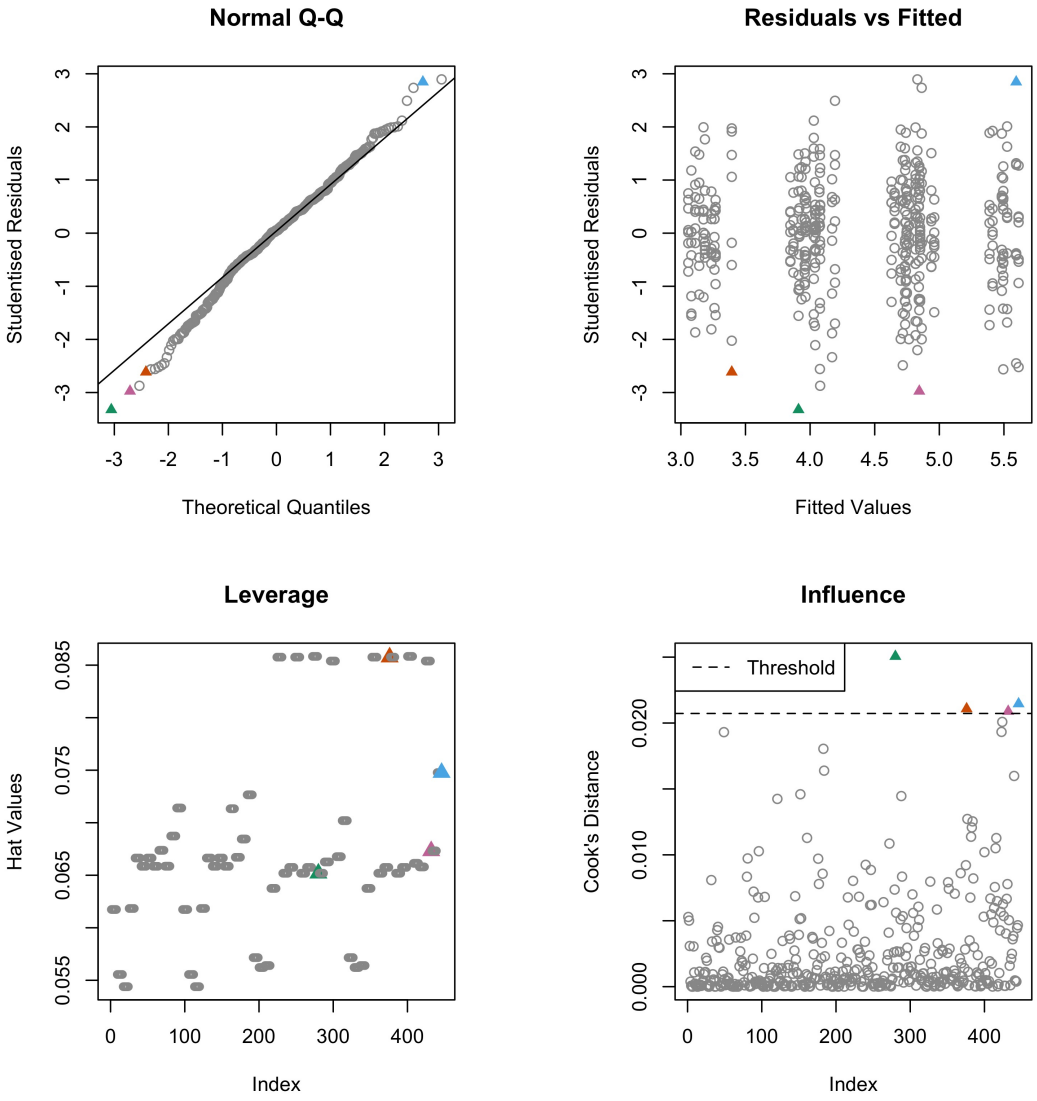

Figure 5: Model assessment and outlier analysis of the linear model. Coloured points have a Cook's distance greater than the threshold $\frac{8}{n-2p}$.

The outliers can be identified with Cook's distance, which measures the influence data points have on the linear fit. Points with influence greater than the threshold $\frac{8}{n-2p}$ can be considered to have 'high' influence, where in this case $n = 446$ and $p = 30$. It is worth remarking that none of the data points had 'high' leverage, which can be considered as greater than twice the mean leverage $\frac{2p}{n}$. Instead, the four data-points observed to have high influence were highly misfit, meaning their standardised residuals were large. These four outliers are described in Table 2 and coloured in Figure 5. We choose not to remove all of these values because we have no reasons to doubt their validity and they are only marginally influential. There are at least three other data points that are similarly influential, but happen to have a Cook's distance just below the threshold for high influence. Nonetheless, we will remove the green data point, because it has the largest misfit and is far more influential than the other outliers.

Table 2: Outliers with high influence on the linear fit.

Time	Distance	Sex	Stroke	Course	Misfit	Leverage	Influence	Colour
48.08	100	M	Butterfly	Short	-3.28	0.0652	0.0251	Green
28.92	50	F	Breaststroke	Short	-2.60	0.0858	0.0211	Orange
122.90	200	F	Medley	Short	-2.95	0.0673	0.0209	Pink
277.79	400	F	Medley	Short	2.82	0.0748	0.0215	Blue

Though not shown here, the Q-Q and residual versus fitted plots are no worse upon removing the most influential outlier from the model fit, four slight outliers remain. In this final model, each term is highly predictive by ANOVA ($p < 0.001$), and only a few of the coefficients are not significant. The significance-level of each coefficient is shown in Table 3. The final model has an adjusted $R^2 = 0.999743$.

3.3 Model interpretation

We would like to be able to interpret the linear model. Firstly, the final model had a residual standard error of 0.0118. Because e^ϵ gave the percent error, we raise e to the residual standard error to obtain the residual standard error in percentage form: 1.19%. This tells us that the model's predictions of race time tended to be off by about one percent.

In order to interpret how the model makes predictions, we turn to the coefficients, which are shown in Table 3. The exponent of the intercept e^α is the predicted time for a female, swimming backstroke in a 50 metre race in a long (50 metre) pool: 27.2 seconds. All of the single-category coefficients, are given precisely relative to this baseline. Each coefficient corresponds with a relative marginal difference, and the marginal differences are multiplied to obtain the final prediction.

For example, to obtain the predicted time for a female swimming a 100 metre butterfly race, one would multiply the baseline 27.2 seconds by the marginal effect of a 100 metre distance (2.16) and the butterfly stroke (0.9453). Furthermore, there is an interaction between the 100 metre distance and the butterfly stroke, such that the predicted time is different than might be expected from knowing just that the race is 100 metres or that the stroke is butterfly, alone. Lastly, multiplying by this interaction (1.0245), obtains the predicted time.

Table 3: Marginal difference of each term, given relative to the time of 27.2 seconds (from the intercept). Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

Category	**Observation**	**Marginal Difference (%)**	**Significance**
Distance	100	116	***
	200	370	***
	400	891	***
Sex	M	-10.8	***
Stroke	Breaststroke	12.0	***
	Butterfly	-5.47	***
	Freestyle	-11.1	***
	Medley	1.28	***
Course	Short	-3.43	***
Distance & Sex	100 & M	0.656	.
	200 & M	1.21	***
	400 & M	2.25	***
Distance & Stroke	100 & Breaststroke	0.786	
	200 & Breaststroke	-0.303	
	400 & Breaststroke	–	–
	100 & Butterfly	2.45	***
	200 & Butterfly	4.80	***
	400 & Butterfly	–	–
	100 & Freestyle	1.49	**
	200 & Freestyle	1.34	**
	400 & Freestyle	1.16	*
	100 & Medley	0.994	*
	200 & Medley	–	–
	400 & Medley	–	–
Sex & Stroke	M & Breaststroke	-0.723	.
	M & Butterfly	0.166	
	M & Freestyle	0.911	*
	M & Medley	0.322	
Sex & Course	M & Short	-1.23	***
Stroke & Course	Short & Breaststroke	1.25	**
	Short & Butterfly	2.40	***
	Short & Freestyle	2.41	***
	Short & Medley	1.65	***

We can see the model's predictions against the actual data in Figure 6, which shows the observed and predicted times for a Backstroke race, for 100 and 200 metre distances, and the different possible sex and course values. There is little variance about the predictions, such that the unexplained, within group variance is clearly much smaller than the explained, between-group variance.

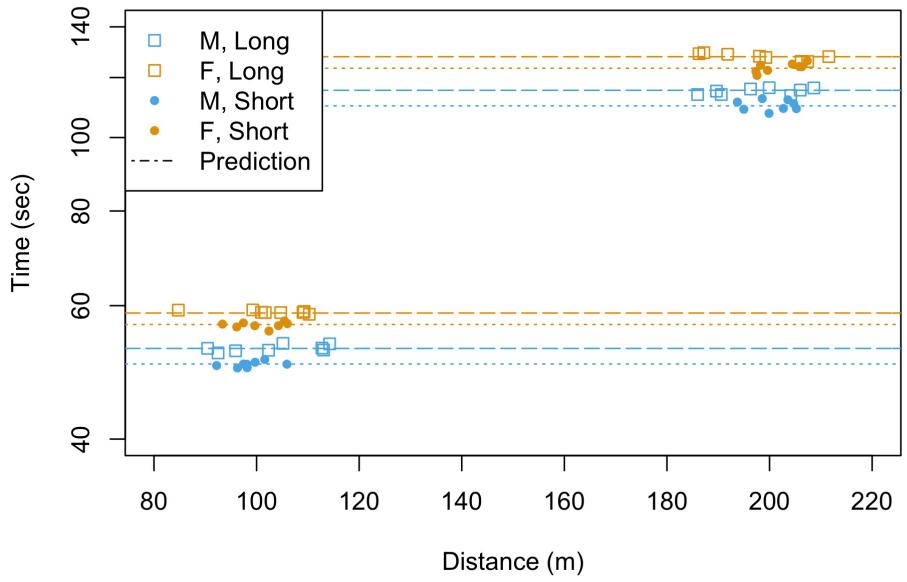

Figure 6: Observed and predicted times for the backstroke race. The race distances are either 100 or 200 metres, though are shown with random 'jitter' to enable vertical comparisons of the data.

3.4 Model Predictions

Let us suppose we are given information about four additional races. We shall use the model to make predictions about these new race times. The predictions are shown in Table 4 along with the lower and upper bounds of the 95% prediction intervals. In general, the prediction intervals are plus or minus about 2.5% relative to the predicted time. Though the 400 metre butterfly prediction has the largest confidence interval as a percentage, likely because there was no observation in the data of a 400 metre butterfly race.

Table 4: Predictions and 95% prediction intervals for the times of four additional races.

Distance	Sex	Stroke	Course	Prediction	Lower	Upper	+/- %
400	F	Freestyle	Long	243.5	237.7	249.4	2.42
50	F	Backstroke	Long	27.2	26.5	27.9	2.53
400	F	Butterfly	Long	255.8	249	262.8	2.73
100	F	Medley	Long	60	58.5	61.5	2.52

We might wish to remark upon a comparison of one of the predictions to actual data. The first set of characteristics (400 metre, female, freestyle, long), has a mean time of 243 seconds and with eight observations, a minimum and maximum time of 236 and 247 seconds, respectively.

4 Conclusions

A linear model was fit between race times and four categories of the race: distance, sex, stroke, and course (long or short). The model predicted the logarithm of time and was improved by treating distance as a factor, along with the other three variables as factors. Furthermore, the model was improved by including all of the interactions between the four factors, which were each highly predictive, except the interaction between distance and course was excluded because it was not as predictive. The model was shown to have normally distributed residuals that were not associated with the fitted values. The model was refit after the removal of one extreme outlier. This final version of the model was interpreted and used to predict the race times of four additional races. The model had a very high goodness of fit obtained random normal errors around one percent. It would be interesting to use this model predict swim times in the 2021 Tokyo Olympics. Systemic error might arise due to a general increase or decrease in performance among finalist swimmers between 2016 and 2021.

Appendix

```
## Load data

sm <- read.csv("data/swim.csv", stringsAsFactors = TRUE)

str(sm)
head(sm)
levels(sm$stroke)


## Feature engineering

sm$log_dist <- log(sm$dist)

sm$dist_factor <- factor(sm$dist)  # distances as factor

sm$log_time <- log(sm$time)  # log transformation of time


## Data exploration

#### Histograms

library(ggplot2)
library(tidyverse)

hist_dist <- sm %>%
  ggplot( aes(x=time, fill=dist_factor)) +
  geom_histogram(color="#e9ecef", alpha=0.6, position = 'identity',
  bins=60) + scale_fill_manual(
  values=c("#000000", "#009E73", "#D55E00", "#CC79A7")) +
  labs(fill="") + xlab("Time (sec)") + ylab("Count") +
 scale_x_continuous(trans = 'log10') + guides(fill=guide_legend(title="Distance (m)"))

hist_dist
ggsave("plots/hist_dist.png", height = 4 , width = 7,
       plot = hist_dist, dpi = 300)

hist_sex <- sm %>%
  ggplot( aes(x=time, fill=sex)) +
  geom_histogram(color="#e9ecef", alpha=0.55, position = 'identity', bins=60) +
  scale_fill_manual(values=c("#E69F00", "#56B4E9")) +
  labs(c("Female", "Male")) + xlab("Time (sec)") + ylab("Count") +
  scale_x_continuous(trans = 'log10') + guides(fill=guide_legend(title="Sex"))

hist_sex
ggsave("plots/hist_sex.png", height = 4 , width = 7,
       plot = hist_sex, dpi = 300)
```

```
hist_course <- sm %>%
  ggplot( aes(x=time, fill=course)) +
  geom_histogram(color="#ffffff", alpha=0.55, position = 'identity', bins=60) +
  scale_fill_manual(values=c("#E69F00", "#56B4E9")) +
  labs(c("Long", "Short")) + xlab("Time (sec)") + ylab("Count") +
 scale_x_continuous(trans = 'log10') + guides(fill=guide_legend(title="Course"))

hist_course


#### Relationship between Sex and Race time

sm %>%
  group_by(sex) %>%
  summarise(
    median = median(time),
    iqr = IQR(time))


#### Summary statistics for 200 m distance

subset(sm, dist_factor == "200") %>%
  group_by(stroke) %>%
  summarise(
    median = median(time),
    q1 = quantile(time, 0.25),
    q3 = quantile(time, 0.75),
    iqr = IQR(time)
  ) %>%
  arrange(desc(median))

subset(sm, dist_factor == "200") %>%
  group_by(sex) %>%
  summarise(
    median = median(time),
    q1 = quantile(time, 0.25),
    q3 = quantile(time, 0.75),
    iqr = IQR(time)
  ) %>%
  arrange(desc(median))

subset(sm, dist_factor == "200") %>%
  group_by(course) %>%
  summarise(
    median = median(time),
    q1 = quantile(time, 0.25),
    q3 = quantile(time, 0.75),
    iqr = IQR(time)
```

```
  ) %>%
  arrange(desc(median))


#### Boxplots of four variables

jpeg("plots/overall_boxplots.jpg", width = 8, height = 8, units = "in", res = 300)
par(mfrow = c(2, 2))
plot(time ~ dist_factor, data = sm, log='y', main = "Time vs Distance (m)", ylab = NULL,
     xlab = NULL)
plot(time ~ stroke, data = sm, log='y', main = "Time vs Stroke", ylab = NULL,
     xlab = NULL, xaxt = "n")

axis(1, at = c(1:5),
     labels = c("Back", "Breast", "Butterfly", "Freestyle", "Medley"),
     las=2)

plot(time ~ course, data = sm, log='y', main = "Time vs Course", ylab = NULL,
     xlab = NULL)
plot(time ~ sex, data = sm, log='y', main = "Time vs Sex", ylab = NULL,
     xlab = NULL)
dev.off()


#### Boxplots of three variables for the 200 m race

jpeg("plots/200m_boxplots.jpg", width = 9, height = 4,
     units = "in", res = 300)
par(mfrow = c(1, 3))

plot(time ~ stroke, data = sm[(sm$dist == 200), ], log='y', ylab = NULL,
     main = "Time vs Stroke", xlab = NULL, xaxt = "n")
axis(1, at = c(1:5),
     labels = c("Back", "Breast", "Butterfly", "Freestyle", "Medley"),
     las=2)

plot(time ~ sex, data = sm[(sm$dist == 200), ], log='y', ylab = NULL,
     main = "Time vs Sex", xlab = NULL)

plot(time ~ course, data = sm[(sm$dist == 200), ], log='y', ylab = NULL,
     main = "Time vs Course", xlab = NULL)
dev.off()


#### Tables of the counts for each two-way interaction

# There are (4 choose 2 = 6) two-way interactions of the four variables
table(sm$sex, sm$stroke)
```

```
table(sm$sex, sm$dist)

table(sm$dist, sm$stroke)

table(sm$sex, sm$course)

table(sm$stroke, sm$course)

table(sm$course, sm$dist)

# data.frame(table(sm$course, sm$dist, sm$sex, sm$stroke))  # four-way

table(sm$dist_factor)  # 200 m has the most observations


#### Get information about the four-way interactions

# Sex is unbiased
which.max(table(sm$sex, sm$stroke, sm$dist_factor, sm$course))
cs <- data.frame(table(sm$sex, sm$stroke, sm$dist_factor, smr$course))
print(unique(cs$Freq))
cs <- cs[cs$Freq != 0, ]
cs

print("Possible:")
print(2 * 5 * 4 * 2)  # possible
print("Actual:")
print(dim(cs)[1])  # actual


## Modeling

# no interactions
sm.lm0.t <- lm(time ~ dist + sex + stroke + course, data = sm)

sm.lm0.d <- lm(log_time ~ dist + sex + stroke + course, data = sm)
sm.lm0.l <- lm(log_time ~ log_dist + sex + stroke + course, data = sm)
sm.lm0 <- lm(log_time ~ dist_factor + sex + stroke + course, data = sm)

# two-way interactions
sm.lm1 <- lm(log_time ~ (dist_factor + sex + stroke + course)^2, data = sm)
sm.lm1.l <- lm(log_time ~ (log_dist + sex + stroke + course)^2, data = sm)

# three-way interactions
sm.lm2 <- lm(log_time ~ (dist_factor + sex + stroke + course)^3, data = sm)


#### Model selection
```

```
# comparing log time
summary(sm.lm0.t)$r.squared
summary(sm.lm0.l)$r.squared
# greater R squared

# comparing distance as factor with no interactions
summary(sm.lm0.l)$r.squared
summary(sm.lm0)$r.squared
# greater R squared, lower sigma

# comparing two way interactions
summary(sm.lm0)$sigma
summary(sm.lm1)$sigma
summary(sm.lm0)$r.squared
summary(sm.lm1)$r.squared
# lower sigma and greater R squared.


#### ANOVA and t-tests

# Performance of no interactions
summary(sm.lm0)
anova(sm.lm0)

# Performance of all two-way interactions
options(digits = 3)
summary(sm.lm1)
anova(sm.lm1)

# Performance of all two-way interactions
options(digits = 2)
summary(sm.lm2)
anova(sm.lm2)

# final, simpler model
sm.lm <- lm(log_time ~ (dist_factor + sex + stroke + course)^2
            - dist_factor:course, data = sm)

# Assess its performance
options(digits=3)
summary(sm.lm)
anova(sm.lm)


# Outlier analysis

n <- dim(sm)[1]
p <- sm.lm$rank
```

```
#### Influence

inf <- cooks.distance(sm.lm) > (8/(n - 2*p))
(sm_inf <- sm[inf, c("time", "sex", "dist_factor", "stroke", "course")])

#
i <- as.integer(inf)
count <- 0
for (j in 1:length(i)) {
  if (i[j] > 0) {
    i[j] <- i[j] + count
    count <- count + 1
  }
}

#### Leverage

lev <- hatvalues(sm.lm) > (2*p/ n)

# no points are considered to have high leverage
sum(lev) == 0

# quantiles for ratio to mean leverage
quantile(hatvalues(sm.lm) / (p / n), c(0.01, 0.5, 0.25, 0.5, 0.75, 0.95, 0.99))


#### Misfit (standardized residual)

mis <- abs(rstandard(sm.lm)) > 2
sm[mis, ]

# 19 misfit values
sum(mis)

sm_inf$mis <- rstandard(sm.lm)[inf]
sm_inf$lev <- hatvalues(sm.lm)[inf]
sm_inf$inf <- cooks.distance(sm.lm)[inf]

sm_inf


#### Plot outliers

pch_value <- 16

jpeg("plots/model_analysis.jpg", width = 8, height = 8, units = "in", res = 400)

colors_outlier <- c("#999999", "#009E73", "#D55E00", "#CC79A7",
"#56B4E9", "#F0E442")
```

```
par(mfrow = c(2, 2), pty="s")

qqnorm(rstudent(sm.lm), main = "Normal Q-Q",
pch = 1 + pch_value*inf, col = colors_outlier[i+1], ylab = "Studentised Residuals")
qqline(rstudent(sm.lm))

plot(fitted.values(sm.lm), main = "Residuals vs Fitted",
     rstudent(sm.lm), pch = 1 + pch_value*inf,
     col = colors_outlier[i+1], xlab = "Fitted Values",
     ylab = "Studentised Residuals")

plot(hatvalues(sm.lm), main = "Leverage",
     pch = 1 + pch_value*inf, col = colors_outlier[i+1], cex=0.8 + 0.7*(i > 0),
     ylab = "Hat Values")
abline(2*p/n, 0, lty = 2)

plot(cooks.distance(sm.lm), main = "Influence",
     pch = 1 + pch_value*inf, col = colors_outlier[i+1],
     ylab = "Cook's Distance")
abline(8/(n - 2*p), 0, lty = 2)
legend("topleft", c("Threshold"), lwd=1, lty = 2)
dev.off()


#### Final model (remove outlier)

smr <- sm[-280]

nr <- dim(smr)[1]

smr.lm <- lm(log_time ~ (dist_factor + sex + stroke + course)^2
               - dist_factor:course, data = smr)

options(digits = 20)
summary(smr.lm)$r.squared
anova(smr.lm)
summary(smr.lm)$sigma
summary(sm.lm)$sigma
summary(sm.lm)$r.squared

options(digits=3)
summary(smr.lm)


#### Check for outliers in the final model

any(i <- cooks.distance(smr.lm) > (8/(nr - 2*p)))
inf <- i
```

```
par(mfrow = c(2, 2))

qqnorm(rstudent(smr.lm), main = NULL, pch = 1 + 15*inf,
col = colors_outlier[i+1])
qqline(rstudent(smr.lm))

plot(fitted.values(smr.lm), rstudent(smr.lm), pch = 1 + 15*inf,
col = colors_outlier[i+1])

plot(hatvalues(smr.lm), pch = 1 + 15*inf, col = colors_outlier[i+1],
cex=0.8 + 0.7*(i > 0))
abline(2*p/n, 0, lty = 2)

help(plot)

plot(cooks.distance(smr.lm), pch = 1 + 15*inf, col = colors_outlier[i+1])
abline(8/(n - 2*p), 0, lty = 2)


## Model interpretation

levels(smr$sex)
levels(smr$stroke)
levels(smr$dist_factor)
levels(smr$course)


#### Prepare interpretive plot

c <- coef(smr.lm)

# already back, female, long_course
# only change is distance factor (100).
(fl_back_100 <- exp(c["(Intercept)"] + c["dist_factor100"]))
(fl_back_200 <- exp(c["(Intercept)"] + c["dist_factor200"]))

(fs_back_100 <- exp(c["(Intercept)"] + c["dist_factor100"] + c["courseShort"]))
(fs_back_200 <- exp(c["(Intercept)"] + c["dist_factor200"] + c["courseShort"]))

(ml_back_100 <- exp(c["(Intercept)"] + c["dist_factor100"]
                    + c["sexM"] + c["dist_factor100:sexM"]))
(ml_back_200 <- exp(c["(Intercept)"] + c["dist_factor200"]
                    + c["sexM"] + c["dist_factor200:sexM"]))

(ms_back_100 <- exp(c["(Intercept)"] + c["dist_factor100"] + c["courseShort"]
               + c["sexM"] + c["dist_factor100:sexM"] + c["sexM:courseShort"]))

(ms_back_200 <- exp(c["(Intercept)"] + c["dist_factor200"] + c["courseShort"]
```

```
               + c["sexM"] + c["dist_factor200:sexM"] + c["sexM:courseShort"]))

stroke_plot <- "Backstroke"
ml <- subset(smr, stroke == stroke_plot & course == "Long" & sex == "M")
fl <- subset(smr, stroke == stroke_plot & course == "Long" & sex == "F")
ms <- subset(smr, stroke == stroke_plot & course == "Short" & sex == "M")
fs <- subset(smr, stroke == stroke_plot & course == "Short" & sex == "F")


#### Make plot
jpeg("plots/interpret.jpg", width = 6.4, height = 4.8, units = "in", res = 300)

pchs <- c(0, 20)
colors <- c("#56B4E9", "#E69F00", "#009E73", "#F0E442", "#0072B2",
"#D55E00", "#CC79A7")

colors2 <- c("#ff7f0e", "#1f77b4", "#2ca02c", "#9467bd", "#8c564b")
set.seed(1)
jitter_val <- 0.8
plot(ml$time ~ jitter(ml$dist, jitter_val), pch = pchs[1], log='y', col = colors[1],
    ylim=c(40, 140), xlim=c(80, 220), xlab="Distance (m)", ylab = "Time (sec)")
abline(ml_back_200, 0, lty = 5, untf = TRUE, col = colors[1])
abline(ml_back_100, 0, lty = 5, untf = TRUE, col = colors[1])

points(ms$time ~ jitter(ms$dist, jitter_val), pch = pchs[2], col = colors[1])
abline(ms_back_200, 0, lty = 3, untf = TRUE, col = colors[1])
abline(ms_back_100, 0, lty = 3, untf = TRUE, col = colors[1])

points(fl$time ~ jitter(fl$dist, jitter_val), pch = pchs[1], col = colors[2])
abline(fl_back_200, 0, lty = 5, untf = TRUE, col = colors[2])
abline(fl_back_100, 0, lty = 5, untf = TRUE, col = colors[2])


points(fs$time ~ jitter(fs$dist, jitter_val), pch = pchs[2], col = colors[2])
abline(fs_back_200, 0, lty = 3, untf = TRUE, col = colors[2])
abline(fs_back_100, 0, lty = 3, untf = TRUE, col = colors[2])


legend("topleft", c("M, Long", "F, Long",
                    "M, Short", "F, Short", "Prediction"),
       pch=c(rep(pchs, each=2), NA), lty = c(NA, NA, NA, NA, 4),
       lwd = c(NA, NA, NA, NA, 1),
       col=c(rep(colors[1:2], 2), 1))
dev.off()


#### Get coefficients

c <- coef(smr.lm)
```

```
cinterp <- exp(c)

df <- data.frame(a = c(cinterp[1], 100*(cinterp[-1] - 1)))

df$a <- round(df$a, 3)


#### Get standard error

options(digits=10)
summary(smr.lm)$sigma
exp(summary(smr.lm)$sigma)



## Predictions

new <- data.frame(dist = c(400, 50, 400, 100),
           stroke = c("Freestyle", "Backstroke", "Butterfly", "Medley"),
           sex = rep("F", 4),
           course = rep("Long", 4), stringsAsFactors = TRUE)
new$dist_factor <- factor(new$dist)

preds <- predict(smr.lm, new, interval = "prediction", level = 0.95)

preds <- data.frame(exp(preds))

library(dplyr)
np <- bind_cols(new, preds)


#### Compare the first prediction to data

pred1 <- subset(sm, (dist == 400) & (sex == "F") &
         (stroke == "Freestyle") &  (course == "Long"))

mean(pred1$time)
sd(pred1$time)
quantile(pred1$time, c(0.025, 0.975))

pred1 %>%
  arrange(time)
```