Linear Models, Marked Practical, Supplementary Code

```
swim <- read.csv("swim.csv")

library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.5
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.0     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 4.0.5
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(patchwork)
```

```
## Warning: package 'patchwork' was built under R version 4.0.5
```

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:patchwork':
##
##     area
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
saved_values_numeric <- data.frame(name = c(), value = c())
saved_values_text <- data.frame(name = c(), value = c())
```

Exploratory Analysis

First we have to transform the data into a suitable form for analysis:

```
# Factorialising categorial variables
swim$stroke <- as.factor(swim$stroke)
swim$sex <- as.factor(swim$sex)
swim$course <- as.factor(swim$course)
# For the moment we are considering distance as a continous value
swim$dist <- as.integer(swim$dist)

# Removing the event column since it is providing information already included in the
# other columns
swim <- swim %>%
  dplyr::select(-event)
```

Basic Features of the Data

The data consists of one continuous variable, time; one ordinal variable, distance; and three categorical variables, sex, course and stroke.

```
summary(swim$stroke)
```

```
##   Backstroke Breaststroke    Butterfly    Freestyle       Medley
##           80           80           79          128           79
```

```
summary(swim$sex)
```

```
##   F   M
## 222 224
```

```
summary(swim$dist)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    50.0   100.0   150.0   169.3   200.0   400.0
```

```
summary(swim$course)
```

```
##  Long Short
##   191   255
```

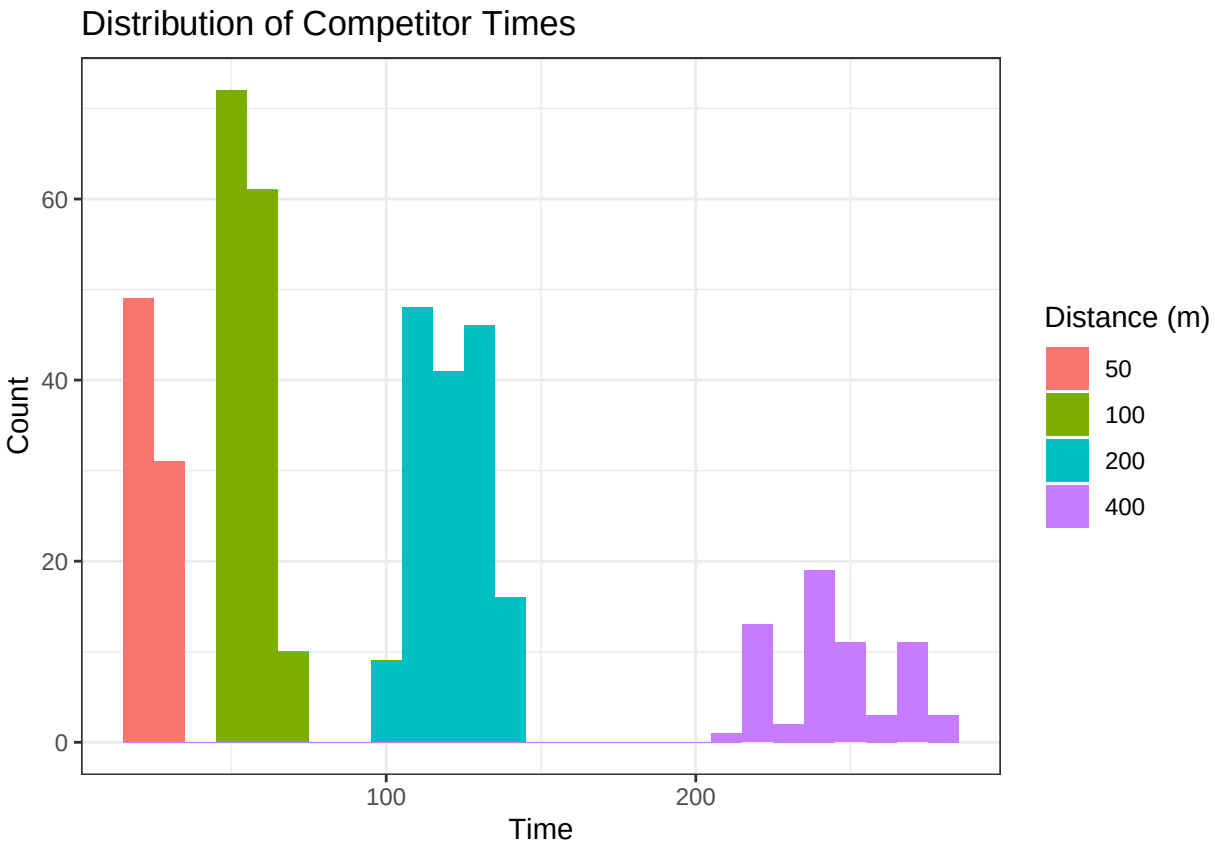
```r
# Summarising swim times and outputting results to an external file
swim_summary <- as.data.frame(as.matrix(summary(swim$time)))
swim_summary$statistic <- row.names(swim_summary)
swim_summary <- swim_summary %>%
  dplyr::select(statistic,V1)
row.names(swim_summary) <- NULL
swim_summary
```

```
##   statistic        V1
## 1      Min.  21.10000
## 2   1st Qu.  50.81500
## 3    Median  84.56500
## 4      Mean  99.94726
## 5   3rd Qu. 126.80500
## 6      Max. 278.06000
```

```r
write.csv(swim_summary,file = "swim_time_summary.csv",quote = FALSE)
```

The overall distribution of the data is visualised in the following histogram:

```r
swim %>%
  mutate(dist = as.factor(dist)) %>%
  ggplot() +
  geom_histogram(aes(x = time, fill = dist),binwidth = 10) +
  theme_bw() +
  labs(title = "Distribution of Competitor Times", x = "Time", y = "Count") +
  scale_fill_discrete(name = "Distance (m)")
```

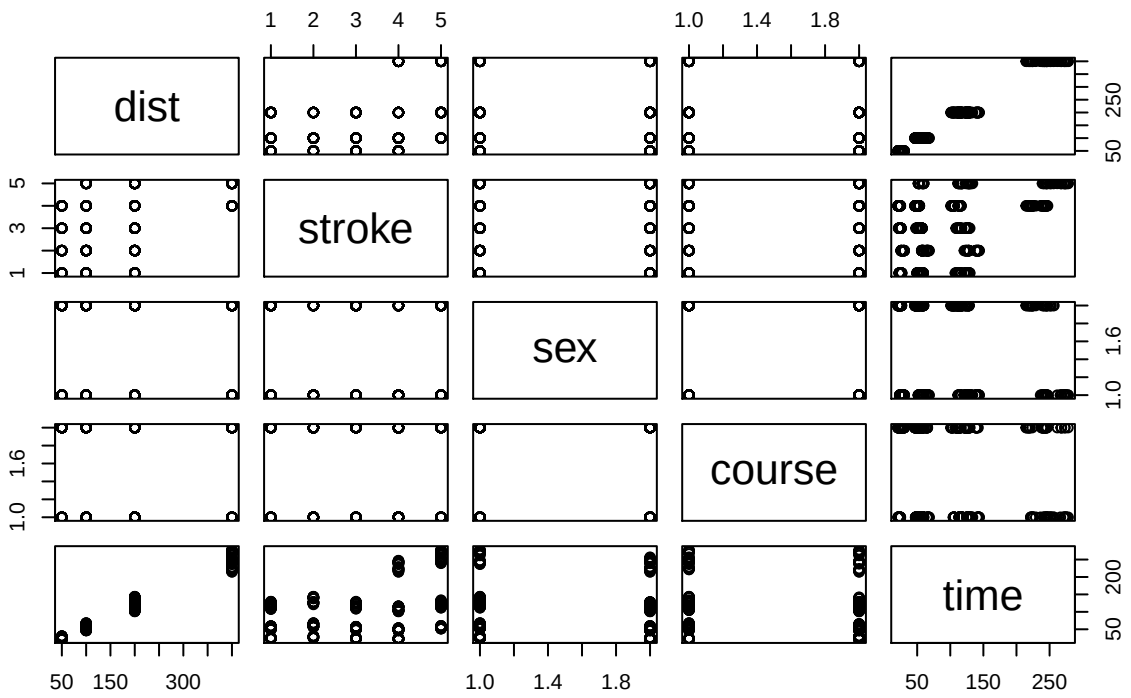


The distribution of times consists of four very distinct peaks, which appear to become wider with increasing time.

Exploratory Plots

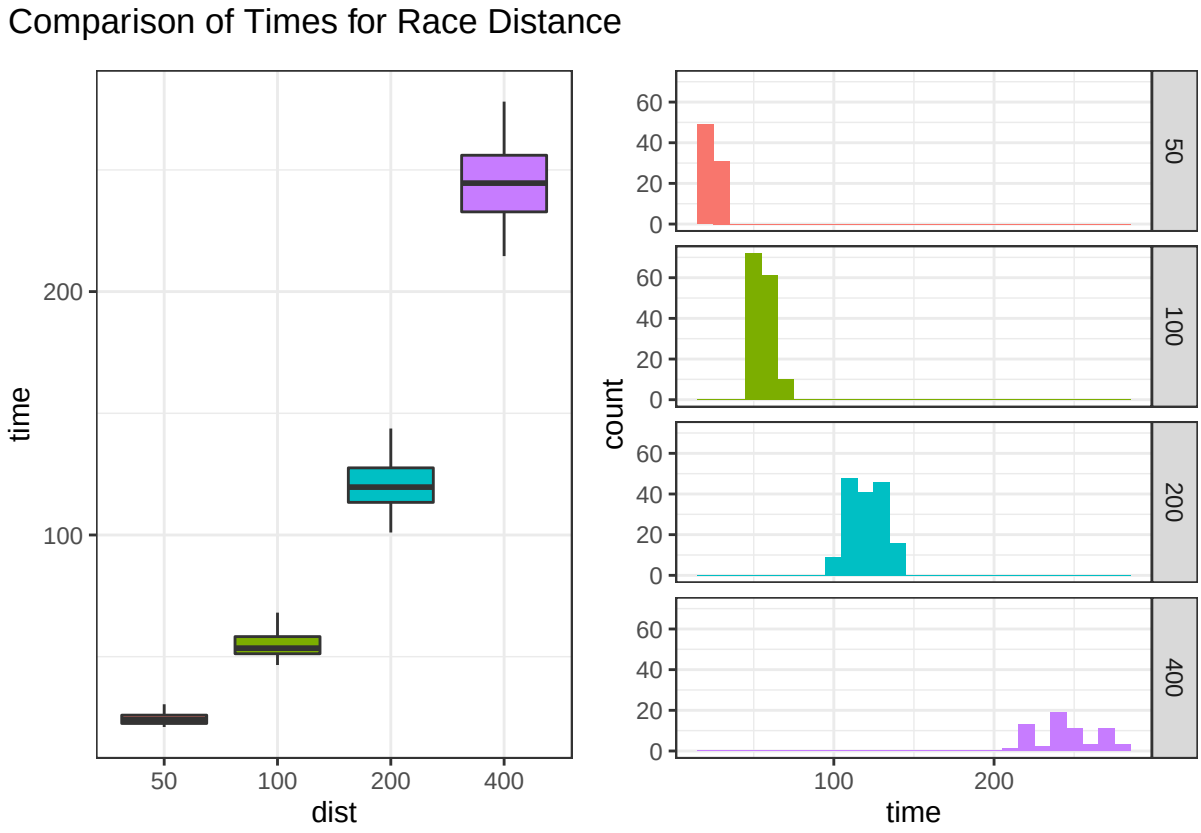
First we look at a plot of the all the variables plotted against each other in a pairs plot:

```
pairs(swim, main = "Pairs Plot of Competitor Data")
```



Each of the categorical variables, distance, stroke, sex and course seem for the most part independent of each other apart from distance and stroke where we can see that some there are there are some distances for which there is no competition for a particular stroke. There are clear relationships between distance, stroke, sex and course and time, which we shall explore in closer detail using some clearer plots.

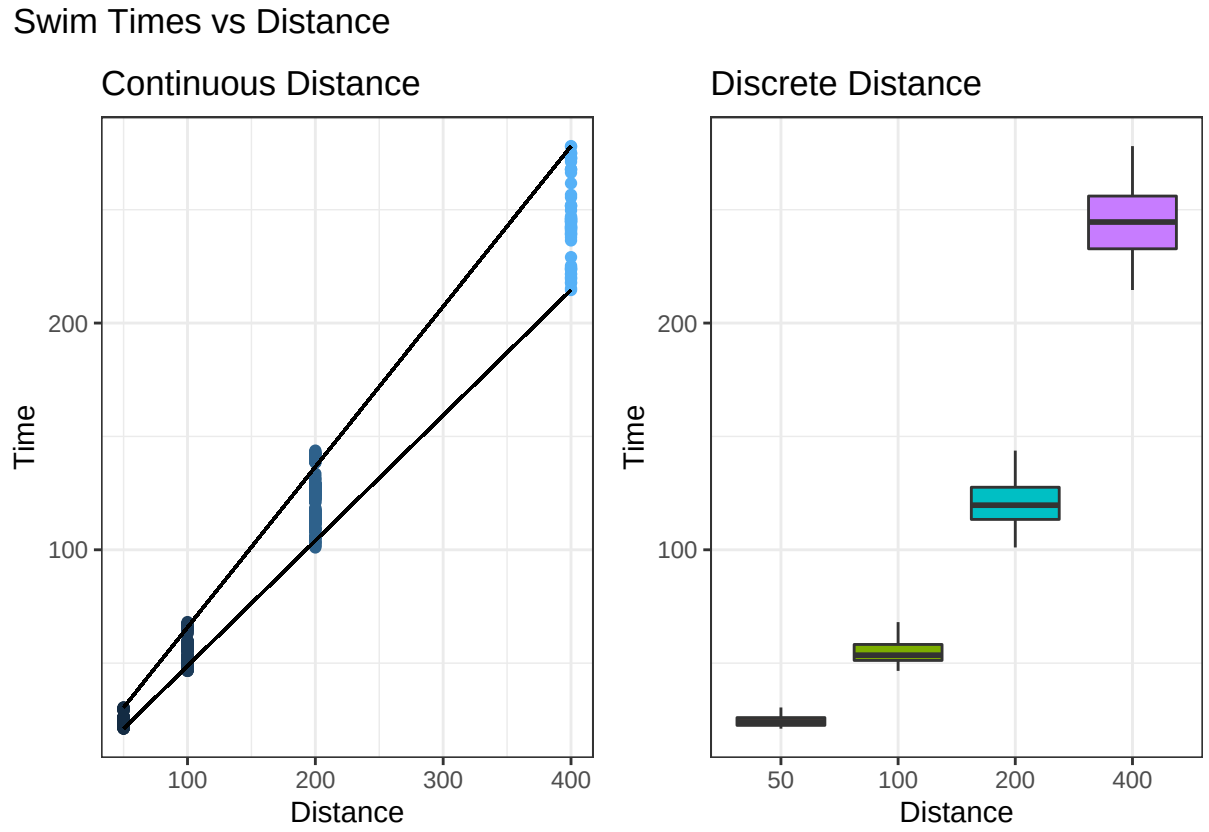
```
dist_boxplot <- swim %>%
  mutate(dist = as.factor(dist)) %>%
  ggplot() +
  geom_boxplot(aes(y = time, x = dist, fill = dist)) +
  guides(fill = "none") +
  theme_bw()

dist_histogram <- swim %>%
  mutate(dist = as.factor(dist)) %>%
  ggplot() +
  geom_histogram(aes(x = time, fill = dist),binwidth = 10) +
  facet_grid(dist ~ .) +
  guides(fill = "none") +
  theme_bw()

dist_boxplot + dist_histogram +
  plot_annotation(
  title = "Comparison of Times for Race Distance"
)
```



Looking at the boxplot and histogram of time plotted separately for each distance category, we can see that distance explains a large part of the variation in the times, with each of the four peaks in the distribution of times comprising of a different distance category. There is a greater variation in distance with time.

```
min_time_50m <- min(filter(swim,dist == 50)$time)
max_time_50m <- max(filter(swim,dist == 50)$time)

min_time_400m <- min(filter(swim,dist == 400)$time)
max_time_400m <- max(filter(swim,dist == 400)$time)

continous_distance_plot <- swim %>%
  ggplot() +
  geom_point(aes(x = dist, y = time, colour = dist)) +
  geom_segment(aes(x = 50, y = min_time_50m,
                   xend = 400, yend = min_time_400m), alpha = 0.7) +
  geom_segment(aes(x = 50, y = max_time_50m,
                   xend = 400, yend = max_time_400m), alpha = 0.7) +
  labs(title = "Continuous Distance", x = "Distance", y = "Time") +
  theme_bw() +
  guides(colour = "none")

discrete_distance_plot <- swim %>%
  mutate(dist = as.factor(dist)) %>%
  ggplot() +
  geom_boxplot(aes(x = dist, y = time, fill = dist)) +
  labs(title = "Discrete Distance", x = "Distance", y = "Time") +
  theme_bw() +
```

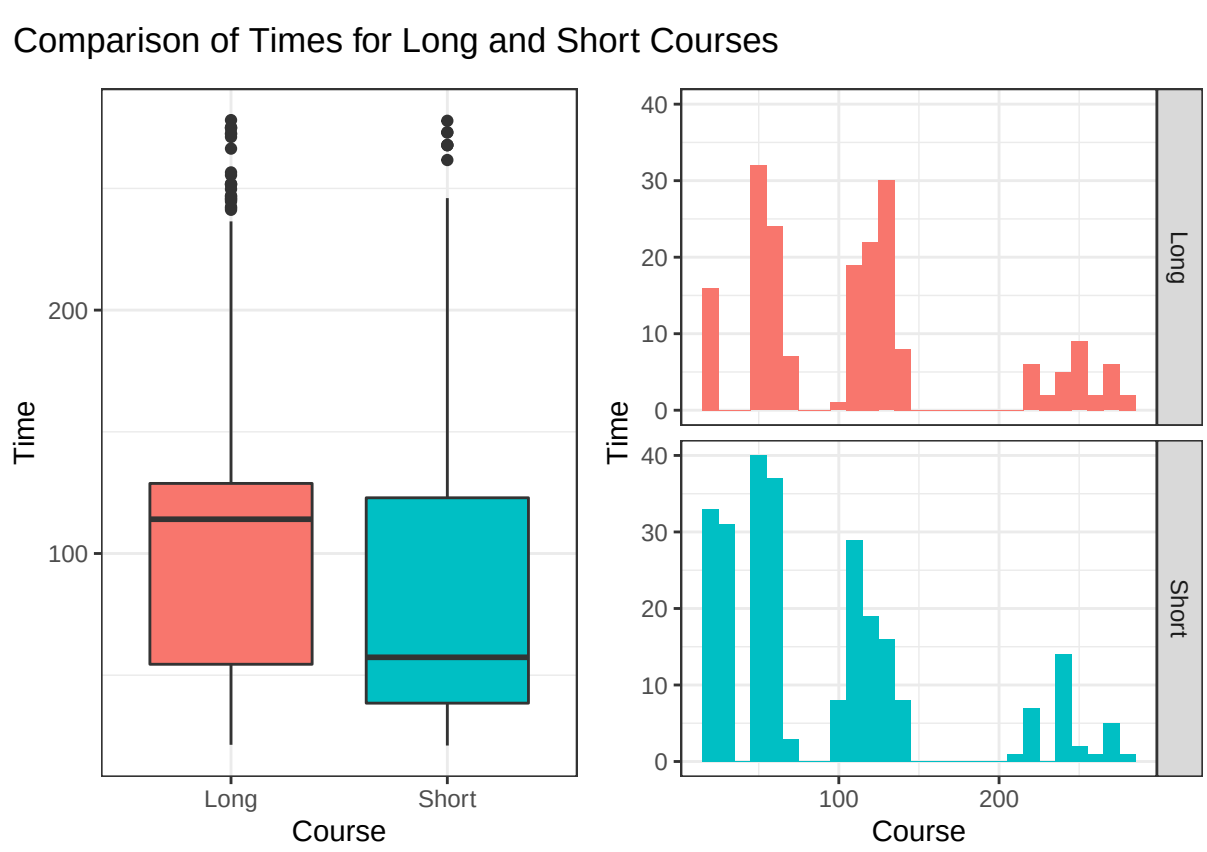
```
  guides(fill = "none")

continous_distance_plot + discrete_distance_plot +
  plot_annotation(
  title = "Swim Times vs Distance"
)
```



We can see that while it appears that there could be a linear relationship between distance and time, the variation in times increases with distance. This can also be observed in the boxplot.

```
course_boxplot <- swim %>%
  dplyr::select(time, course) %>%
  ggplot() +
  geom_boxplot(aes(y = time, x = course, fill = course)) +
  guides(fill = "none") +
  theme_bw() +
  labs(x = "Course", y = "Time")

course_histogram <- swim %>%
  ggplot() +
  geom_histogram(aes(x = time, fill = course),binwidth = 10) +
  facet_grid(course ~ .) +
  guides(fill = "none") +
  theme_bw() +
  labs(x = "Course", y = "Time")
```

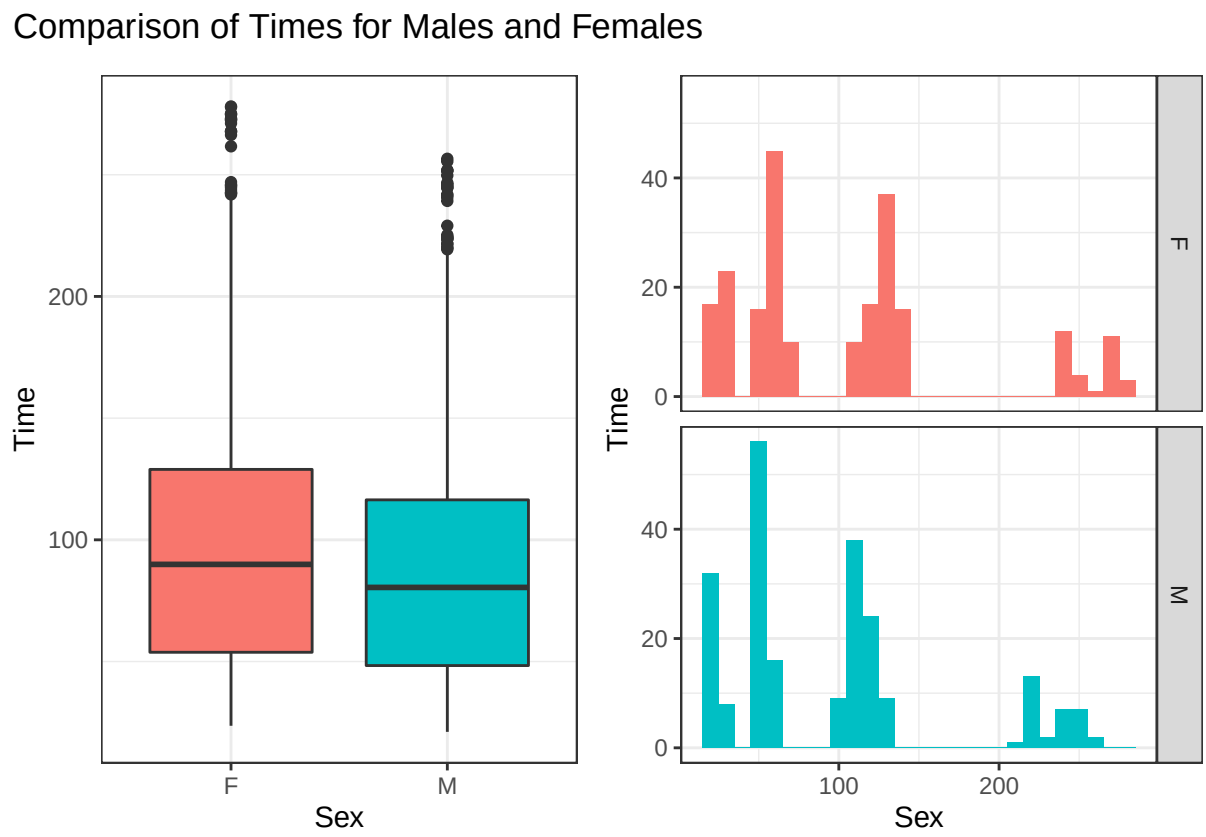
```
course_boxplot + course_histogram +
  plot_annotation(
  title = "Comparison of Times for Long and Short Courses"
)
```



From these plots, we can see that the median and modal times are higher for long courses and that there are some distributional differences for each different type of course. Looking at the histograms side by side however, we can see that the variation in time explained by the course type is much smaller than for distance.

```
sex_boxplot <- swim %>%
  dplyr::select(time, sex) %>%
  ggplot() +
  geom_boxplot(aes(y = time, x = sex, fill = sex)) +
  guides(fill = "none") +
  theme_bw() +
  labs(x = "Sex", y = "Time")

sex_histogram <- swim %>%
  ggplot() +
  geom_histogram(aes(x = time, fill = sex),binwidth = 10) +
  facet_grid(sex ~ .) +
  guides(fill = "none") +
  theme_bw() +
  labs(x = "Sex", y = "Time")

sex_boxplot + sex_histogram +
```

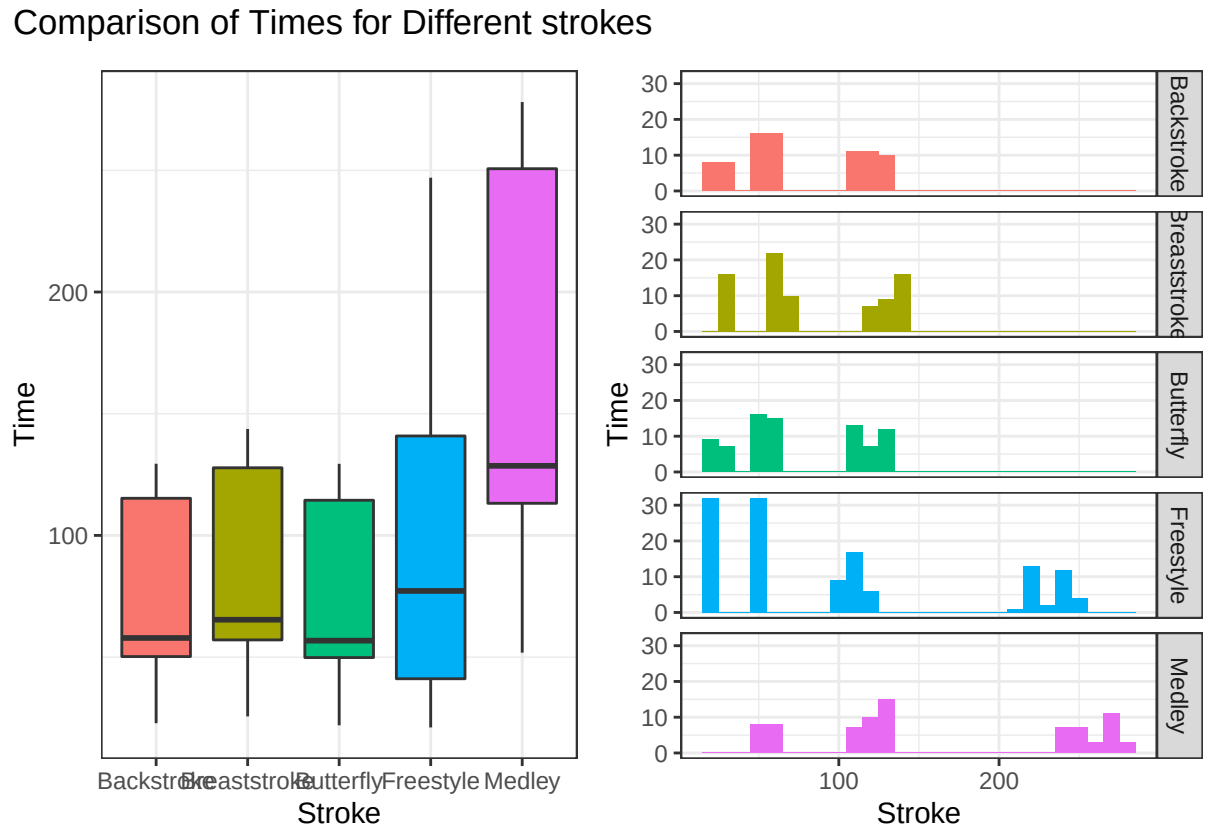
```
  plot_annotation(
  title = "Comparison of Times for Males and Females"
)
```



We see that there are some differences between the times for females and males, with females having slightly longer times overall.

```
stroke_boxplot <- swim %>%
  dplyr::select(time, stroke) %>%
  ggplot() +
  geom_boxplot(aes(y = time, x = stroke, fill = stroke)) +
  guides(fill = "none") +
  theme_bw() +
  labs(x = "Stroke", y = "Time")

stroke_histogram <- swim %>%
  ggplot() +
  geom_histogram(aes(x = time, fill = stroke),binwidth = 10) +
  guides(fill = "none") +
  facet_grid(stroke ~ .) +
  theme_bw() +
  labs(x = "Stroke", y = "Time")

stroke_boxplot + stroke_histogram +
  plot_annotation(
```

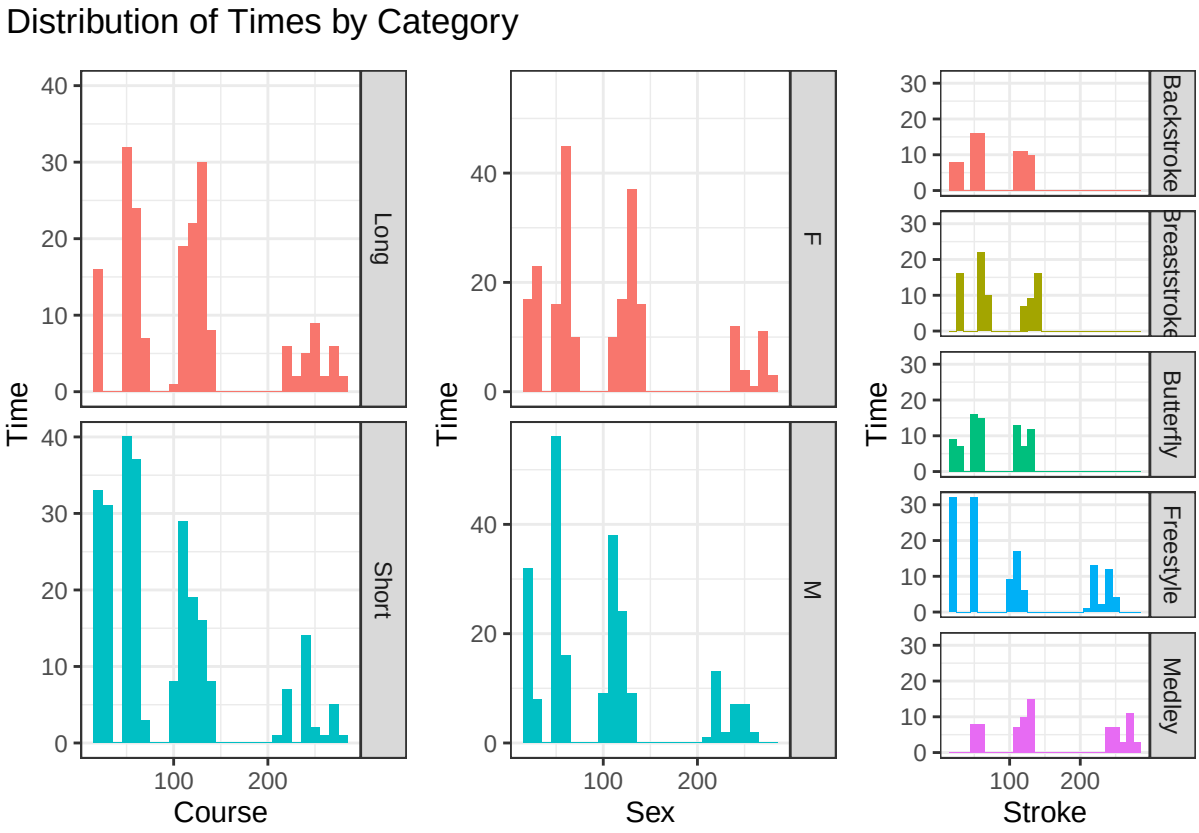
```
  title = "Comparison of Times for Different strokes"
)
```



We can see there are a few differences in the distributions of the stroke times. In particular, as we previously observed, most of the strokes do not participate in all distances, which appears to be the variable contributing most to the variation of the times for each of the strokes.

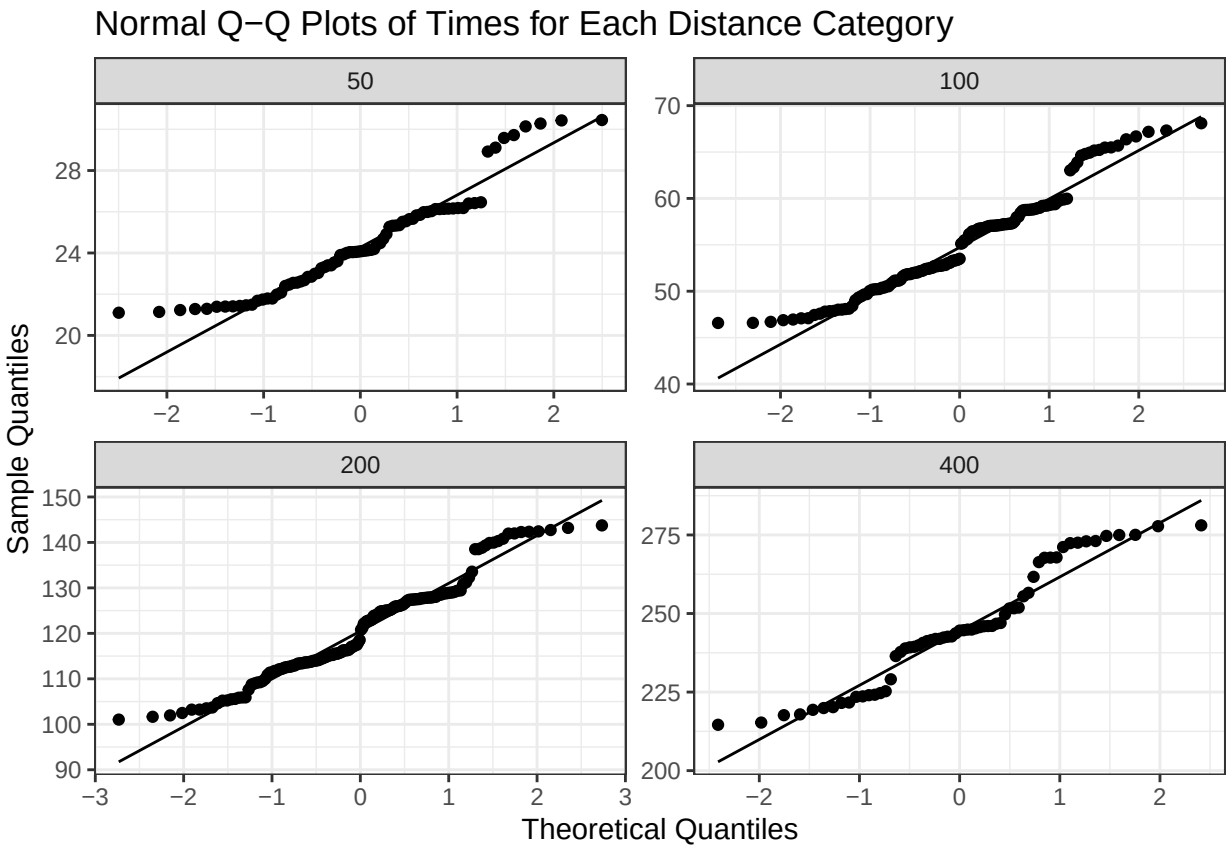
The following graph summarises all of the distributional observations for each of the categorical variables

```
(course_histogram + sex_histogram) + stroke_histogram +
  plot_layout(widths = c(2,2, 1.5)) +
  plot_annotation(title = "Distribution of Times by Category")
```



Of interest is the normality of the data. The histogram of the data clearly shows that the data is not normal, however within each distance category, variation appears normal, as the points mostly close to the ab-line, with some residual variation, and thus it may be appropriate to apply a normal linear model to the data.

```
swim %>%
  ggplot(aes(sample = time)) +
  geom_qq() +
  geom_qq_line() +
  facet_wrap(dist ~ .,scales = "free") +
  labs(title = "Normal Q-Q Plots of Times for Each Distance Category",
       x = "Theoretical Quantiles",
       y = "Sample Quantiles") +
  theme_bw()
```



Model Choice

The least parsimonious normal linear model applied to the data produces residuals that are clearly associated with the fitted values, which violates model assumptions:

```
swim_lm <- lm(time ~ dist*stroke*sex*course, data = swim)

summary(swim_lm)
```

```
##
## Call:
## lm(formula = time ~ dist * stroke * sex * course, data = swim)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6519 -1.1229 -0.0039  1.1405  8.4681
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       -10.122500   1.488599  -6.800 3.75e-11
## dist                                0.689837   0.009415  73.272  < 2e-16
## strokeBreaststroke                  0.693750   2.105197   0.330   0.7419
## strokeButterfly                    -3.015000   2.164509  -1.393   0.1644
## strokeFreestyle                     1.290163   1.594835   0.809   0.4190
```

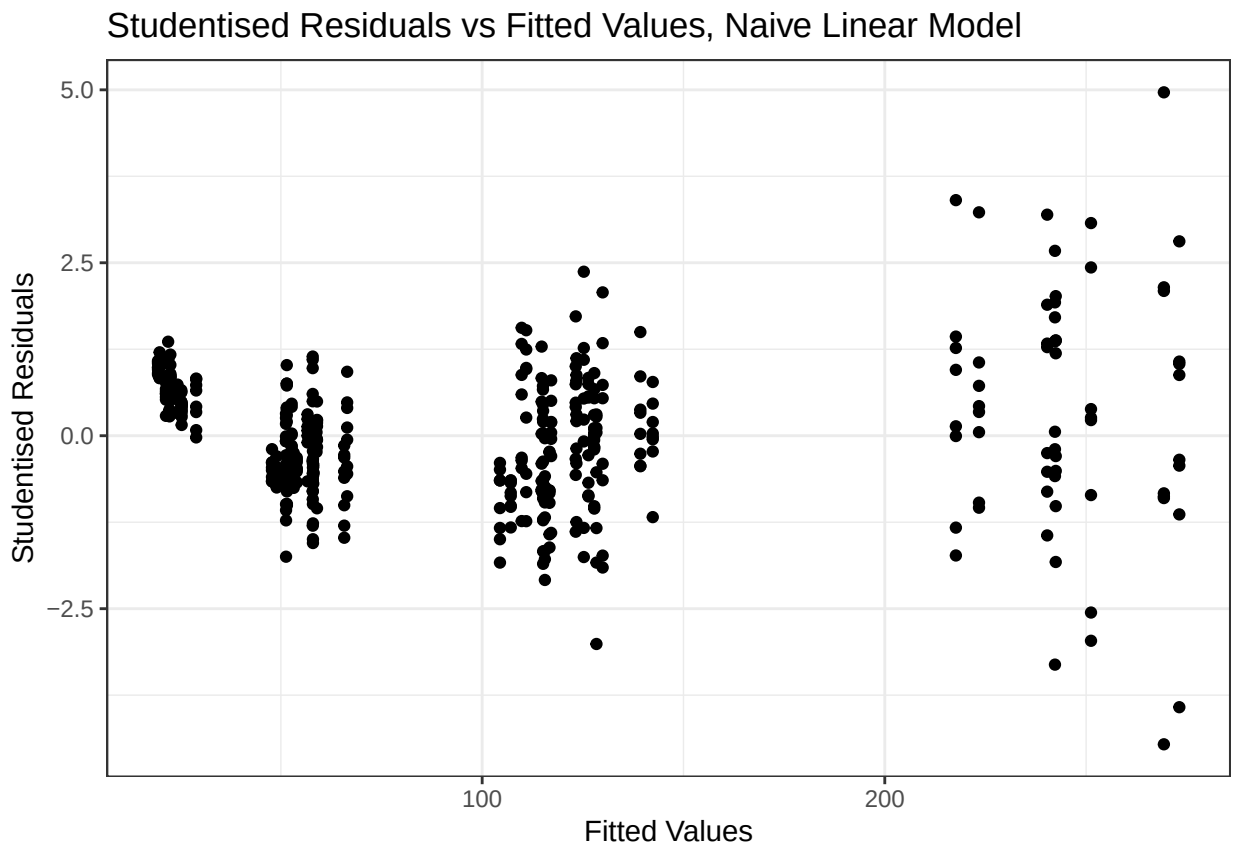
```
## strokeMedley                                -3.180000   2.105197  -1.511   0.1317
## sexM                                         0.396250   2.105197   0.188   0.8508
## courseShort                                  2.740625   1.697264   1.615   0.1071
## dist:strokeBreaststroke                      0.069163   0.013314   5.195 3.25e-07
## dist:strokeButterfly                         0.007838   0.013550   0.578   0.5633
## dist:strokeFreestyle                        -0.062088   0.009737  -6.377 4.93e-10
## dist:strokeMedley                            0.026300   0.010526   2.499   0.0129
## dist:sexM                                   -0.065737   0.013314  -4.937 1.16e-06
## strokeBreaststroke:sexM                     -0.822500   2.977197  -0.276   0.7825
## strokeButterfly:sexM                         0.523750   3.019429   0.173   0.8624
## strokeFreestyle:sexM                        -0.752772   2.255437  -0.334   0.7387
## strokeMedley:sexM                           -4.055000   2.977197  -1.362   0.1739
## dist:courseShort                            -0.036711   0.011253  -3.262   0.0012
## strokeBreaststroke:courseShort              -1.123125   2.400293  -0.468   0.6401
## strokeButterfly:courseShort                  1.089375   2.452480   0.444   0.6571
## strokeFreestyle:courseShort                 -3.090734   1.880385  -1.644   0.1010
## strokeMedley:courseShort                    -1.992620   2.403555  -0.829   0.4076
## sexM:courseShort                             0.177500   2.400293   0.074   0.9411
## dist:strokeBreaststroke:sexM                -0.004938   0.018829  -0.262   0.7933
## dist:strokeButterfly:sexM                    0.003012   0.018997   0.159   0.8741
## dist:strokeFreestyle:sexM                    0.019457   0.013770   1.413   0.1584
## dist:strokeMedley:sexM                       0.020056   0.014886   1.347   0.1786
## dist:strokeBreaststroke:courseShort          0.013225   0.015914   0.831   0.4064
## dist:strokeButterfly:courseShort             0.011807   0.016111   0.733   0.4641
## dist:strokeFreestyle:courseShort             0.032760   0.011788   2.779   0.0057
## dist:strokeMedley:courseShort                0.025264   0.012613   2.003   0.0458
## dist:sexM:courseShort                       -0.004284   0.015914  -0.269   0.7879
## strokeBreaststroke:sexM:courseShort          0.734375   3.394527   0.216   0.8288
## strokeButterfly:sexM:courseShort            -0.154375   3.431627  -0.045   0.9641
## strokeFreestyle:sexM:courseShort             0.540435   2.659266   0.203   0.8391
## strokeMedley:sexM:courseShort                3.849495   3.396834   1.133   0.2578
## dist:strokeBreaststroke:sexM:courseShort    -0.006839   0.022505  -0.304   0.7614
## dist:strokeButterfly:sexM:courseShort       -0.009214   0.022646  -0.407   0.6843
## dist:strokeFreestyle:sexM:courseShort       -0.007037   0.016671  -0.422   0.6732
## dist:strokeMedley:sexM:courseShort          -0.018106   0.017815  -1.016   0.3101
##
## (Intercept)                                ***
## dist                                       ***
## strokeBreaststroke
## strokeButterfly
## strokeFreestyle
## strokeMedley
## sexM
## courseShort
## dist:strokeBreaststroke                    ***
## dist:strokeButterfly
## dist:strokeFreestyle                       ***
## dist:strokeMedley                          *
## dist:sexM                                  ***
## strokeBreaststroke:sexM
## strokeButterfly:sexM
## strokeFreestyle:sexM
## strokeMedley:sexM
## dist:courseShort                           **
```

```
## strokeBreaststroke:courseShort
## strokeButterfly:courseShort
## strokeFreestyle:courseShort
## strokeMedley:courseShort
## sexM:courseShort
## dist:strokeBreaststroke:sexM
## dist:strokeButterfly:sexM
## dist:strokeFreestyle:sexM
## dist:strokeMedley:sexM
## dist:strokeBreaststroke:courseShort
## dist:strokeButterfly:courseShort
## dist:strokeFreestyle:courseShort         **
## dist:strokeMedley:courseShort            *
## dist:sexM:courseShort
## strokeBreaststroke:sexM:courseShort
## strokeButterfly:sexM:courseShort
## strokeFreestyle:sexM:courseShort
## strokeMedley:sexM:courseShort
## dist:strokeBreaststroke:sexM:courseShort
## dist:strokeButterfly:sexM:courseShort
## dist:strokeFreestyle:sexM:courseShort
## dist:strokeMedley:sexM:courseShort
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.883 on 406 degrees of freedom
## Multiple R-squared:  0.9993, Adjusted R-squared:  0.9993
## F-statistic: 1.579e+04 on 39 and 406 DF,  p-value: < 2.2e-16
```

```
#Residual standard error of the model
sqrt(deviance(swim_lm)/df.residual(swim_lm))
```

```
## [1] 1.882945
```

```
saved_values_numeric <- data.frame(name = "Residual Standard Error, Naive Model",
                                   value = sqrt(deviance(swim_lm)/df.residual(swim_lm)))

data.frame(fitted.values = swim_lm$fitted.values, studentised.residuals = rstudent(swim_lm)) %>%
  ggplot() +
  geom_point(aes(x = fitted.values, y = studentised.residuals)) +
  theme_bw() +
  labs(x = "Fitted Values", y = "Studentised Residuals",
       title = "Studentised Residuals vs Fitted Values, Naive Linear Model")
```

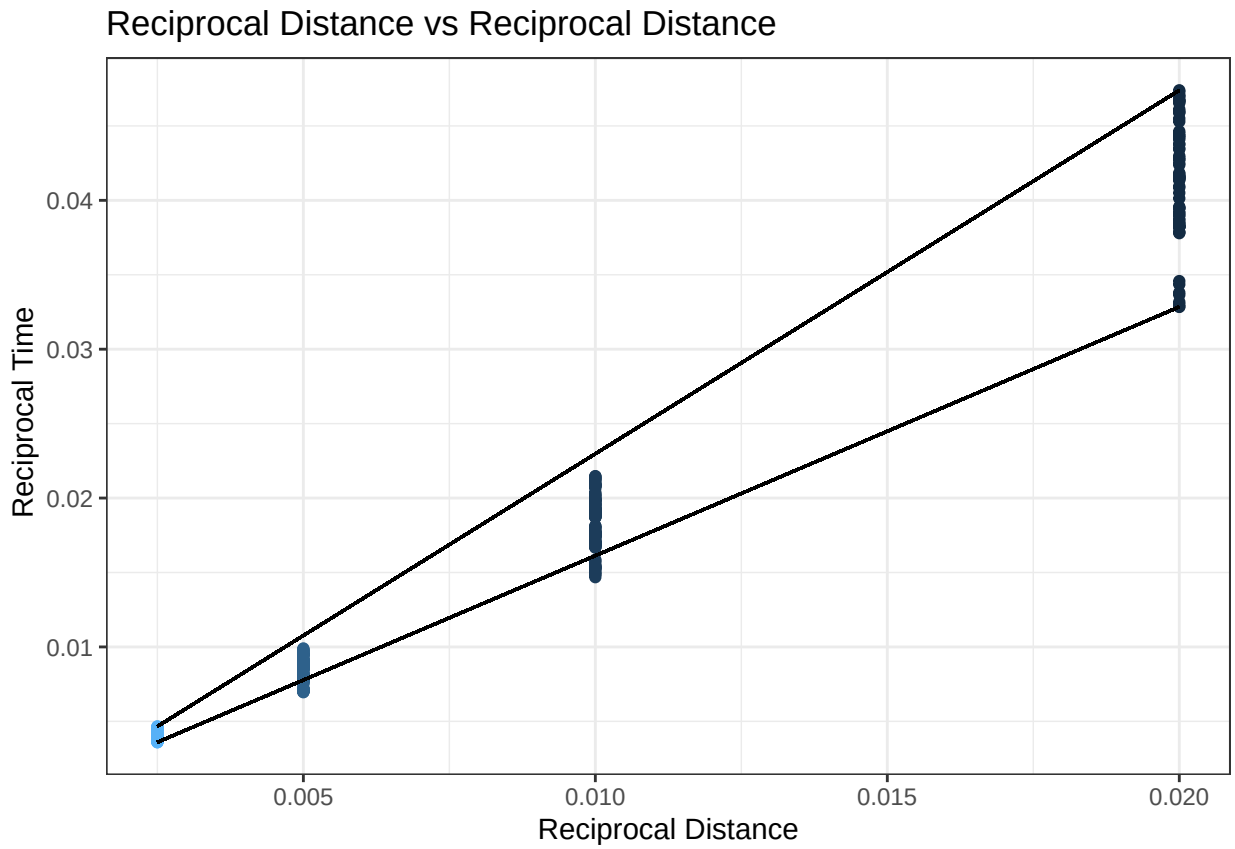


Looking at the plausibility of a physical model:

```
swim %>%
  mutate(reciprocal_time = 1/time, reciprocal_dist = 1/dist) -> swim

min_time_50m <- min(
  filter(swim,reciprocal_dist == min(swim$reciprocal_dist))$reciprocal_time)
max_time_50m <- max(
  filter(swim,reciprocal_dist == min(swim$reciprocal_dist))$reciprocal_time)

min_time_400m <- min(
  filter(swim,reciprocal_dist == max(swim$reciprocal_dist))$reciprocal_time)
max_time_400m <- max(
  filter(swim,reciprocal_dist == max(swim$reciprocal_dist))$reciprocal_time)



continous_distance_plot <- swim %>%
  ggplot() +
  geom_point(aes(x = reciprocal_dist, y = reciprocal_time, colour = dist)) +
  geom_segment(aes(x = min(swim$reciprocal_dist), y = min_time_50m,
                   xend = max(swim$reciprocal_dist), yend = min_time_400m), alpha = 0.7) +
  geom_segment(aes(x = min(swim$reciprocal_dist), y = max_time_50m,
                   xend = max(swim$reciprocal_dist), yend = max_time_400m), alpha = 0.7) +
  labs(title = "Reciprocal Distance vs Reciprocal Distance",
       x = "Reciprocal Distance", y = " Reciprocal Time") +
```

```
  theme_bw() +
  guides(colour = "none")

continous_distance_plot
```

```
## Warning: Use of ‘swim$reciprocal_dist‘ is discouraged. Use ‘reciprocal_dist‘
## instead.

## Warning: Use of ‘swim$reciprocal_dist‘ is discouraged. Use ‘reciprocal_dist‘
## instead.

## Warning: Use of ‘swim$reciprocal_dist‘ is discouraged. Use ‘reciprocal_dist‘
## instead.

## Warning: Use of ‘swim$reciprocal_dist‘ is discouraged. Use ‘reciprocal_dist‘
## instead.
```



Fitting weighted regressons 1 and 2:

```
weights1 <- swim$dist^2
swim_lm_weights1 <- lm(I(1/time) ~ (stroke + sex + course)*I(1/dist),
                       data = swim, weights = weights1)

weights2 <- 1/swim$dist^2
swim_lm_weights2 <- lm(time ~ (stroke + sex + course)*dist,
```

```
                      data = swim, weights = weights2)

summary(swim_lm_weights1)
```

```
##
## Call:
## lm(formula = I(1/time) ~ (stroke + sex + course) * I(1/dist),
##     data = swim, weights = weights1)
##
## Weighted Residuals:
##       Min        1Q    Median        3Q       Max
## -0.127966 -0.031701 -0.000053  0.031332  0.151546
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -1.686e-03  1.031e-04 -16.355  < 2e-16 ***
## strokeBreaststroke            3.478e-04  1.324e-04   2.628 0.008897 **
## strokeButterfly              -4.538e-04  1.325e-04  -3.426 0.000671 ***
## strokeFreestyle               6.778e-04  1.007e-04   6.728 5.47e-11 ***
## strokeMedley                  1.236e-03  1.054e-04  11.729  < 2e-16 ***
## sexM                         -2.557e-04  4.352e-05  -5.875 8.44e-09 ***
## courseShort                  -1.619e-04  4.554e-05  -3.555 0.000419 ***
## I(1/dist)                     1.892e+00  1.608e-02 117.632  < 2e-16 ***
## strokeBreaststroke:I(1/dist) -2.485e-01  1.895e-02 -13.113  < 2e-16 ***
## strokeButterfly:I(1/dist)     8.398e-02  1.900e-02   4.420 1.25e-05 ***
## strokeFreestyle:I(1/dist)     1.167e-01  1.571e-02   7.429 5.92e-13 ***
## strokeMedley:I(1/dist)       -2.522e-01  1.900e-02 -13.275  < 2e-16 ***
## sexM:I(1/dist)                2.395e-01  8.769e-03  27.316  < 2e-16 ***
## courseShort:I(1/dist)         8.488e-02  9.533e-03   8.904  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05023 on 432 degrees of freedom
## Multiple R-squared:  0.9979, Adjusted R-squared:  0.9979
## F-statistic: 1.601e+04 on 13 and 432 DF,  p-value: < 2.2e-16
```

```
summary(swim_lm_weights2)
```

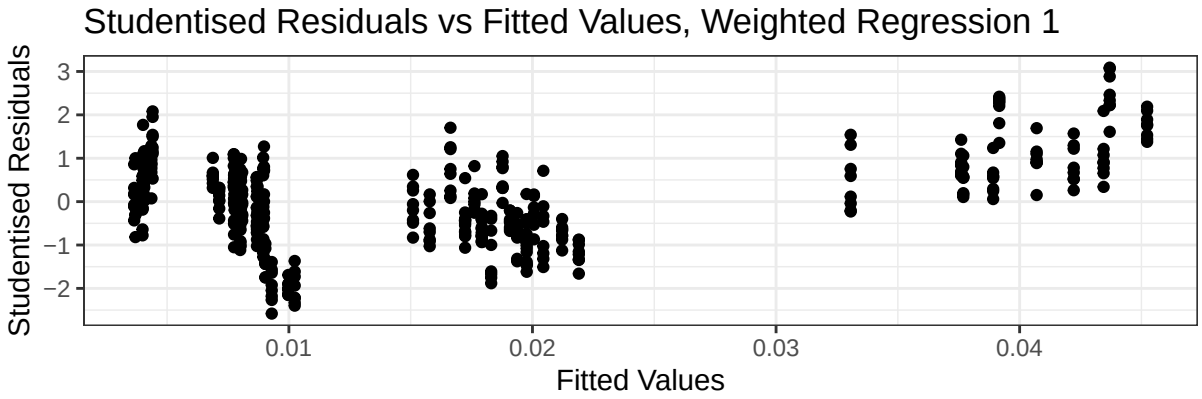
```
##
## Call:
## lm(formula = time ~ (stroke + sex + course) * dist, data = swim,
##     weights = weights2)
##
## Weighted Residuals:
##       Min        1Q    Median        3Q       Max
## -0.032632 -0.009030  0.000610  0.008994  0.034395
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -7.624996   0.363348 -20.985  < 2e-16 ***
## strokeBreaststroke      -0.444792   0.360198  -1.235  0.21756
## strokeButterfly         -1.153333   0.360198  -3.202  0.00147 **
```

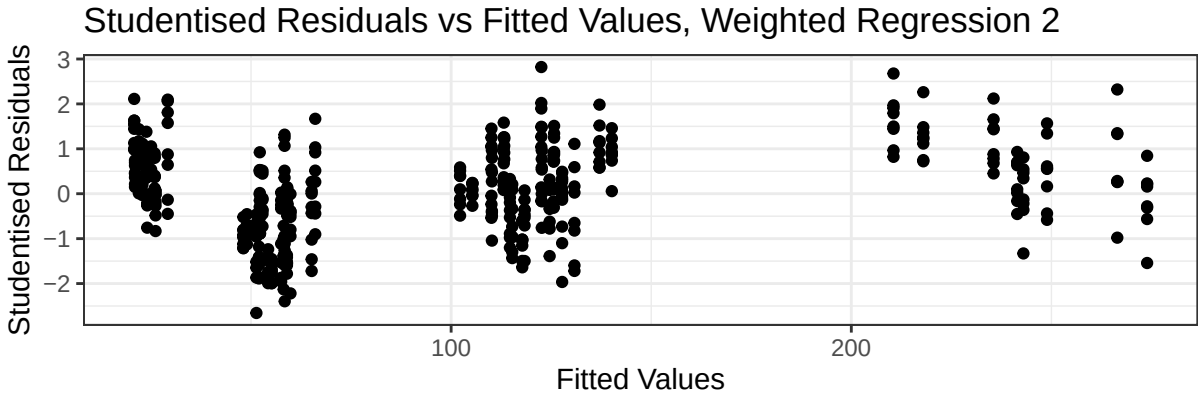
```
## strokeFreestyle          0.172795   0.323870   0.534  0.59394
## strokeMedley            -4.657222   0.577891  -8.059 7.60e-15 ***
## sexM                     0.088067   0.203475   0.433  0.66536
## courseShort              1.296839   0.258121   5.024 7.42e-07 ***
## dist                     0.666790   0.003566 186.981  < 2e-16 ***
## strokeBreaststroke:dist  0.074352   0.004107  18.104  < 2e-16 ***
## strokeButterfly:dist     0.005190   0.004110   1.263  0.20738
## strokeFreestyle:dist    -0.040653   0.003559 -11.424  < 2e-16 ***
## strokeMedley:dist        0.048708   0.004175  11.665  < 2e-16 ***
## sexM:dist               -0.062746   0.002169 -28.924  < 2e-16 ***
## courseShort:dist        -0.021867   0.002359  -9.269  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01248 on 432 degrees of freedom
## Multiple R-squared:  0.9988, Adjusted R-squared:  0.9987
## F-statistic: 2.686e+04 on 13 and 432 DF,  p-value: < 2.2e-16
```

```
# Saving RSS values
saved_values_numeric <- rbind(saved_values_numeric,
  data.frame(name = c("Residual Standard Error, Weights 1",
                      "Residual Standard Error, Weights 2"),
             value = c(sqrt(deviance(swim_lm_weights1)/df.residual(swim_lm_weights1)),
                       sqrt(deviance(swim_lm_weights2)/df.residual(swim_lm_weights2)))))
```

Errors are still problematic with these models:

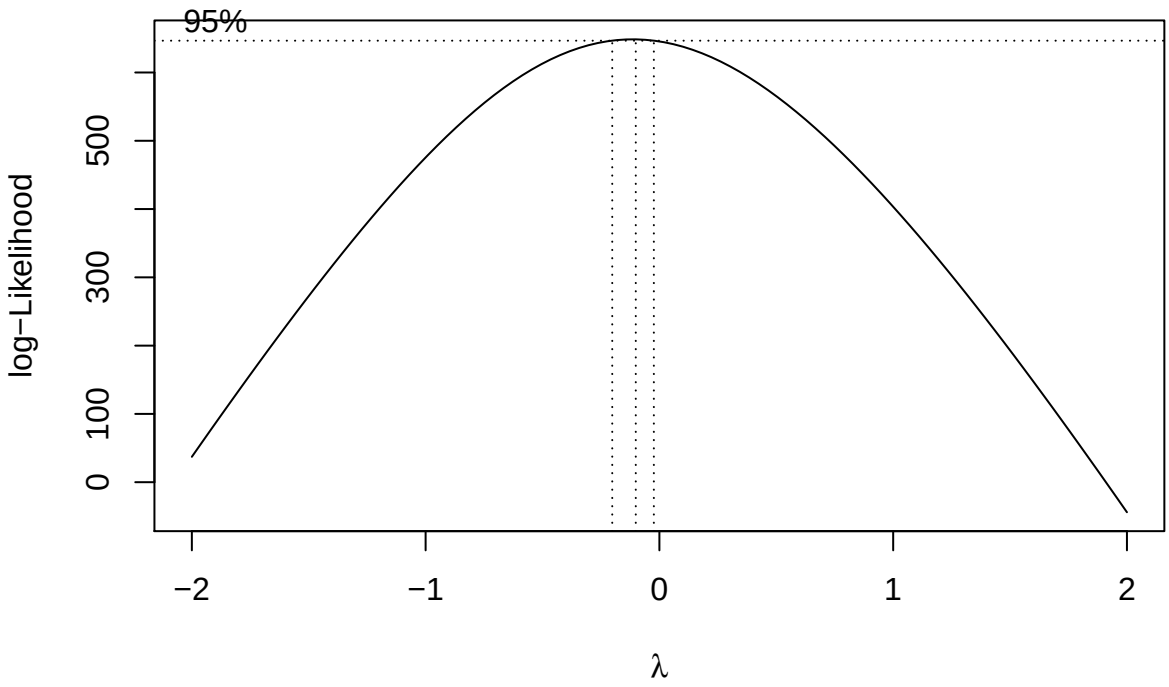
```
swim_lm_weights1_errors <- data.frame(fitted_values = fitted.values(swim_lm_weights1),
                                      studentised.residuals = rstudent(swim_lm_weights1))
swim_lm_weights2_errors <- data.frame(fitted_values = fitted.values(swim_lm_weights2),
                                      studentised.residuals = rstudent(swim_lm_weights2))

swim_lm_weights1_error_plot <- swim_lm_weights1_errors %>%
  ggplot() +
  geom_point(aes(x = fitted_values, y = studentised.residuals)) +
  theme_bw() +
  labs(x = "Fitted Values", y = "Studentised Residuals",
       title = "Studentised Residuals vs Fitted Values, Weighted Regression 1")

swim_lm_weights2_error_plot <- swim_lm_weights2_errors %>%
  ggplot() +
  geom_point(aes(x = fitted_values, y = studentised.residuals)) +
  theme_bw() +
  labs(x = "Fitted Values", y = "Studentised Residuals",
       title = "Studentised Residuals vs Fitted Values, Weighted Regression 2")

swim_lm_weights1_error_plot / swim_lm_weights2_error_plot
```





Now we consider a Box-Cox transformation:

```
swim$dist_fact <- as.factor(swim$dist)
swim_lm_discrete <- lm(time ~ dist_fact*stroke*sex*course, swim)

boxcox(swim_lm_discrete)
```



$\lambda = 0$ is best for interpretability.

Fitting the transformed model:

```
swim_lm_boxcox <- lm(log(time) ~ dist_fact*stroke*sex*course,data = swim)

summary(swim_lm_boxcox)
```

```
##
## Call:
## lm(formula = log(time) ~ dist_fact * stroke * sex * course, data = swim)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.033319 -0.006640  0.000408  0.007481  0.029191
##
## Coefficients: (24 not defined because of singularities)
##                                                    Estimate Std. Error t value
## (Intercept)                                        3.290468   0.014056 234.100
## dist_fact100                                       0.784707   0.014673  53.480
## dist_fact200                                       1.560306   0.013410 116.352
## dist_fact400                                       2.303469   0.012016 191.706
## strokeBreaststroke                                 0.115658   0.010314  11.213
## strokeButterfly                                   -0.060036   0.010314  -5.821
## strokeFreestyle                                   -0.102859   0.013410  -7.670
## strokeMedley                                       0.016022   0.005955   2.691
## sexM                                              -0.107424   0.019814  -5.422
```

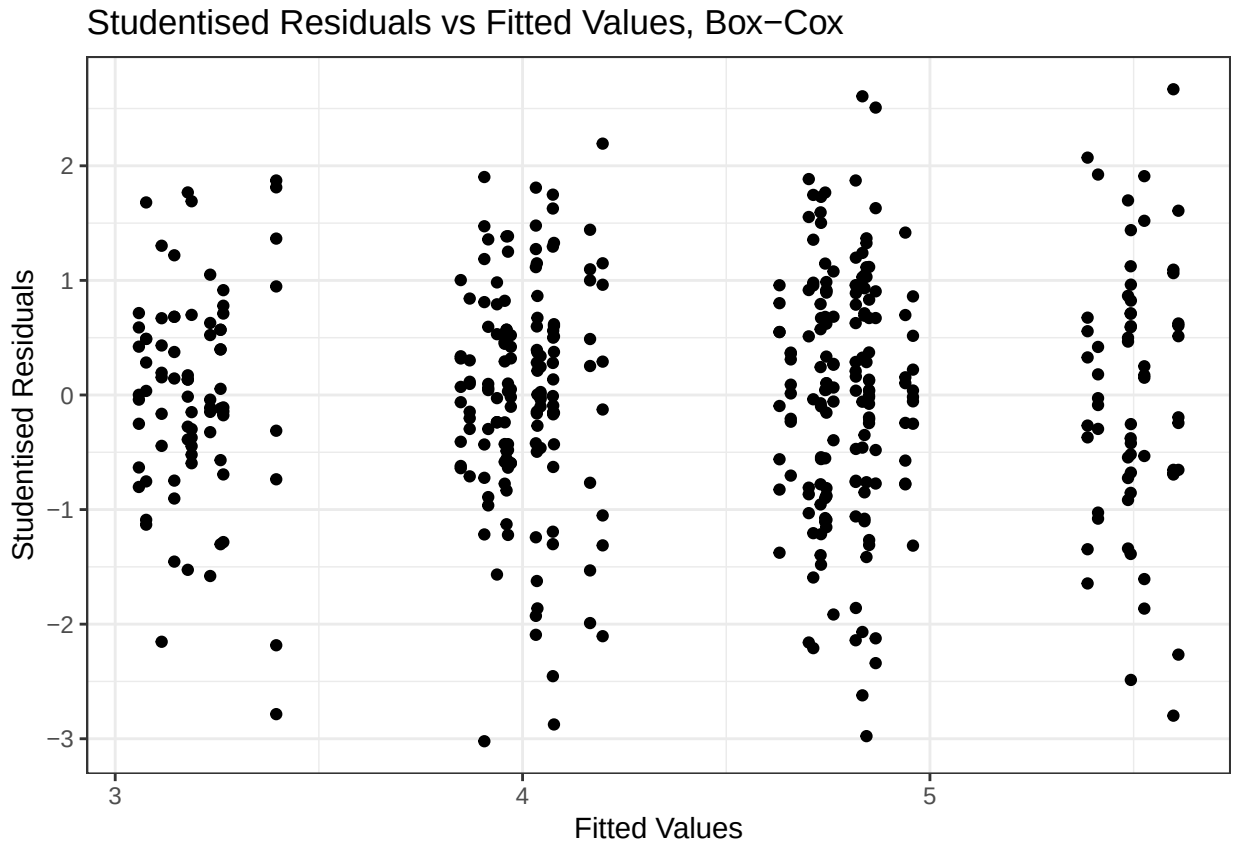
```
## courseShort                                    -0.025027   0.013410  -1.866
## dist_fact100:strokeBreaststroke                 0.005826   0.011910   0.489
## dist_fact200:strokeBreaststroke                -0.008015   0.008421  -0.952
## dist_fact400:strokeBreaststroke                       NA         NA      NA
## dist_fact100:strokeButterfly                    0.021358   0.012016   1.777
## dist_fact200:strokeButterfly                    0.048646   0.008421   5.776
## dist_fact400:strokeButterfly                          NA         NA      NA
## dist_fact100:strokeFreestyle                   -0.001067   0.014673  -0.073
## dist_fact200:strokeFreestyle                   -0.001956   0.013410  -0.146
## dist_fact400:strokeFreestyle                    0.002224   0.010436   0.213
## dist_fact100:strokeMedley                       0.012846   0.008421   1.525
## dist_fact200:strokeMedley                             NA         NA      NA
## dist_fact400:strokeMedley                             NA         NA      NA
## dist_fact100:sexM                              -0.003488   0.020690  -0.169
## dist_fact200:sexM                               0.002357   0.018898   0.125
## dist_fact400:sexM                               0.022191   0.016918   1.312
## strokeBreaststroke:sexM                        -0.012816   0.014586  -0.879
## strokeButterfly:sexM                            0.015120   0.014586   1.037
## strokeFreestyle:sexM                           -0.004190   0.018898  -0.222
## strokeMedley:sexM                               0.001513   0.008421   0.180
## dist_fact100:courseShort                       -0.014994   0.014673  -1.022
## dist_fact200:courseShort                       -0.007954   0.012016  -0.662
## dist_fact400:courseShort                        0.002346   0.008421   0.279
## strokeBreaststroke:courseShort                  0.014200   0.008421   1.686
## strokeButterfly:courseShort                     0.027833   0.008421   3.305
## strokeFreestyle:courseShort                     0.015636   0.012016   1.301
## strokeMedley:courseShort                        0.010396   0.008421   1.235
## sexM:courseShort                               -0.012888   0.018898  -0.682
## dist_fact100:strokeBreaststroke:sexM            0.004139   0.016843   0.246
## dist_fact200:strokeBreaststroke:sexM            0.010010   0.011910   0.840
## dist_fact400:strokeBreaststroke:sexM                  NA         NA      NA
## dist_fact100:strokeButterfly:sexM              -0.003486   0.016918  -0.206
## dist_fact200:strokeButterfly:sexM              -0.006550   0.011910  -0.550
## dist_fact400:strokeButterfly:sexM                     NA         NA      NA
## dist_fact100:strokeFreestyle:sexM               0.014184   0.020690   0.686
## dist_fact200:strokeFreestyle:sexM               0.021765   0.018898   1.152
## dist_fact400:strokeFreestyle:sexM               0.008790   0.014673   0.599
## dist_fact100:strokeMedley:sexM                  0.002146   0.011910   0.180
## dist_fact200:strokeMedley:sexM                        NA         NA      NA
## dist_fact400:strokeMedley:sexM                        NA         NA      NA
## dist_fact100:strokeBreaststroke:courseShort    -0.005134   0.011910  -0.431
## dist_fact200:strokeBreaststroke:courseShort           NA         NA      NA
## dist_fact400:strokeBreaststroke:courseShort           NA         NA      NA
## dist_fact100:strokeButterfly:courseShort        0.008420   0.012016   0.701
## dist_fact200:strokeButterfly:courseShort              NA         NA      NA
## dist_fact400:strokeButterfly:courseShort              NA         NA      NA
## dist_fact100:strokeFreestyle:courseShort        0.009349   0.014673   0.637
## dist_fact200:strokeFreestyle:courseShort        0.003025   0.012016   0.252
## dist_fact400:strokeFreestyle:courseShort              NA         NA      NA
## dist_fact100:strokeMedley:courseShort                 NA         NA      NA
## dist_fact200:strokeMedley:courseShort                 NA         NA      NA
## dist_fact400:strokeMedley:courseShort                 NA         NA      NA
## dist_fact100:sexM:courseShort                   0.004381   0.020690   0.212
## dist_fact200:sexM:courseShort                   0.002815   0.016918   0.166
```

```
## dist_fact400:sexM:courseShort                     -0.010229   0.011910  -0.859
## strokeBreaststroke:sexM:courseShort               -0.003520   0.011910  -0.296
## strokeButterfly:sexM:courseShort                  -0.014030   0.011910  -1.178
## strokeFreestyle:sexM:courseShort                   0.004506   0.016918   0.266
## strokeMedley:sexM:courseShort                      0.002067   0.011910   0.174
## dist_fact100:strokeBreaststroke:sexM:courseShort   0.010078   0.016843   0.598
## dist_fact200:strokeBreaststroke:sexM:courseShort         NA         NA      NA
## dist_fact400:strokeBreaststroke:sexM:courseShort         NA         NA      NA
## dist_fact100:strokeButterfly:sexM:courseShort     -0.004727   0.016918  -0.279
## dist_fact200:strokeButterfly:sexM:courseShort            NA         NA      NA
## dist_fact400:strokeButterfly:sexM:courseShort            NA         NA      NA
## dist_fact100:strokeFreestyle:sexM:courseShort     -0.002999   0.020690  -0.145
## dist_fact200:strokeFreestyle:sexM:courseShort     -0.007840   0.016918  -0.463
## dist_fact400:strokeFreestyle:sexM:courseShort            NA         NA      NA
## dist_fact100:strokeMedley:sexM:courseShort               NA         NA      NA
## dist_fact200:strokeMedley:sexM:courseShort               NA         NA      NA
## dist_fact400:strokeMedley:sexM:courseShort               NA         NA      NA
##                                                  Pr(>|t|)
## (Intercept)                                       < 2e-16 ***
## dist_fact100                                      < 2e-16 ***
## dist_fact200                                      < 2e-16 ***
## dist_fact400                                      < 2e-16 ***
## strokeBreaststroke                                < 2e-16 ***
## strokeButterfly                                  1.22e-08 ***
## strokeFreestyle                                  1.39e-13 ***
## strokeMedley                                      0.00744 **
## sexM                                             1.04e-07 ***
## courseShort                                       0.06276 .
## dist_fact100:strokeBreaststroke                   0.62498
## dist_fact200:strokeBreaststroke                   0.34179
## dist_fact400:strokeBreaststroke                        NA
## dist_fact100:strokeButterfly                      0.07627 .
## dist_fact200:strokeButterfly                     1.56e-08 ***
## dist_fact400:strokeButterfly                           NA
## dist_fact100:strokeFreestyle                      0.94207
## dist_fact200:strokeFreestyle                      0.88411
## dist_fact400:strokeFreestyle                      0.83137
## dist_fact100:strokeMedley                         0.12798
## dist_fact200:strokeMedley                              NA
## dist_fact400:strokeMedley                              NA
## dist_fact100:sexM                                 0.86620
## dist_fact200:sexM                                 0.90080
## dist_fact400:sexM                                 0.19040
## strokeBreaststroke:sexM                           0.38016
## strokeButterfly:sexM                              0.30058
## strokeFreestyle:sexM                              0.82466
## strokeMedley:sexM                                 0.85754
## dist_fact100:courseShort                          0.30748
## dist_fact200:courseShort                          0.50838
## dist_fact400:courseShort                          0.78069
## strokeBreaststroke:courseShort                    0.09256 .
## strokeButterfly:courseShort                       0.00104 **
## strokeFreestyle:courseShort                       0.19392
## strokeMedley:courseShort                          0.21776
```

```
## sexM:courseShort                                    0.49566
## dist_fact100:strokeBreaststroke:sexM                0.80599
## dist_fact200:strokeBreaststroke:sexM                0.40117
## dist_fact400:strokeBreaststroke:sexM                     NA
## dist_fact100:strokeButterfly:sexM                   0.83685
## dist_fact200:strokeButterfly:sexM                   0.58268
## dist_fact400:strokeButterfly:sexM                        NA
## dist_fact100:strokeFreestyle:sexM                   0.49340
## dist_fact200:strokeFreestyle:sexM                   0.25015
## dist_fact400:strokeFreestyle:sexM                   0.54946
## dist_fact100:strokeMedley:sexM                      0.85710
## dist_fact200:strokeMedley:sexM                           NA
## dist_fact400:strokeMedley:sexM                           NA
## dist_fact100:strokeBreaststroke:courseShort         0.66667
## dist_fact200:strokeBreaststroke:courseShort              NA
## dist_fact400:strokeBreaststroke:courseShort              NA
## dist_fact100:strokeButterfly:courseShort            0.48388
## dist_fact200:strokeButterfly:courseShort                 NA
## dist_fact400:strokeButterfly:courseShort                 NA
## dist_fact100:strokeFreestyle:courseShort            0.52439
## dist_fact200:strokeFreestyle:courseShort            0.80139
## dist_fact400:strokeFreestyle:courseShort                 NA
## dist_fact100:strokeMedley:courseShort                    NA
## dist_fact200:strokeMedley:courseShort                    NA
## dist_fact400:strokeMedley:courseShort                    NA
## dist_fact100:sexM:courseShort                       0.83240
## dist_fact200:sexM:courseShort                       0.86796
## dist_fact400:sexM:courseShort                       0.39095
## strokeBreaststroke:sexM:courseShort                 0.76770
## strokeButterfly:sexM:courseShort                    0.23951
## strokeFreestyle:sexM:courseShort                    0.79014
## strokeMedley:sexM:courseShort                       0.86231
## dist_fact100:strokeBreaststroke:sexM:courseShort    0.54997
## dist_fact200:strokeBreaststroke:sexM:courseShort         NA
## dist_fact400:strokeBreaststroke:sexM:courseShort         NA
## dist_fact100:strokeButterfly:sexM:courseShort       0.78007
## dist_fact200:strokeButterfly:sexM:courseShort            NA
## dist_fact400:strokeButterfly:sexM:courseShort            NA
## dist_fact100:strokeFreestyle:sexM:courseShort       0.88481
## dist_fact200:strokeFreestyle:sexM:courseShort       0.64332
## dist_fact400:strokeFreestyle:sexM:courseShort            NA
## dist_fact100:strokeMedley:sexM:courseShort               NA
## dist_fact200:strokeMedley:sexM:courseShort               NA
## dist_fact400:strokeMedley:sexM:courseShort               NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01191 on 390 degrees of freedom
## Multiple R-squared:  0.9998, Adjusted R-squared:  0.9997
## F-statistic: 3.104e+04 on 55 and 390 DF,  p-value: < 2.2e-16
```

```
swim_lm_boxcox_errors <- data.frame(fitted_values = fitted.values(swim_lm_boxcox),
                                    studentised.residuals = rstudent(swim_lm_boxcox))
```

```
swim_lm_boxcox_error_plot <- swim_lm_boxcox_errors %>%
  ggplot() +
  geom_point(aes(x = fitted_values, y = studentised.residuals)) +
  theme_bw() +
  labs(x = "Fitted Values", y = "Studentised Residuals",
       title = "Studentised Residuals vs Fitted Values, Box-Cox")

swim_lm_boxcox_error_plot
```




```
saved_values_numeric <- rbind(saved_values_numeric,
  data.frame(name = c("Residual Standard Error, Box-Cox Transformation Model"),
             value = c(sqrt(deviance(swim_lm_boxcox)/df.residual(swim_lm_boxcox)))))
```

Comparison of all fits with just distance:

```
swim_lm_dist <- lm(time ~ dist, data = swim)

swim_lm_boxcox_dist <- lm(log(time) ~ dist_fact,data = swim)


weights1 <- swim$dist^2
swim_lm_weights1_dist <- lm(I(1/time) ~ I(1/dist), data = swim, weights = weights1)

weights2 <- 1/swim$dist^2
```

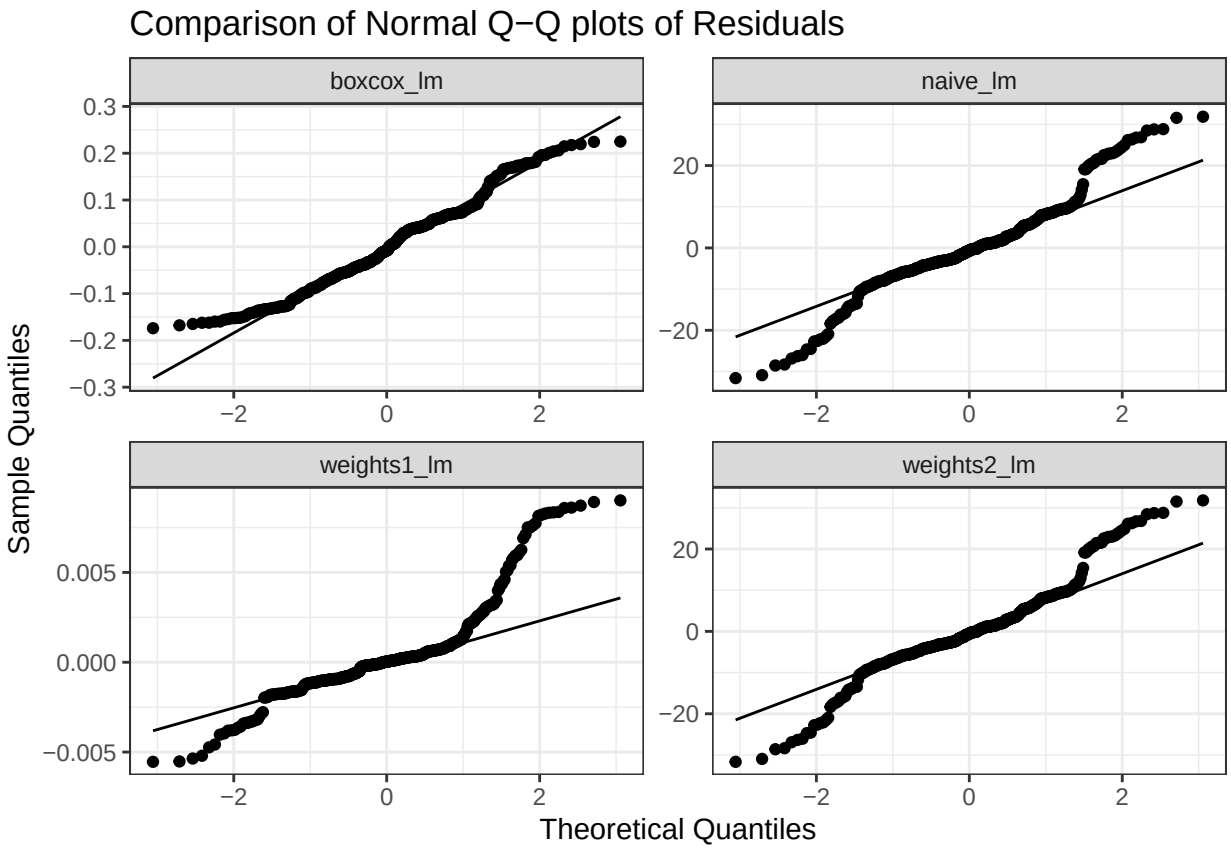
```r
swim_lm_weights2_dist <- lm(time ~ dist, data = swim, weights = weights2)

dist_fitted_values <- data.frame(naive_lm = swim_lm_dist$fitted.values,
                                 weights1_lm = 1/swim_lm_weights1_dist$fitted.values,
                                 weights2_lm = swim_lm_weights2_dist$fitted.values,
                                 boxcox_lm = exp(swim_lm_boxcox_dist$fitted.values),
                                 time = swim$time)


dist_fitted_values <- dist_fitted_values %>%
  mutate(naive_lm_residuals = (naive_lm - time)^2,
         weights1_lm_residuals = (weights1_lm - time)^2,
         weights2_lm_residuals = (weights2_lm - time)^2,
         boxcox_lm_residuals = (boxcox_lm - time)^2) %>%
  dplyr::select(naive_lm_residuals,weights1_lm_residuals,
                weights2_lm_residuals,boxcox_lm_residuals) %>%
  summarise(naive_RSS = sum(naive_lm_residuals),
            weights1_RSS = sum(weights1_lm_residuals),
            weights2_RSS = sum(weights2_lm_residuals),
            boxcox_RSS = sum(boxcox_lm_residuals))

dist_fitted_values <- signif(dist_fitted_values, digits = 4)

dist_fitted_values
```

```
##   naive_RSS weights1_RSS weights2_RSS boxcox_RSS
## 1     42740        64380        42740      42350
```

```r
write.csv(dist_fitted_values, file = "RSS.csv",quote = FALSE)
```

Comparison of QQ-plots:

```r
error_plot <- data.frame(data.frame(naive_lm = swim_lm_dist$residuals,
                                    weights1_lm = swim_lm_weights1_dist$residuals,
                                    weights2_lm = swim_lm_weights2_dist$residuals,
                                    boxcox_lm = swim_lm_boxcox_dist$residuals))

error_plot <- error_plot %>%
  pivot_longer(cols = 1:4, names_to = "Model", values_to = "Residuals")

error_plot %>%
  ggplot(aes(sample = Residuals)) +
  geom_qq() +
  geom_qq_line() +
  facet_wrap(Model ~ ., scales = "free") +
  theme_bw() +
  labs(title = "Comparison of Normal Q-Q plots of Residuals",
       x = "Theoretical Quantiles",y = "Sample Quantiles")
```



Variable Selection

Now that we have arrived at a chosen model structure we can look at variable selection.

Returning to the saturated Box-Cox model we can look at the significance of fitted value for a cursory view on which factors may be significant or not:

```
summary(swim_lm_boxcox)
```

```
##
## Call:
## lm(formula = log(time) ~ dist_fact * stroke * sex * course, data = swim)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.033319 -0.006640  0.000408  0.007481  0.029191
##
## Coefficients: (24 not defined because of singularities)
##                                                 Estimate Std. Error t value
## (Intercept)                                     3.290468   0.014056 234.100
## dist_fact100                                    0.784707   0.014673  53.480
## dist_fact200                                    1.560306   0.013410 116.352
## dist_fact400                                    2.303469   0.012016 191.706
## strokeBreaststroke                              0.115658   0.010314  11.213
```

```
## strokeButterfly                                -0.060036   0.010314  -5.821
## strokeFreestyle                                -0.102859   0.013410  -7.670
## strokeMedley                                    0.016022   0.005955   2.691
## sexM                                           -0.107424   0.019814  -5.422
## courseShort                                    -0.025027   0.013410  -1.866
## dist_fact100:strokeBreaststroke                 0.005826   0.011910   0.489
## dist_fact200:strokeBreaststroke                -0.008015   0.008421  -0.952
## dist_fact400:strokeBreaststroke                       NA         NA      NA
## dist_fact100:strokeButterfly                    0.021358   0.012016   1.777
## dist_fact200:strokeButterfly                    0.048646   0.008421   5.776
## dist_fact400:strokeButterfly                          NA         NA      NA
## dist_fact100:strokeFreestyle                   -0.001067   0.014673  -0.073
## dist_fact200:strokeFreestyle                   -0.001956   0.013410  -0.146
## dist_fact400:strokeFreestyle                    0.002224   0.010436   0.213
## dist_fact100:strokeMedley                       0.012846   0.008421   1.525
## dist_fact200:strokeMedley                             NA         NA      NA
## dist_fact400:strokeMedley                             NA         NA      NA
## dist_fact100:sexM                              -0.003488   0.020690  -0.169
## dist_fact200:sexM                               0.002357   0.018898   0.125
## dist_fact400:sexM                               0.022191   0.016918   1.312
## strokeBreaststroke:sexM                        -0.012816   0.014586  -0.879
## strokeButterfly:sexM                            0.015120   0.014586   1.037
## strokeFreestyle:sexM                           -0.004190   0.018898  -0.222
## strokeMedley:sexM                               0.001513   0.008421   0.180
## dist_fact100:courseShort                       -0.014994   0.014673  -1.022
## dist_fact200:courseShort                       -0.007954   0.012016  -0.662
## dist_fact400:courseShort                        0.002346   0.008421   0.279
## strokeBreaststroke:courseShort                  0.014200   0.008421   1.686
## strokeButterfly:courseShort                     0.027833   0.008421   3.305
## strokeFreestyle:courseShort                     0.015636   0.012016   1.301
## strokeMedley:courseShort                        0.010396   0.008421   1.235
## sexM:courseShort                               -0.012888   0.018898  -0.682
## dist_fact100:strokeBreaststroke:sexM            0.004139   0.016843   0.246
## dist_fact200:strokeBreaststroke:sexM            0.010010   0.011910   0.840
## dist_fact400:strokeBreaststroke:sexM                  NA         NA      NA
## dist_fact100:strokeButterfly:sexM              -0.003486   0.016918  -0.206
## dist_fact200:strokeButterfly:sexM              -0.006550   0.011910  -0.550
## dist_fact400:strokeButterfly:sexM                     NA         NA      NA
## dist_fact100:strokeFreestyle:sexM               0.014184   0.020690   0.686
## dist_fact200:strokeFreestyle:sexM               0.021765   0.018898   1.152
## dist_fact400:strokeFreestyle:sexM               0.008790   0.014673   0.599
## dist_fact100:strokeMedley:sexM                  0.002146   0.011910   0.180
## dist_fact200:strokeMedley:sexM                        NA         NA      NA
## dist_fact400:strokeMedley:sexM                        NA         NA      NA
## dist_fact100:strokeBreaststroke:courseShort    -0.005134   0.011910  -0.431
## dist_fact200:strokeBreaststroke:courseShort           NA         NA      NA
## dist_fact400:strokeBreaststroke:courseShort           NA         NA      NA
## dist_fact100:strokeButterfly:courseShort        0.008420   0.012016   0.701
## dist_fact200:strokeButterfly:courseShort              NA         NA      NA
## dist_fact400:strokeButterfly:courseShort              NA         NA      NA
## dist_fact100:strokeFreestyle:courseShort        0.009349   0.014673   0.637
## dist_fact200:strokeFreestyle:courseShort        0.003025   0.012016   0.252
## dist_fact400:strokeFreestyle:courseShort              NA         NA      NA
## dist_fact100:strokeMedley:courseShort                 NA         NA      NA
```

```
## dist_fact200:strokeMedley:courseShort                  NA         NA      NA
## dist_fact400:strokeMedley:courseShort                  NA         NA      NA
## dist_fact100:sexM:courseShort                       0.004381   0.020690   0.212
## dist_fact200:sexM:courseShort                       0.002815   0.016918   0.166
## dist_fact400:sexM:courseShort                      -0.010229   0.011910  -0.859
## strokeBreaststroke:sexM:courseShort                -0.003520   0.011910  -0.296
## strokeButterfly:sexM:courseShort                   -0.014030   0.011910  -1.178
## strokeFreestyle:sexM:courseShort                    0.004506   0.016918   0.266
## strokeMedley:sexM:courseShort                       0.002067   0.011910   0.174
## dist_fact100:strokeBreaststroke:sexM:courseShort    0.010078   0.016843   0.598
## dist_fact200:strokeBreaststroke:sexM:courseShort          NA         NA      NA
## dist_fact400:strokeBreaststroke:sexM:courseShort          NA         NA      NA
## dist_fact100:strokeButterfly:sexM:courseShort      -0.004727   0.016918  -0.279
## dist_fact200:strokeButterfly:sexM:courseShort             NA         NA      NA
## dist_fact400:strokeButterfly:sexM:courseShort             NA         NA      NA
## dist_fact100:strokeFreestyle:sexM:courseShort      -0.002999   0.020690  -0.145
## dist_fact200:strokeFreestyle:sexM:courseShort      -0.007840   0.016918  -0.463
## dist_fact400:strokeFreestyle:sexM:courseShort             NA         NA      NA
## dist_fact100:strokeMedley:sexM:courseShort                NA         NA      NA
## dist_fact200:strokeMedley:sexM:courseShort                NA         NA      NA
## dist_fact400:strokeMedley:sexM:courseShort                NA         NA      NA
##                                                   Pr(>|t|)
## (Intercept)                                        < 2e-16 ***
## dist_fact100                                       < 2e-16 ***
## dist_fact200                                       < 2e-16 ***
## dist_fact400                                       < 2e-16 ***
## strokeBreaststroke                                 < 2e-16 ***
## strokeButterfly                                   1.22e-08 ***
## strokeFreestyle                                   1.39e-13 ***
## strokeMedley                                       0.00744 **
## sexM                                              1.04e-07 ***
## courseShort                                        0.06276 .
## dist_fact100:strokeBreaststroke                    0.62498
## dist_fact200:strokeBreaststroke                    0.34179
## dist_fact400:strokeBreaststroke                         NA
## dist_fact100:strokeButterfly                       0.07627 .
## dist_fact200:strokeButterfly                      1.56e-08 ***
## dist_fact400:strokeButterfly                            NA
## dist_fact100:strokeFreestyle                       0.94207
## dist_fact200:strokeFreestyle                       0.88411
## dist_fact400:strokeFreestyle                       0.83137
## dist_fact100:strokeMedley                          0.12798
## dist_fact200:strokeMedley                               NA
## dist_fact400:strokeMedley                               NA
## dist_fact100:sexM                                  0.86620
## dist_fact200:sexM                                  0.90080
## dist_fact400:sexM                                  0.19040
## strokeBreaststroke:sexM                            0.38016
## strokeButterfly:sexM                               0.30058
## strokeFreestyle:sexM                               0.82466
## strokeMedley:sexM                                  0.85754
## dist_fact100:courseShort                           0.30748
## dist_fact200:courseShort                           0.50838
## dist_fact400:courseShort                           0.78069
```

```
## strokeBreaststroke:courseShort                    0.09256 .
## strokeButterfly:courseShort                       0.00104 **
## strokeFreestyle:courseShort                       0.19392
## strokeMedley:courseShort                          0.21776
## sexM:courseShort                                  0.49566
## dist_fact100:strokeBreaststroke:sexM              0.80599
## dist_fact200:strokeBreaststroke:sexM              0.40117
## dist_fact400:strokeBreaststroke:sexM                   NA
## dist_fact100:strokeButterfly:sexM                 0.83685
## dist_fact200:strokeButterfly:sexM                 0.58268
## dist_fact400:strokeButterfly:sexM                      NA
## dist_fact100:strokeFreestyle:sexM                 0.49340
## dist_fact200:strokeFreestyle:sexM                 0.25015
## dist_fact400:strokeFreestyle:sexM                 0.54946
## dist_fact100:strokeMedley:sexM                    0.85710
## dist_fact200:strokeMedley:sexM                         NA
## dist_fact400:strokeMedley:sexM                         NA
## dist_fact100:strokeBreaststroke:courseShort       0.66667
## dist_fact200:strokeBreaststroke:courseShort            NA
## dist_fact400:strokeBreaststroke:courseShort            NA
## dist_fact100:strokeButterfly:courseShort          0.48388
## dist_fact200:strokeButterfly:courseShort               NA
## dist_fact400:strokeButterfly:courseShort               NA
## dist_fact100:strokeFreestyle:courseShort          0.52439
## dist_fact200:strokeFreestyle:courseShort          0.80139
## dist_fact400:strokeFreestyle:courseShort               NA
## dist_fact100:strokeMedley:courseShort                  NA
## dist_fact200:strokeMedley:courseShort                  NA
## dist_fact400:strokeMedley:courseShort                  NA
## dist_fact100:sexM:courseShort                     0.83240
## dist_fact200:sexM:courseShort                     0.86796
## dist_fact400:sexM:courseShort                     0.39095
## strokeBreaststroke:sexM:courseShort               0.76770
## strokeButterfly:sexM:courseShort                  0.23951
## strokeFreestyle:sexM:courseShort                  0.79014
## strokeMedley:sexM:courseShort                     0.86231
## dist_fact100:strokeBreaststroke:sexM:courseShort  0.54997
## dist_fact200:strokeBreaststroke:sexM:courseShort       NA
## dist_fact400:strokeBreaststroke:sexM:courseShort       NA
## dist_fact100:strokeButterfly:sexM:courseShort     0.78007
## dist_fact200:strokeButterfly:sexM:courseShort          NA
## dist_fact400:strokeButterfly:sexM:courseShort          NA
## dist_fact100:strokeFreestyle:sexM:courseShort     0.88481
## dist_fact200:strokeFreestyle:sexM:courseShort     0.64332
## dist_fact400:strokeFreestyle:sexM:courseShort          NA
## dist_fact100:strokeMedley:sexM:courseShort             NA
## dist_fact200:strokeMedley:sexM:courseShort             NA
## dist_fact400:strokeMedley:sexM:courseShort             NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01191 on 390 degrees of freedom
## Multiple R-squared:  0.9998, Adjusted R-squared:  0.9997
## F-statistic: 3.104e+04 on 55 and 390 DF,  p-value: < 2.2e-16
```

```r
anova(swim_lm_boxcox)
```

```
## Analysis of Variance Table
##
## Response: log(time)
##                              Df  Sum Sq Mean Sq    F value    Pr(>F)
## dist_fact                     3 238.686  79.562 5.6092e+05 < 2.2e-16 ***
## stroke                        4   2.006   0.501 3.5356e+03 < 2.2e-16 ***
## sex                           1   1.338   1.338 9.4362e+03 < 2.2e-16 ***
## course                        1   0.059   0.059 4.1535e+02 < 2.2e-16 ***
## dist_fact:stroke              8   0.017   0.002 1.4619e+01 < 2.2e-16 ***
## dist_fact:sex                 3   0.009   0.003 2.0251e+01 3.233e-12 ***
## stroke:sex                    4   0.004   0.001 6.5202e+00 4.358e-05 ***
## dist_fact:course              3   0.002   0.001 4.2757e+00  0.005492 **
## stroke:course                 4   0.006   0.001 9.9974e+00 1.035e-07 ***
## sex:course                    1   0.004   0.004 2.7716e+01 2.326e-07 ***
## dist_fact:stroke:sex          8   0.001   0.000 7.8820e-01  0.613282
## dist_fact:stroke:course       4   0.000   0.000 4.1750e-01  0.796044
## dist_fact:sex:course          3   0.000   0.000 6.0600e-01  0.611439
## stroke:sex:course             4   0.001   0.000 1.5641e+00  0.183126
## dist_fact:stroke:sex:course   4   0.000   0.000 2.5560e-01  0.906168
## Residuals                   390   0.055   0.000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see that the main influences seem to be the baseline influences of each factor, apart from the course length.

We use automatic model selection using the Akaike information criterion to find a model that balances fit and parsimony:

```r
# The minimal model we can fit
null_lm <- lm(log(time) ~ 1, data = swim)
# Saturated model
sat_lm <- swim_lm_boxcox
```

Forward selection:

```r
forward_selection <- stepAIC(null_lm, scope = list(lower = null_lm, upper = sat_lm),
        data = swim, direction = "forward")
```

```
## Start:  AIC=-270.33
## log(time) ~ 1
##
##              Df Sum of Sq     RSS      AIC
## + dist_fact   3   238.686   3.501 -2153.83
## + stroke      4    34.909 207.278  -331.75
## + course      1     9.715 232.472  -286.59
## + sex         1     1.248 240.938  -270.64
## <none>                    242.187  -270.33
##
## Step:  AIC=-2153.83
```

```
## log(time) ~ dist_fact
##
##          Df Sum of Sq    RSS     AIC
## + stroke  4   2.00598 1.4954 -2525.3
## + sex     1   1.33470 2.1667 -2365.9
## <none>                3.5014 -2153.8
## + course  1   0.01406 3.4873 -2153.6
##
## Step:  AIC=-2525.27
## log(time) ~ dist_fact + stroke
##
##                     Df Sum of Sq     RSS     AIC
## + sex                1   1.33845 0.15698 -3528.6
## + course             1   0.05848 1.43695 -2541.1
## <none>                           1.49543 -2525.3
## + dist_fact:stroke   8   0.01875 1.47668 -2514.9
##
## Step:  AIC=-3528.57
## log(time) ~ dist_fact + stroke + sex
##
##                     Df Sum of Sq      RSS     AIC
## + course             1  0.058914 0.098068 -3736.4
## + dist_fact:stroke   8  0.018286 0.138695 -3567.8
## + dist_fact:sex      3  0.008410 0.148572 -3547.1
## + stroke:sex         4  0.006391 0.150591 -3539.1
## <none>                           0.156982 -3528.6
##
## Step:  AIC=-3736.4
## log(time) ~ dist_fact + stroke + sex + course
##
##                     Df Sum of Sq      RSS     AIC
## + dist_fact:stroke   8 0.0165885 0.081480 -3803.0
## + dist_fact:sex      3 0.0086399 0.089428 -3771.5
## + sex:course         1 0.0065670 0.091501 -3765.3
## + stroke:sex         4 0.0064948 0.091573 -3759.0
## + stroke:course      4 0.0045313 0.093537 -3749.5
## <none>                           0.098068 -3736.4
## + dist_fact:course   3 0.0011317 0.096936 -3735.6
##
## Step:  AIC=-3803.04
## log(time) ~ dist_fact + stroke + sex + course + dist_fact:stroke
##
##                     Df Sum of Sq      RSS     AIC
## + dist_fact:sex      3 0.0086174 0.072862 -3846.9
## + sex:course         1 0.0066297 0.074850 -3838.9
## + stroke:course      4 0.0071271 0.074352 -3835.9
## + stroke:sex         4 0.0064734 0.075006 -3832.0
## + dist_fact:course   3 0.0018468 0.079633 -3807.3
## <none>                           0.081480 -3803.0
##
## Step:  AIC=-3846.9
## log(time) ~ dist_fact + stroke + sex + course + dist_fact:stroke +
##     dist_fact:sex
##
```

```
##                     Df Sum of Sq      RSS     AIC
## + stroke:course      4 0.0071156 0.065747 -3884.7
## + sex:course         1 0.0045278 0.068334 -3873.5
## + stroke:sex         4 0.0036993 0.069163 -3862.1
## + dist_fact:course   3 0.0018094 0.071053 -3852.1
## <none>                           0.072862 -3846.9
##
## Step:  AIC=-3884.73
## log(time) ~ dist_fact + stroke + sex + course + dist_fact:stroke +
##     dist_fact:sex + stroke:course
##
##                     Df Sum of Sq      RSS     AIC
## + sex:course         1 0.0045535 0.061193 -3914.7
## + stroke:sex         4 0.0036993 0.062047 -3902.6
## <none>                           0.065747 -3884.7
## + dist_fact:course   3 0.0003733 0.065373 -3881.3
##
## Step:  AIC=-3914.74
## log(time) ~ dist_fact + stroke + sex + course + dist_fact:stroke +
##     dist_fact:sex + stroke:course + sex:course
##
##                     Df  Sum of Sq      RSS     AIC
## + stroke:sex         4 0.00307829 0.058115 -3929.8
## <none>                            0.061193 -3914.7
## + dist_fact:course   3 0.00037302 0.060820 -3911.5
##
## Step:  AIC=-3929.76
## log(time) ~ dist_fact + stroke + sex + course + dist_fact:stroke +
##     dist_fact:sex + stroke:course + sex:course + stroke:sex
##
##                         Df  Sum of Sq      RSS     AIC
## <none>                                0.058115 -3929.8
## + dist_fact:course       3 0.00037485 0.057740 -3926.6
## + stroke:sex:course      4 0.00052279 0.057592 -3925.8
## + dist_fact:stroke:sex   8 0.00089438 0.057220 -3920.7
```

Backward selection:

```
backward_selection <- stepAIC(sat_lm, scope = list(lower = null_lm, upper = sat_lm),
        data = swim, direction = "backward")
```

```
## Start:  AIC=-3899.76
## log(time) ~ dist_fact * stroke * sex * course
##
##                                Df  Sum of Sq      RSS     AIC
## - dist_fact:stroke:sex:course   4 0.00014503 0.055463 -3906.6
## <none>                                      0.055318 -3899.8
##
## Step:  AIC=-3906.59
## log(time) ~ dist_fact + stroke + sex + course + dist_fact:stroke +
##     dist_fact:sex + stroke:sex + dist_fact:course + stroke:course +
##     sex:course + dist_fact:stroke:sex + dist_fact:stroke:course +
##     dist_fact:sex:course + stroke:sex:course
```

```
##
##                            Df  Sum of Sq      RSS     AIC
## - dist_fact:stroke:sex      8 0.00094454 0.056408 -3915.1
## - dist_fact:stroke:course   4 0.00024058 0.055704 -3912.7
## - dist_fact:sex:course      3 0.00033733 0.055801 -3909.9
## - stroke:sex:course         4 0.00088742 0.056351 -3907.5
## <none>                                   0.055463 -3906.6
##
## Step:  AIC=-3915.06
## log(time) ~ dist_fact + stroke + sex + course + dist_fact:stroke +
##     dist_fact:sex + stroke:sex + dist_fact:course + stroke:course +
##     sex:course + dist_fact:stroke:course + dist_fact:sex:course +
##     stroke:sex:course
##
##                            Df  Sum of Sq      RSS     AIC
## - dist_fact:stroke:course   4 0.00024037 0.056648 -3921.2
## - stroke:sex:course         4 0.00058698 0.056995 -3918.4
## - dist_fact:sex:course      3 0.00057237 0.056980 -3916.6
## <none>                                   0.056408 -3915.1
##
## Step:  AIC=-3921.16
## log(time) ~ dist_fact + stroke + sex + course + dist_fact:stroke +
##     dist_fact:sex + stroke:sex + dist_fact:course + stroke:course +
##     sex:course + dist_fact:sex:course + stroke:sex:course
##
##                          Df Sum of Sq      RSS     AIC
## - stroke:sex:course       4 0.0005841 0.057232 -3924.6
## - dist_fact:sex:course    3 0.0005723 0.057221 -3922.7
## <none>                                0.056648 -3921.2
## - dist_fact:stroke        8 0.0194739 0.076122 -3805.4
##
## Step:  AIC=-3924.59
## log(time) ~ dist_fact + stroke + sex + course + dist_fact:stroke +
##     dist_fact:sex + stroke:sex + dist_fact:course + stroke:course +
##     sex:course + dist_fact:sex:course
##
##                          Df Sum of Sq      RSS     AIC
## - dist_fact:sex:course    3 0.0005076 0.057740 -3926.6
## <none>                                0.057232 -3924.6
## - stroke:sex              4 0.0033238 0.060556 -3907.4
## - stroke:course           4 0.0056818 0.062914 -3890.4
## - dist_fact:stroke        8 0.0194173 0.076650 -3810.3
##
## Step:  AIC=-3926.65
## log(time) ~ dist_fact + stroke + sex + course + dist_fact:stroke +
##     dist_fact:sex + stroke:sex + dist_fact:course + stroke:course +
##     sex:course
##
##                        Df Sum of Sq      RSS     AIC
## - dist_fact:course      3 0.0003748 0.058115 -3929.8
## <none>                              0.057740 -3926.6
## - stroke:sex            4 0.0030801 0.060820 -3911.5
## - dist_fact:sex         3 0.0039088 0.061649 -3903.4
## - sex:course            1 0.0039313 0.061671 -3899.3
```

```
## - stroke:course     4 0.0056971 0.063437 -3892.7
## - dist_fact:stroke  8 0.0194372 0.077177 -3813.2
##
## Step:  AIC=-3929.76
## log(time) ~ dist_fact + stroke + sex + course + dist_fact:stroke +
##     dist_fact:sex + stroke:sex + stroke:course + sex:course
##
##                    Df Sum of Sq      RSS     AIC
## <none>                           0.058115 -3929.8
## - stroke:sex        4 0.0030783 0.061193 -3914.7
## - dist_fact:sex     3 0.0039135 0.062028 -3906.7
## - sex:course        1 0.0039324 0.062047 -3902.6
## - stroke:course     4 0.0071410 0.065256 -3886.1
## - dist_fact:stroke  8 0.0192368 0.077352 -3818.2
```

```
backward_selection
```

```
##
## Call:
## lm(formula = log(time) ~ dist_fact + stroke + sex + course +
##     dist_fact:stroke + dist_fact:sex + stroke:sex + stroke:course +
##     sex:course, data = swim)
##
## Coefficients:
##                       (Intercept)                       dist_fact100
##                          3.303498                           0.768452
##                       dist_fact200                       dist_fact400
##                          1.547379                           2.297152
##               strokeBreaststroke                   strokeButterfly
##                          0.112923                          -0.056225
##                  strokeFreestyle                       strokeMedley
##                         -0.117268                           0.012709
##                              sexM                        courseShort
##                         -0.114135                          -0.034938
## dist_fact100:strokeBreaststroke  dist_fact200:strokeBreaststroke
##                          0.007825                          -0.003034
## dist_fact400:strokeBreaststroke     dist_fact100:strokeButterfly
##                                NA                           0.024210
##    dist_fact200:strokeButterfly     dist_fact400:strokeButterfly
##                          0.046831                                NA
##    dist_fact100:strokeFreestyle     dist_fact200:strokeFreestyle
##                          0.014787                           0.013316
##    dist_fact400:strokeFreestyle        dist_fact100:strokeMedley
##                          0.011547                           0.009891
##       dist_fact200:strokeMedley        dist_fact400:strokeMedley
##                                NA                                NA
##                dist_fact100:sexM                 dist_fact200:sexM
##                          0.006544                           0.011990
##                dist_fact400:sexM          strokeBreaststroke:sexM
##                          0.022201                          -0.007253
##             strokeButterfly:sexM             strokeFreestyle:sexM
##                          0.001656                           0.009069
##                strokeMedley:sexM   strokeBreaststroke:courseShort
##                          0.003211                           0.012393
```

```
##     strokeButterfly:courseShort      strokeFreestyle:courseShort
##                        0.023738                         0.023773
##        strokeMedley:courseShort                 sexM:courseShort
##                        0.016357                        -0.012416
```

ANOVA of all second order terms:

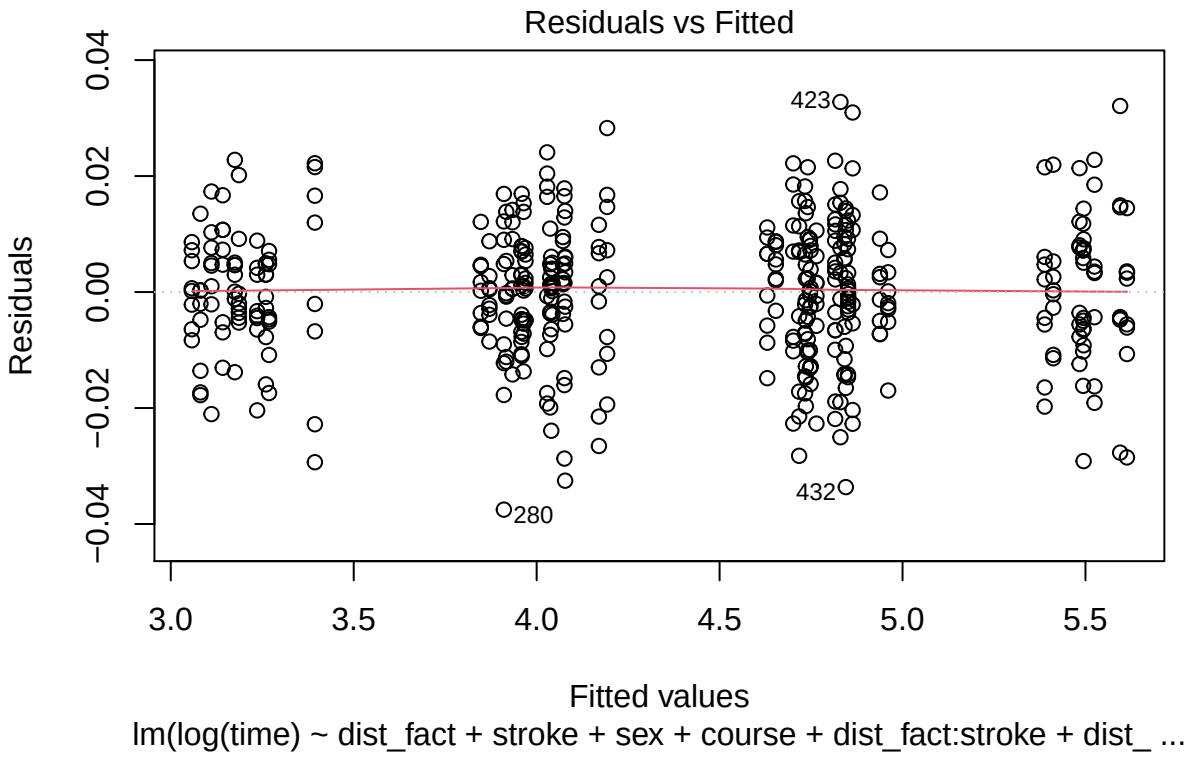
```
swim_lm_second_order <- lm(log(time) ~ dist_fact + stroke + sex + course
                           + dist_fact:stroke + dist_fact:sex + stroke:sex
                           + stroke:course + sex:course + dist_fact:course,
                           data = swim)
anova(swim_lm_second_order)
```

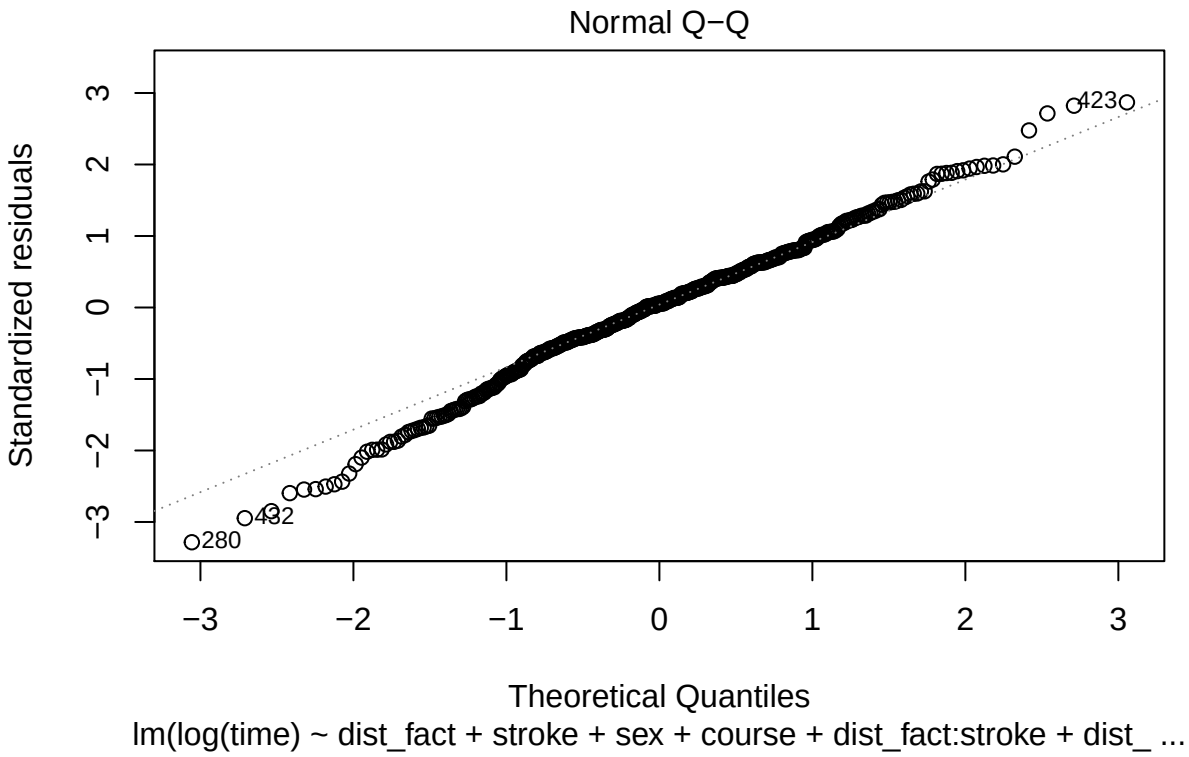
```
## Analysis of Variance Table
##
## Response: log(time)
##                    Df  Sum Sq Mean Sq    F value    Pr(>F)
## dist_fact           3 238.686  79.562 5.6909e+05 < 2.2e-16 ***
## stroke              4   2.006   0.501 3.5871e+03 < 2.2e-16 ***
## sex                 1   1.338   1.338 9.5736e+03 < 2.2e-16 ***
## course              1   0.059   0.059 4.2139e+02 < 2.2e-16 ***
## dist_fact:stroke    8   0.017   0.002 1.4832e+01 < 2.2e-16 ***
## dist_fact:sex       3   0.009   0.003 2.0546e+01 1.991e-12 ***
## stroke:sex          4   0.004   0.001 6.6151e+00 3.622e-05 ***
## stroke:course       4   0.007   0.002 1.2724e+01 8.947e-10 ***
## sex:course          1   0.004   0.004 2.8128e+01 1.855e-07 ***
## dist_fact:course    3   0.000   0.000 8.9370e-01    0.4444
## Residuals         413   0.058   0.000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

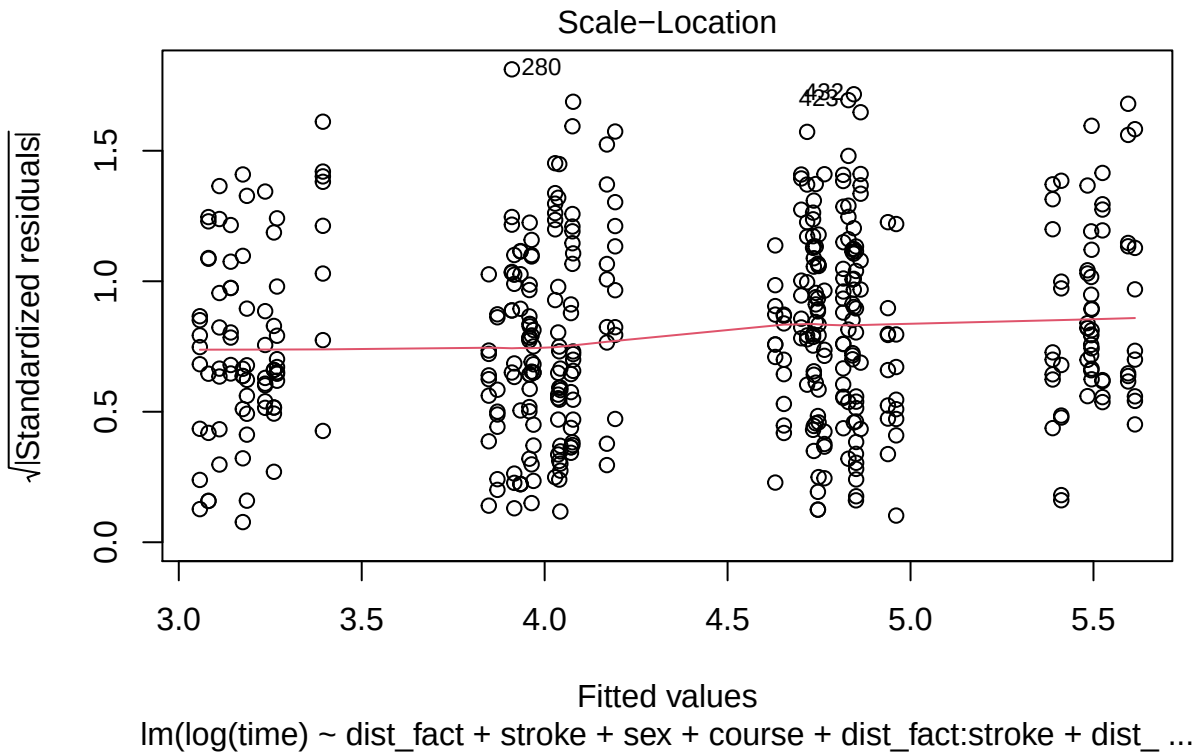
Outlier detection

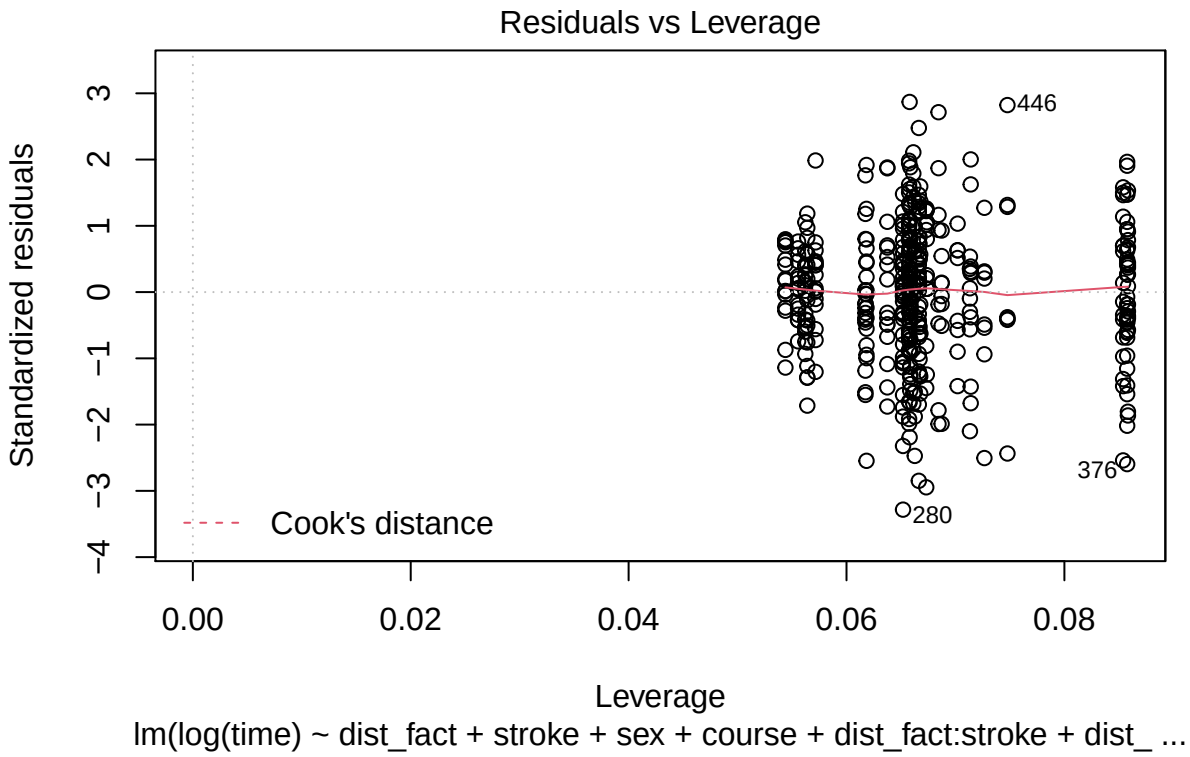
Let us look at the plot diagnostics:

```
swim_lm_selected <- lm(formula = log(time) ~ dist_fact + stroke + sex + course
                       + dist_fact:stroke + dist_fact:sex + stroke:sex
                       + stroke:course + sex:course, data = swim)
plot(swim_lm_selected)
```


Residuals vs Fitted
Residuals
0.04
0.02
0.00
−0.02
−0.04
423
280
432
3.0
3.5
4.0
4.5
5.0
5.5
Fitted values
lm(log(time) ~ dist_fact + stroke + sex + course + dist_fact:stroke + dist_ ...


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(log(time) ~ dist_fact + stroke + sex + course + dist_fact:stroke + dist_ ...
423
432
280
−3
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
280
432
423
3.0
3.5
4.0
4.5
5.0
5.5
Fitted values
lm(log(time) ~ dist_fact + stroke + sex + course + dist_fact:stroke + dist_ ...



Max leverage value is less than the value we should be concerned.

```
max(hatvalues(swim_lm_selected))
```

```
## [1] 0.08583473
```

```
2*length(swim_lm_selected$coefficients)/dim(swim)[1]
```

```
## [1] 0.1524664
```

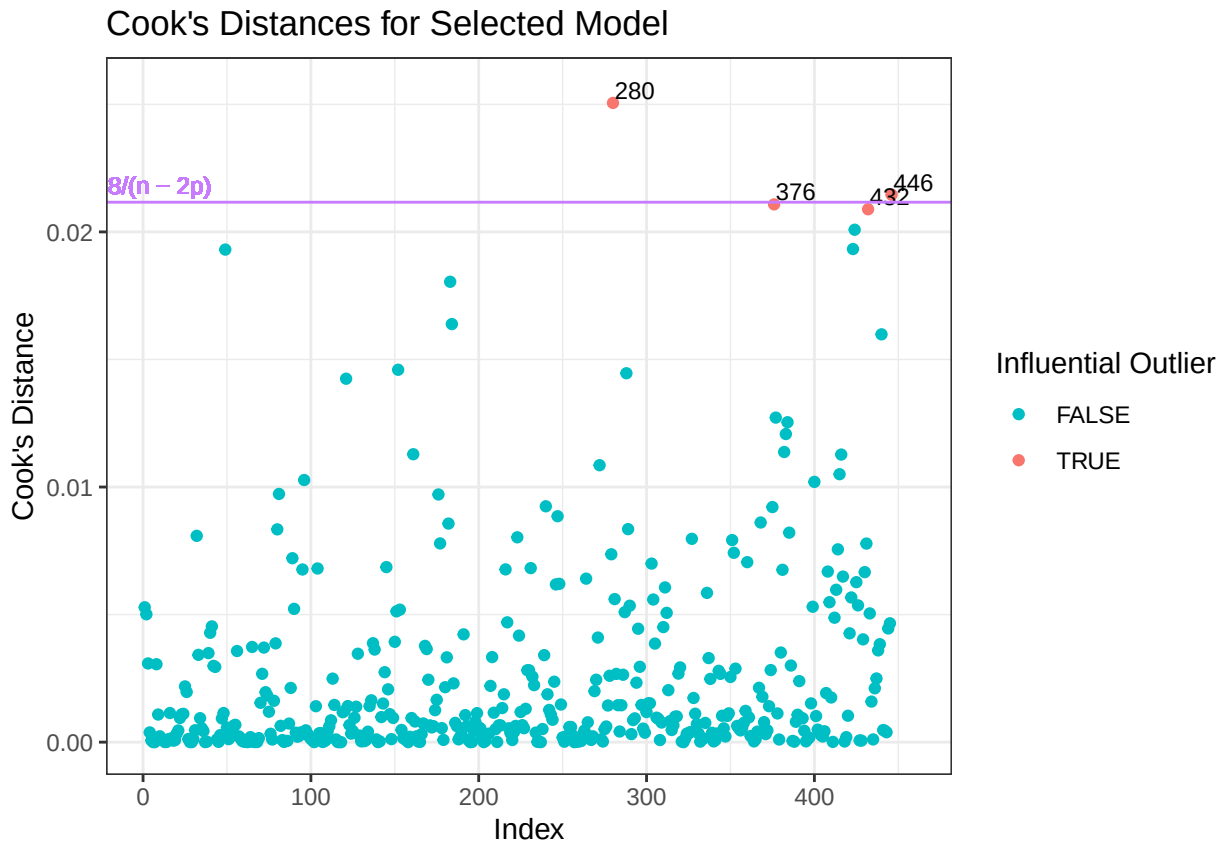
```
saved_values_numeric <- rbind(saved_values_numeric,
  data.frame(name = c("Maximum Leverage",
                      "2p/n"),
             value = c(max(hatvalues(swim_lm_selected)),
                       2*length(swim_lm_selected$coefficients)/dim(swim)[1])))
```

Cook's distances:

```
cooks_bound <-  8 / (dim(swim)[1] - 2 * length(swim_lm_selected$coefficients))

data.frame(cooks_distance = cooks.distance(swim_lm_selected), index = 1:dim(swim)[1]) %>%
  mutate(influential_outlier = cooks_distance >= cooks_bound - 0.001) %>%
  ggplot() +
  geom_point(aes(x = index, y = cooks_distance, colour = influential_outlier)) +
```

```
  geom_text(aes(x = index, y = cooks_distance,
                label = ifelse(influential_outlier,index,NA)),
            size = 3, nudge_x = 13, nudge_y = 0.0005) +
  geom_hline(aes(yintercept = cooks_bound), colour = "#C77CFF") +
  geom_text(aes( x = 10, y = cooks_bound, label = "8/(n - 2p)", vjust = -0.5),
            size = 3,colour = "#C77CFF") +
  labs(title = "Cook's Distances for Selected Model",
       x = "Index", y = "Cook's Distance") +
  theme_bw() +
  scale_colour_manual(name = "Influential Outlier", values = c("#00BFC4","#F8766D"))
```

```
## Warning: Removed 442 rows containing missing values (geom_text).
```




```
influential_outliers <- (data.frame(cooks_distance = cooks.distance(swim_lm_selected),
                                    index = 1:dim(swim)[1]) %>%
                           mutate(influential_outlier = cooks_distance >= cooks_bound) %>%
                           filter(influential_outlier == TRUE))$index

saved_values_text <- rbind(
  saved_values_text,
  data.frame(name = "Influential Outliers",
             value = paste(
               paste(influential_outliers[1:(length(influential_outliers)-1)], sep = ","),
               influential_outliers[length(influential_outliers)],
```

```
              sep = " and ")))

near_influential_outliers <- (data.frame(cooks_distance = cooks.distance(swim_lm_selected),
                                         index = 1:dim(swim)[1]) %>%
                                mutate(influential_outlier =
                                         cooks_distance <= cooks_bound
                                       & cooks_distance >= cooks_bound - 0.001) %>%
                                filter(influential_outlier == TRUE))$index

saved_values_text <- rbind(
  saved_values_text,
  data.frame(name = "Near Influential Outliers",
             value = paste(
               paste(near_influential_outliers[1:(length(near_influential_outliers)-1)],
                     sep = ", "),
               influential_outliers[length(influential_outliers)],
               sep = " and ")))
```

Examining points:

```
influential_outlier_points <- swim %>%
  slice(c(influential_outliers,near_influential_outliers))

rbind(mutate(influential_outlier_points, Series = "influential_outliers"),
      mutate(swim, Series = "data_points")) %>%
  ggplot() +
  geom_point(aes(x = dist, y = time, colour = Series,size = Series)) +
  theme_bw() +
  labs(title = "Plot of Distance vs Time Highlighting Influential Outliers",
       x = "Distance", y = "Time") +
  scale_colour_discrete(labels = c("Data points","Potential influential outliers")) +
  guides(size = FALSE)
```

```
## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =
## "none")` instead.
```

```
## Warning: Using size for a discrete variable is not advised.
```

Model Interpretation

```
coefficients <- as.data.frame(summary(swim_lm_selected)$coefficients)
coefficients <- signif(coefficients[,1:dim(coefficients)[2]],digits = 3)
coefficients$term <- rownames(coefficients)
coefficients <- coefficients %>%
  mutate(term = str_replace(term,"dist_fact","dist")) %>%
  mutate(term = str_replace(term, ":", "*")) %>%
  mutate(significance = ifelse(`Pr(>|t|)`<0.001,
                               "***",
                               ifelse(`Pr(>|t|)`<0.01,
                                      "**",
                                      ifelse(`Pr(>|t|)`<0.05,
                                             "*",
                                             ifelse(`Pr(>|t|)`<0.1,
                                                    ".",
                                                    "")))))
colnames(coefficients) <- str_replace(colnames(coefficients), " ", "_")
coefficients
```

```
##                              Estimate Std._Error t_value Pr(>|t|)
## (Intercept)                   3.30000    0.00465 710.000 0.00e+00
## dist_fact100                  0.76800    0.00426 180.000 0.00e+00
```

```
## dist_fact200                      1.55000    0.00426 363.000 0.00e+00
## dist_fact400                      2.30000    0.00541 425.000 0.00e+00
## strokeBreaststroke                0.11300    0.00620  18.200 5.97e-55
## strokeButterfly                  -0.05620    0.00622  -9.040 6.10e-18
## strokeFreestyle                  -0.11700    0.00510 -23.000 4.03e-76
## strokeMedley                      0.01270    0.00415   3.070 2.32e-03
## sexM                             -0.11400    0.00409 -27.900 2.67e-97
## courseShort                      -0.03490    0.00318 -11.000 7.48e-25
## dist_fact100:strokeBreaststroke   0.00782    0.00553   1.420 1.58e-01
## dist_fact200:strokeBreaststroke -0.00303    0.00553  -0.549 5.83e-01
## dist_fact100:strokeButterfly      0.02420    0.00553   4.380 1.52e-05
## dist_fact200:strokeButterfly      0.04680    0.00553   8.470 4.38e-16
## dist_fact100:strokeFreestyle      0.01480    0.00490   3.020 2.70e-03
## dist_fact200:strokeFreestyle      0.01330    0.00490   2.720 6.85e-03
## dist_fact400:strokeFreestyle      0.01150    0.00573   2.010 4.47e-02
## dist_fact100:strokeMedley         0.00989    0.00490   2.020 4.43e-02
## dist_fact100:sexM                 0.00654    0.00340   1.930 5.47e-02
## dist_fact200:sexM                 0.01200    0.00340   3.520 4.75e-04
## dist_fact400:sexM                 0.02220    0.00444   5.000 8.67e-07
## strokeBreaststroke:sexM          -0.00725    0.00374  -1.940 5.30e-02
## strokeButterfly:sexM              0.00166    0.00375   0.442 6.59e-01
## strokeFreestyle:sexM              0.00907    0.00351   2.580 1.01e-02
## strokeMedley:sexM                 0.00321    0.00405   0.794 4.28e-01
## strokeBreaststroke:courseShort    0.01240    0.00418   2.970 3.20e-03
## strokeButterfly:courseShort       0.02370    0.00420   5.660 2.87e-08
## strokeFreestyle:courseShort       0.02380    0.00362   6.570 1.52e-10
## strokeMedley:courseShort          0.01640    0.00420   3.900 1.13e-04
## sexM:courseShort                 -0.01240    0.00234  -5.310 1.83e-07
##                                                          term significance
## (Intercept)                                       (Intercept)          ***
## dist_fact100                                          dist100          ***
## dist_fact200                                          dist200          ***
## dist_fact400                                          dist400          ***
## strokeBreaststroke                          strokeBreaststroke          ***
## strokeButterfly                                strokeButterfly          ***
## strokeFreestyle                                strokeFreestyle          ***
## strokeMedley                                      strokeMedley           **
## sexM                                                     sexM          ***
## courseShort                                       courseShort          ***
## dist_fact100:strokeBreaststroke     dist100*strokeBreaststroke
## dist_fact200:strokeBreaststroke     dist200*strokeBreaststroke
## dist_fact100:strokeButterfly           dist100*strokeButterfly          ***
## dist_fact200:strokeButterfly           dist200*strokeButterfly          ***
## dist_fact100:strokeFreestyle           dist100*strokeFreestyle           **
## dist_fact200:strokeFreestyle           dist200*strokeFreestyle           **
## dist_fact400:strokeFreestyle           dist400*strokeFreestyle            *
## dist_fact100:strokeMedley                 dist100*strokeMedley            *
## dist_fact100:sexM                                dist100*sexM            .
## dist_fact200:sexM                                dist200*sexM          ***
## dist_fact400:sexM                                dist400*sexM          ***
## strokeBreaststroke:sexM                strokeBreaststroke*sexM            .
## strokeButterfly:sexM                      strokeButterfly*sexM
## strokeFreestyle:sexM                      strokeFreestyle*sexM            *
## strokeMedley:sexM                            strokeMedley*sexM
```

```
## strokeBreaststroke:courseShort  strokeBreaststroke*courseShort           **
## strokeButterfly:courseShort        strokeButterfly*courseShort          ***
## strokeFreestyle:courseShort        strokeFreestyle*courseShort          ***
## strokeMedley:courseShort              strokeMedley*courseShort          ***
## sexM:courseShort                             sexM*courseShort          ***
```

```
write.csv(coefficients,row.names = F,file = "coefficients.csv", quote = F)
```

Transformed coefficients:

```
coefficients_transformed <- coefficients %>%
  dplyr::select(term,Estimate) %>%
  mutate(transformed_estimate = signif(exp(Estimate),digits = 4))


write.csv(coefficients_transformed,row.names = F,
          file = "coefficients_transformed.csv", quote = F)
```

Prediction

Loading predictors

```
write("name dist stroke sex course
RaceA 400 Freestyle F Long
RaceB 50 Backstroke F Long
RaceC 400 Butterfly F Long
RaceD 100 Medley F Long","predictors")
predictors <- read.table("predictors", header = T)
predictors <- predictors %>%
  mutate(sex = ifelse(sex,"M","F"),
         dist_fact = as.factor(dist))
predictors
```

```
##    name dist     stroke sex course dist_fact
## 1 RaceA  400  Freestyle   F   Long       400
## 2 RaceB   50 Backstroke   F   Long        50
## 3 RaceC  400  Butterfly   F   Long       400
## 4 RaceD  100     Medley   F   Long       100
```

Predictions:

```
predict.lm(swim_lm_selected,predictors,interval = "prediction")
```

```
## Warning in predict.lm(swim_lm_selected, predictors, interval = "prediction"):
## prediction from a rank-deficient fit may be misleading
```

```
##        fit      lwr      upr
## 1 5.494929 5.470988 5.518870
## 2 3.303498 3.278529 3.328466
## 3 5.544425 5.517536 5.571315
## 4 4.094550 4.069619 4.119481
```

```
predictions <- data.frame(predictors,predict.lm(swim_lm_selected,predictors,
                                                interval = "prediction"))
```

```
## Warning in predict.lm(swim_lm_selected, predictors, interval = "prediction"):
## prediction from a rank-deficient fit may be misleading
```

```
predictions <- predictions %>%
  mutate(across(7:9, exp, .names = "exp_{.col}")) %>%
  mutate(across(7:12,signif,digits = 4))
write.csv(predictions,"predictions.csv", quote = F)
```

Misc

```
saved_values_numeric <- saved_values_numeric %>%
  mutate(name = str_replace_all(name," ","_"),
         value = signif(value, digits = 3))

saved_values_text <- saved_values_text %>%
  mutate(name = str_replace_all(name," ","_"))


write.table(saved_values_numeric, file = "saved_values_numeric.txt",
            row.names = FALSE, quote = F)
write.csv(saved_values_text, file = "saved_values_text.txt",
          row.names = FALSE, quote = F)
```