Marked Practical: Linear Models

MSc in Statistical Science

P699

11/16/2019

Contents

1 Introduction

The dataset `swim` is a 446 observations of swimming time in competitions along with other variables as the distance, stroke, sex and length of the course.

We will explore the data, find a model with explanatory variables, remove outliers, interpret the model and predict the time for some new observations.

2 Data Exploration

We first notice that the variable `dist` is stored as a numerical variable even though it only takes a few fixed values. Let's start by looking at the distribution of time with respect to the other variables:

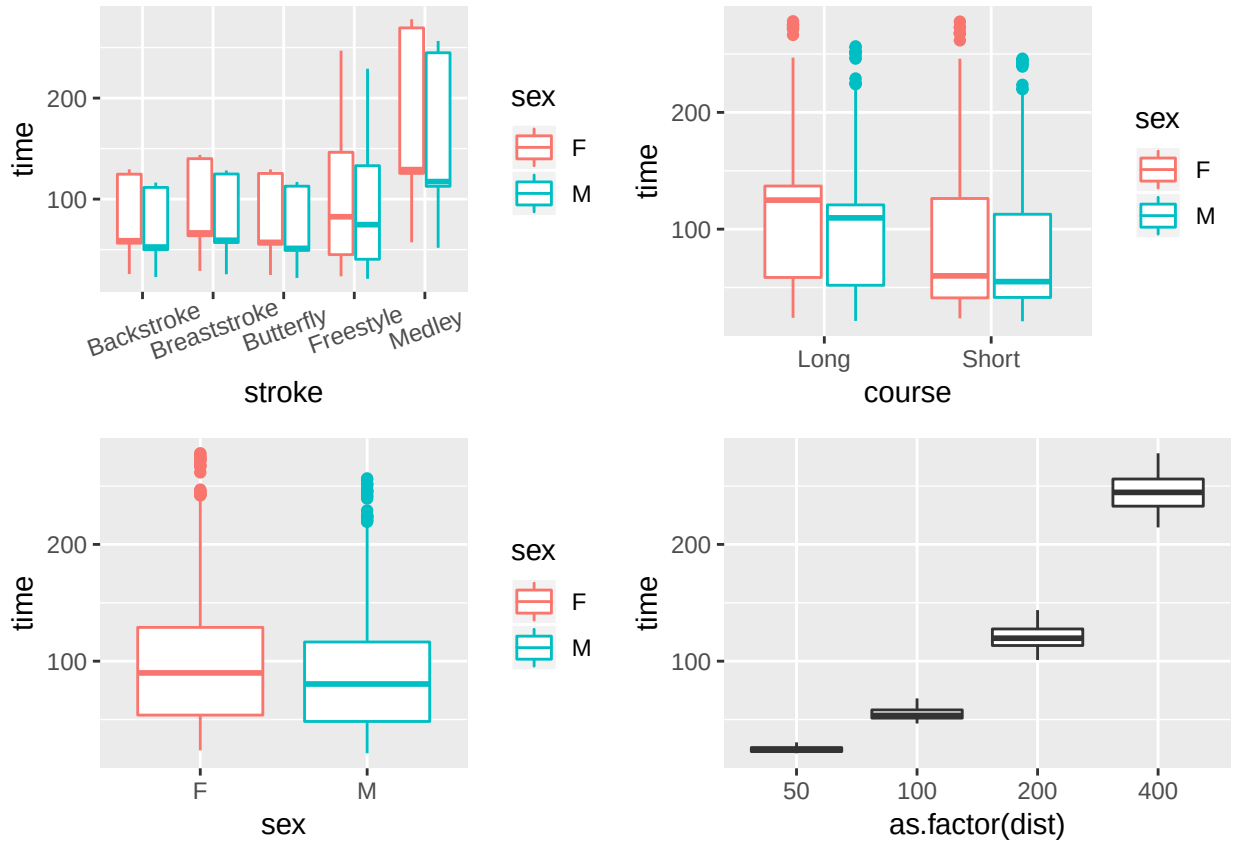


Figure 1: ScatterPlots of time against other variables

We clearly see a higher performance of men in all kind of events. Also, the `Medley` swim requests more energy, hence the higher time and slightly higher variance.

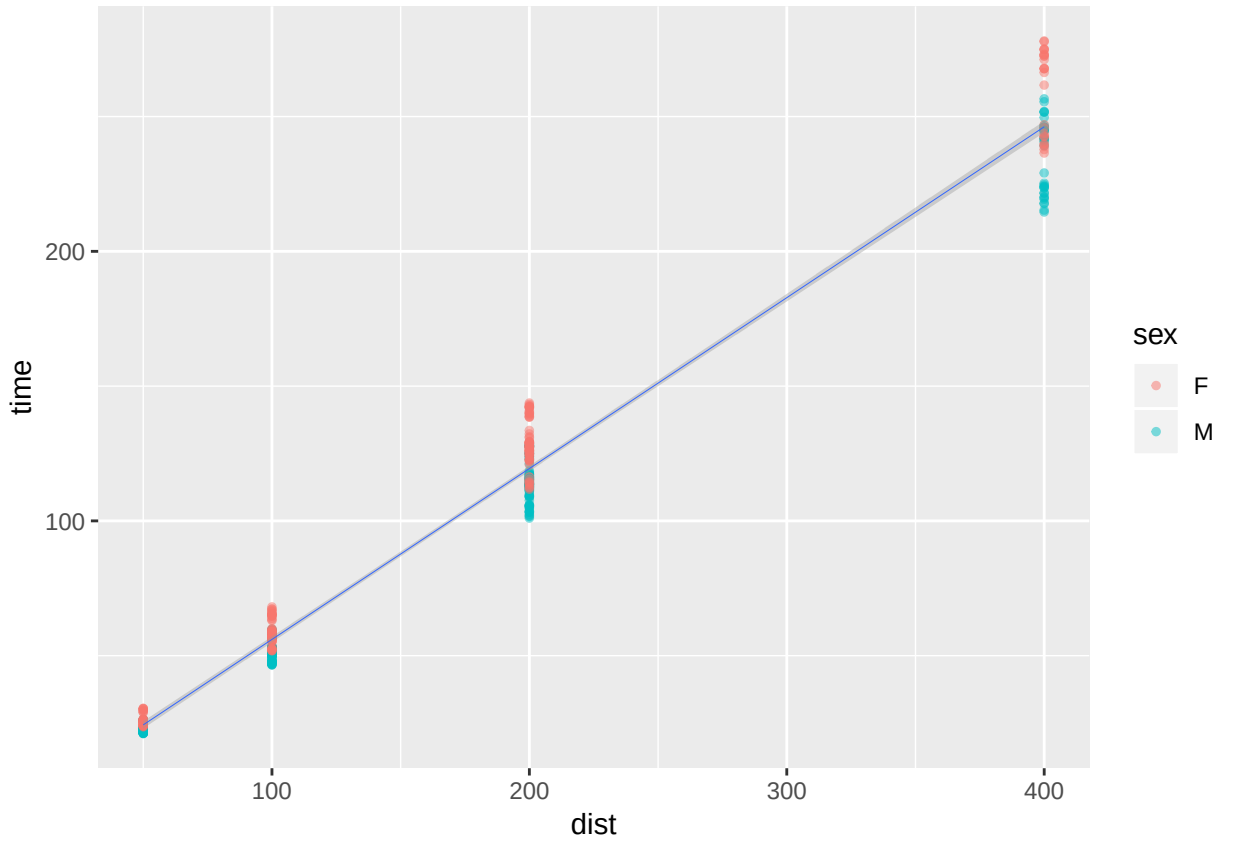
We should also notice that the standard deviation of `time` clearly increases with `dist`:

50	100	200	400
2.447	5.235	10.38	18.31

It seems that it is proportional to `dist`. We will examine this in the next sections. We can also notice that the mean is also proportional to 'dist':

50	100	200	400
24.43	55.03	120.7	245.2

The exponential trend in the plot of `time` against `as.factor(dist)` is only a manifestation of the transformation to factor (being aware that time $= 50 \times 2^k, k = 0..3$). In the next plot, there is a clear linear dependance between 'time' and 'dist'.



There is no easy way to check that the data is normal because there are too many discrete variables which need to be fixed to observe the Gaussian shape. For instance, if we plot `time` without fixing `sex` we see 2 Gaussian-like shapes corresponding to "Male" and "Female".

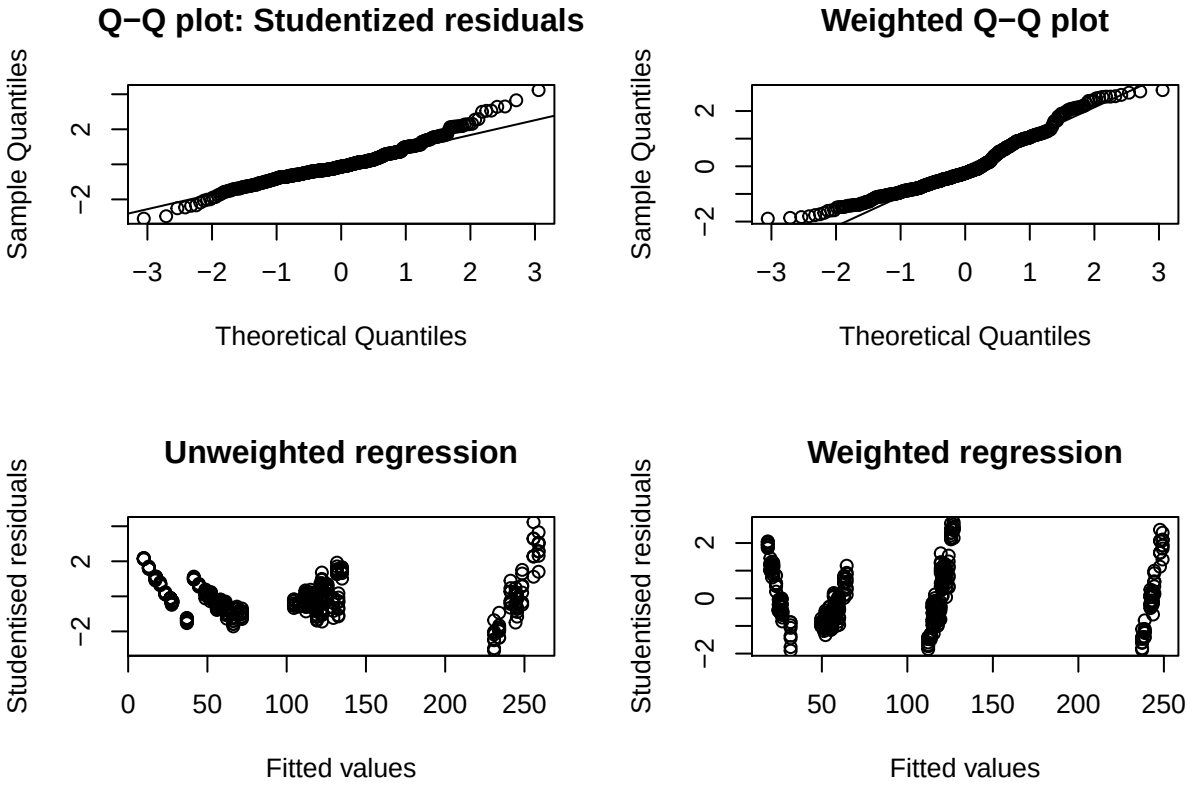
3 Model Selection

Further plots show that the dependance of the variance of `time` with respect to `dist` is a problematic trend if we wish to restrict ourselves to simple linear models.

Weighted Regression

In this subsection, we will take into account the variability of the variance of `time`. Given that $Var(time) \sim dist^2$, we can easily construct the vector of weights such that $\sigma_i^2 = \sigma^2/w_i$.

Let's compare the weighted regression with the unweighted regression:



We see that the studentised residuals have a better behaviour as there is no tren of the variance with respect to the value of time. The normality of the model is also slightly improved.

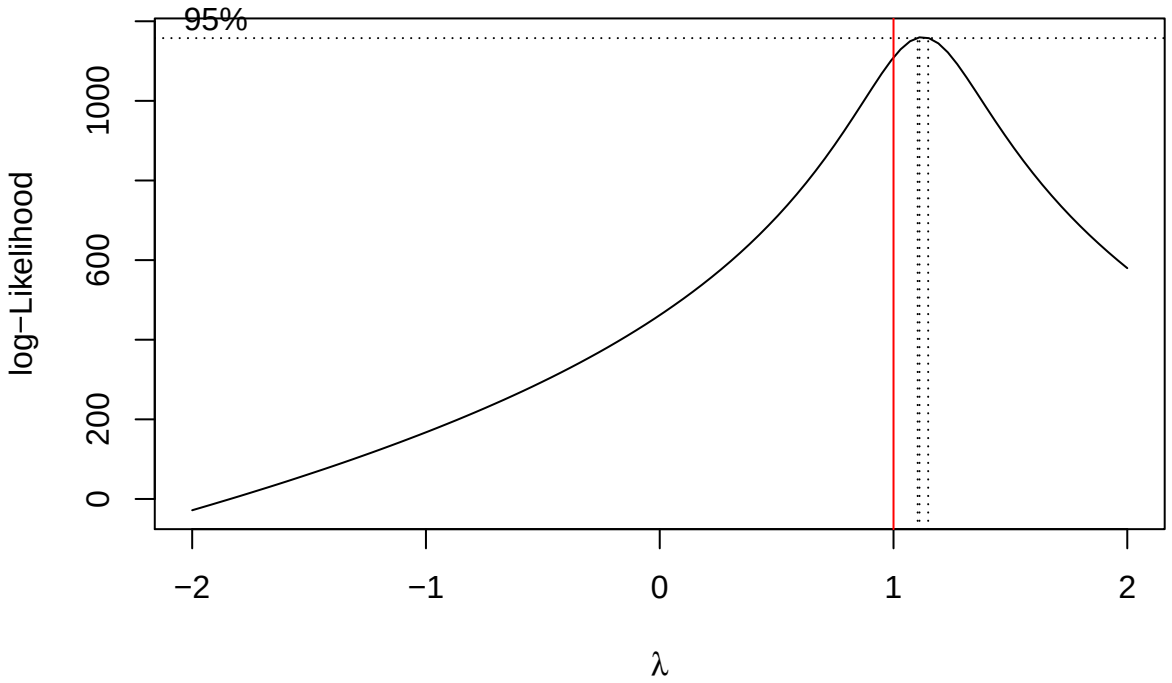
When we compare the models using `summary()`, we see that all the β_is become significant unlike in the unweighted case. Nevertheless, the R-Squared and the F-statistic decrease a little in the weighted regression, respectively from 0.9942 to 0.9925 and from 10,770 to 8,250.

The model doesn't seem this convincing. This may be because the standard deviation of `time` really mainly depends on `dist` (as we showed, they are proportional).

Multiplicative model

In this case, it might be more natural to have a model of the form $time = dist \times (e^{\sum other}) \times (randomness)$. In fact, the randomness here will be lognormal and to model this we are going to fit a linear model of the following type : `log(time) ~ log(dist)+stroke+sex+course`.

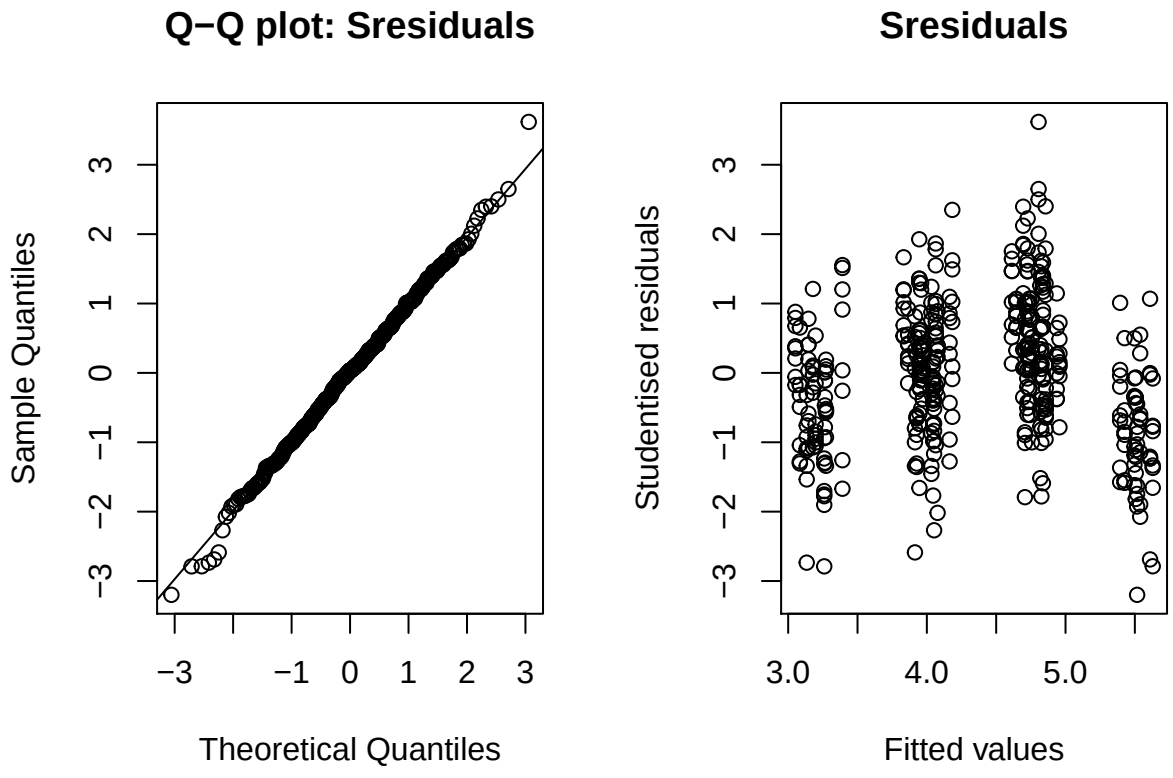
First, let us confirm that the dependance will be linear in this case, to do so, we maximise the likelihood of the Box-Cox family using the model: `log(time) ~ log(dist)+stroke+sex+course`.



We see that the right dependence is linear. In fact the multiplicative model is far better that the last two models as all the variables are now significant with mostly $p_{val} \approx 10^{-16}$ and the R-Squared and F-statistic are now far better (0.9995 and 1.191e+05 respectively).

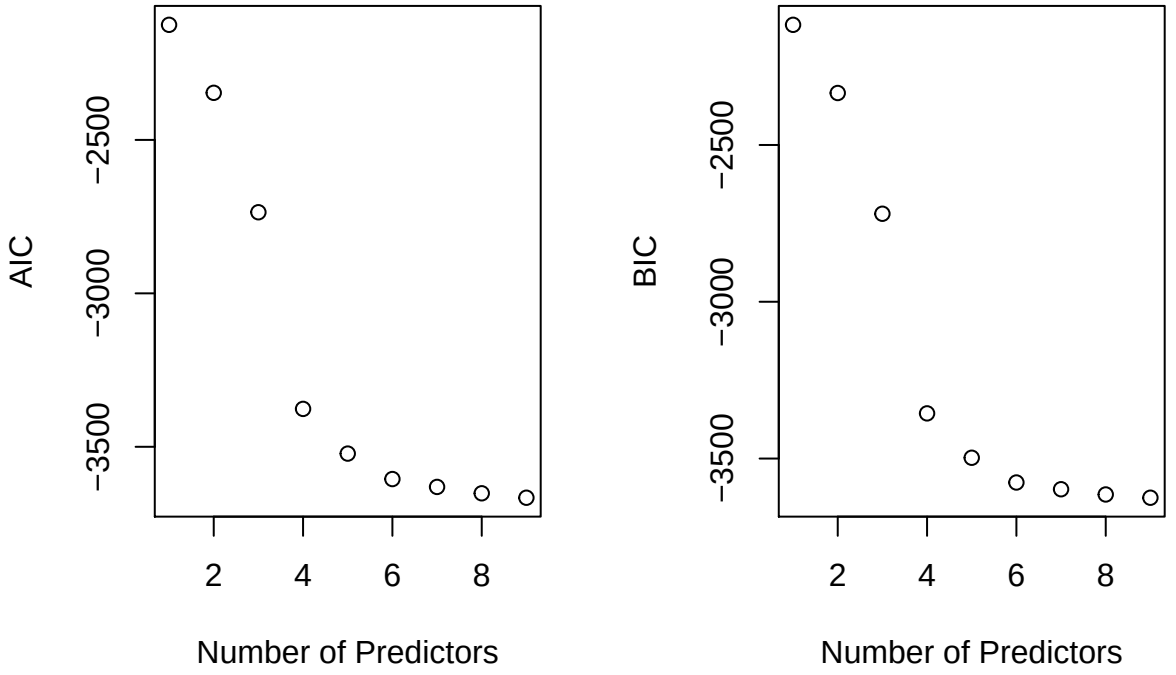
We can now try to explore the interaction between the variables, after some trials we conclude that the significant interactions are `sex:log(dist)` and `sex:course`.

We now have a good model with even better summary values and all variables significant. Below, we check that the model is linear and that the residuals show no trend.



Variables selection

We can now use the AIC and BIC to check that the best model contains all the included variables:



Indeed, the best model seems to be the full

`log(time) ~ log(dist)+course+sex+stroke+(log(dist)+course)*sex`.

A further test is computed using the ANOVA:

Table 3: Anova

	Df	Sum.Sq	Mean.Sq	F.value	Pr..F.
log(dist)	1	238.4	238.4	905093	0
course	1	0.02871	0.02871	109	6.34e-23
sex	1	1.338	1.338	5081	2.085e-242
stroke	4	2.274	0.5686	2159	9.209e-286
log(dist):sex	1	0.007988	0.007988	30.32	6.253e-08
course:sex	1	0.004307	0.004307	16.35	6.223e-05
Residuals	436	0.1149	0.0002634	NA	NA

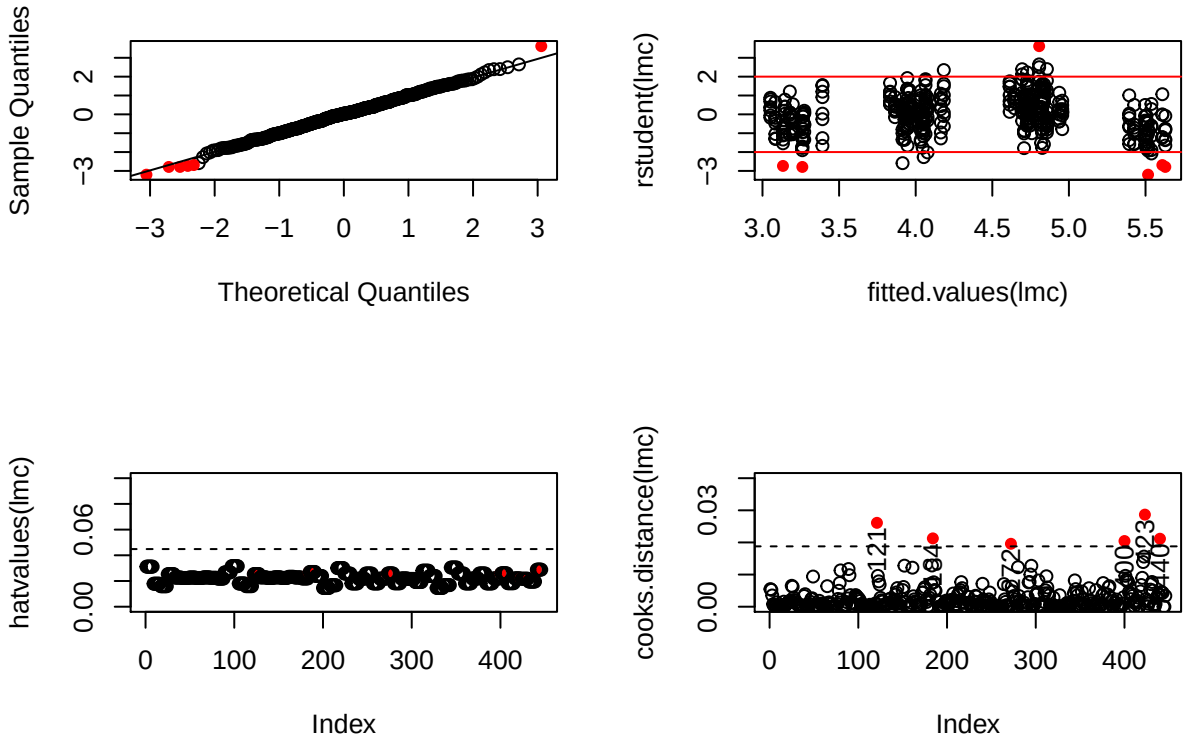
The dependance of `sex` with `course` and `dist` is one of the reasons that some competitions are divided by sex. Now that we have fixed the model with the right parameters, we can move on to the outlier analysis.

Outlier Analysis

We now look for outliers, we plot in red obesrvations with Cook's distance higher than $8/(n-2p)$, n being the number of observations and p the number of explanatory variables.

We notice that the outliers are in the extreme tail of the gaussian in the Q-Q plot, they are also further from the theoretical quantiles of the Gaussian.

From the Studentised residuals plot, they all have a high misfit, but we see that none of them has a high leverage (higher than $2p/n$). This is enough to make of them outliers:



We now delete those outliers, update the dataset and the (n, p). When we look for outliers again, we find that there is none.

The new model is even better than the previous ones, we have an R-Squared of 0.9996.

4 Interpretation

We thought about a model equivalent to $time = dist \times (e^{\sum other}) \times (randomness)$ with $randomness$ being a lognormal, $\mathbb{E}(randomness) = 1$. In these settings: $\sigma_{time} \sim dist \times \sigma_{randomness}$, which reproduces well our first findings, physically, it is also quite natural as we know that $t = \frac{1}{V} \times d$, the more the distance is large the more dispersion we will have in the scores.

In fact, $\beta_{log(dist)} \approx 1.1$ so we have a model of the type $time = dist^{1.1} \times \dots$ which is explained by the fact that the speed V is not constant as swimmers tend to get tired more when they swim larger distances hence the decrease of speed, this is well modelled by the power 1.1: if we double the distance, the time is multiplied by 2.2. Finally, we can notice that this dependance is equivalent to $time^{0.9} \sim dist \times \dots$ which is goes along with the value returned by the Box-Cox likelihood maximisation ≈ 0.9.

The other variables appear as multiplicative factors of the type e^{β_i}, as we can expect these are very small so that a change in any of the categories "Stroke" or "Course" doesn't change the time a lot. In fact, the largest β_i of all is β_{sexM} which suggets a lower time for men. Let's order the strokes from the fastest to the slowest:

	Freestyle	Butterfly	Backstroke	Medley	Breaststroke
exp_beta	0.9105	0.9879	1	1.02	1.126

`Backstroke` being the base stroke, the way we interpret it is: in the same conditions, a time of 1 using

Backstroke will be done in 0.9105 using Freestyle ...

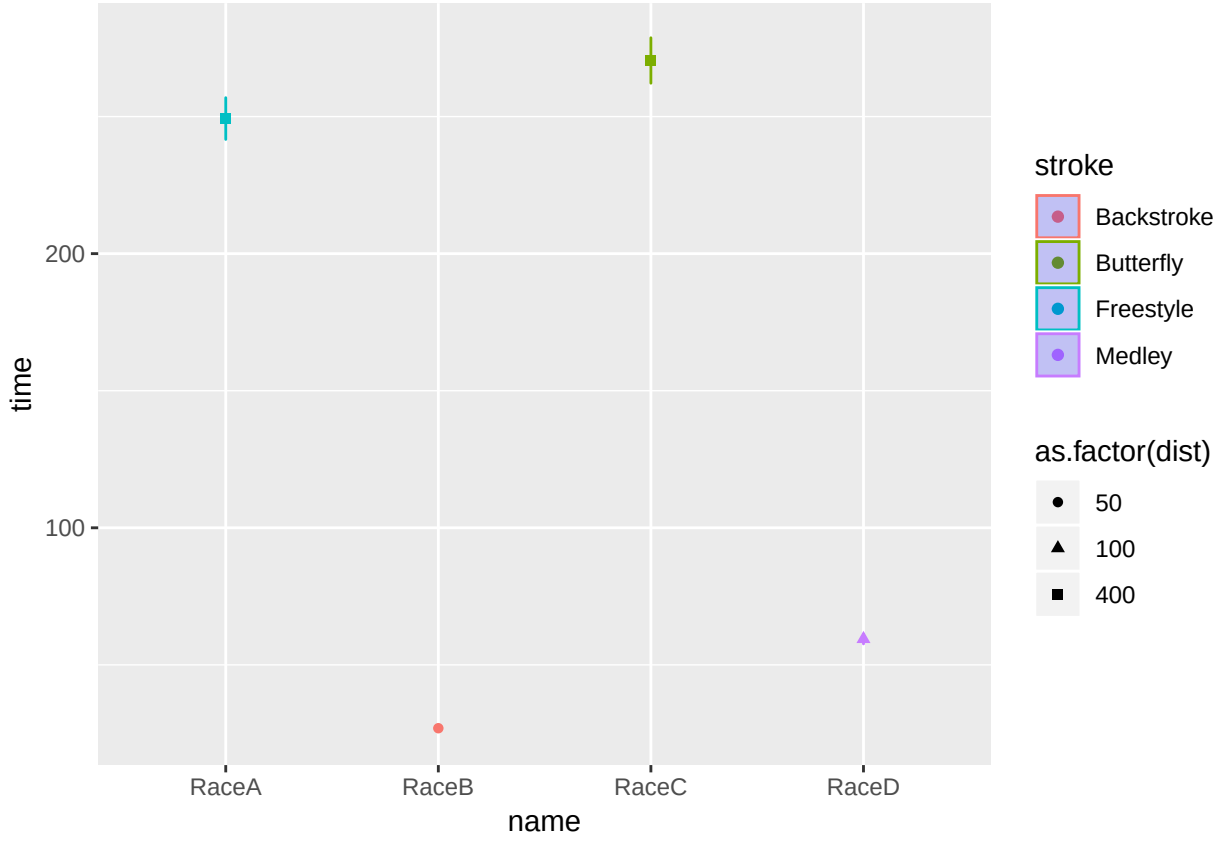
We also find that for the same distance, we have a better score in short courses, maybe because we use the wall to increase the speed and save energy each 25m instead of each 50m ...

The smallest β_is are those of the interactions (`sex:log(dist)` and `sex:course`). $\beta_{courseShort:sexM} \approx -10^{-2}$ which means that men have a higher increase of time when we move from "short" to "long" courses than women. Similarly, in these high performances, men have higher rate of increase of time when we increase distance that women as $\beta_{log(dist):sexM} \approx +10^{-2}$.

5 Prediction

Here is a plot showing the estimated times and their prediction intervals for each new observation:

```
## [1] "Backstroke"   "Breaststroke" "Butterfly"    "Freestyle"
## [5] "Medley"
```



It is clearer here that the length of the prediction interval is of the type $length \approx length_0 \times 2^k$ the same as `dist`:

name	dist	stroke	sex	course	lwr	time	upr
RaceA	400	Freestyle	F	Long	241.7	249.2	256.9
RaceB	50	Backstroke	F	Long	26.08	26.89	27.73
RaceC	400	Butterfly	F	Long	262.2	270.3	278.7
RaceD	100	Medley	F	Long	57.65	59.43	61.28

For RaceA, we have 7 observations in our dataset and 6 of them fall into the 95% interval of prediction which is not so bad.

6 Conclusions

Finally, the model we obtain considers the tiredness of the swimmers (i.e. the speed is not constant) via the dependance $\sim dist^{1.1}$, this also takes into account the proportionality of the dispertion of scores with respect to the distance. The model includes all the explanatory variables and the interaction of the `sex` with the distance and the course.

R Code

```
# Loading the Data

library(ggplot2)
library(gridExtra)
library(ggExtra)
library(knitr)
library(MASS)
library(leaps)
library(pander)
swim = read.csv("swim.csv")
summary(swim)
head(swim)
str(swim)
n = dim(swim)[1]

# Exploration of data

g = ggplot(swim, aes(y = time))
g1 = g + geom_boxplot(aes(x = stroke, colour = sex)) + theme(axis.text.x = element_text(angle=20, vjust=
g2 = g + geom_boxplot(aes(x = course, colour = sex))
g3 = g + geom_boxplot(aes(x = sex, colour = sex))
g4 = g + geom_boxplot(aes(x = as.factor(dist)))
grid.arrange(g1, g2,g3,g4, nrow=2, ncol=2)

#ggplot(swim, aes(stroke, time, colour = stroke)) + geom_violin(scale = "area") +  facet_grid(cols = va

pander(with(swim, tapply(time, dist, sd)))
pander(with(swim, tapply(time, dist, mean)))

ggplot(swim, aes(x = dist, y = time)) + geom_point(aes(colour = sex), size = 1, alpha = .5) + geom_smoo

#Model selection
#Weighted Regression

w = (50/swim$dist)**2
#cbind(w,swim$dist)
lm = lm(time ~ +dist+stroke+sex+course , data = swim)
lmw = lm(time ~ +dist+stroke+sex+course , data = swim, weights = w)
```

```r
#plot(lm)
#plot(lmw)
par(mfrow = c(2, 2))
qqnorm(rstudent(lm), main = "Q-Q plot: Studentized residuals")
qqline(rstudent(lm))

qqnorm(rstudent(lmw), main = "Weighted Q-Q plot")
qqline(rstudent(lmw))

plot(rstudent(lm) ~ fitted(lm), main = "Unweighted regression",
     xlab = "Fitted values", ylab = "Studentised residuals")
plot(rstudent(lmw) ~ fitted(lmw), main = "Weighted regression",
     xlab = "Fitted values", ylab = "Studentised residuals")

summary(lm)
summary(lmw)

##Multiplicative model
# Box-Cox
putts.bc <- boxcox(log(time) ~ log(dist)+stroke+sex+course, data = swim)
abline(v = 1, col = "red")

summary(lm(log(time) ~ log(dist)+stroke+sex+course, data = swim))

lmc = lm(log(time) ~ log(dist)+course+sex+stroke+(log(dist)+course)*sex, data = swim)

#plot(lmc)
par(mfrow=c(1,2))
qqnorm(rstudent(lmc), main = "Q-Q plot: Sresiduals")
qqline(rstudent(lmc))

plot(rstudent(lmc) ~ fitted(lmc), main = "Sresiduals",
     xlab = "Fitted values", ylab = "Studentised residuals")

summary(lmc)

## Variables selection

b <- regsubsets(log(time) ~ log(dist)+course+sex+stroke+(log(dist)+course)*sex,
                data = swim, nvmax = 10)
rs <- summary(b)
AIC <- n*log(rs$rss/n) + (2:10)*2
BIC <- n*log(rs$rss/n) + (2:10)*log(n)
par(mfrow=c(1,2))
plot(AIC ~ I(1:9), ylab = "AIC", xlab = "Number of Predictors")
plot(BIC ~ I(1:9), ylab = "BIC", xlab = "Number of Predictors")
df<-data.frame(anova(lmc))

#options(knitr.kable.NA = '')
#kable(df, col.names = c("Residual df", "Residual SS","Df", "Deviance","F","P value"),
#      digits = 10)
pander(df, caption = "Anova")
```

```r
## Outlier Analysis

p = lmc$rank
i <- cooks.distance(lmc) > (8/(n - 2*p))

num_outliers = sum(i)

#pairs(subset(swim, select = -event),lower.panel = NULL, pch = 1 + 9*i, col = 1 + i)
#plot(lmc)

par(mfrow = c(2, 2))

qqnorm(rstudent(lmc), main = NULL, pch = 1 + 15*i, col = 1 + i)
qqline(rstudent(lmc))

plot(fitted.values(lmc), rstudent(lmc), pch = 1 + 15*i, col = 1 + i)
abline(h = c(-2,2), col = "red")

plot(hatvalues(lmc), ylim = c(0, 0.1), pch = 1 + 15*i, col = 1 + i)
abline(2*p/n, 0, lty = 2)

plt = plot(cooks.distance(lmc), ylim = c(0, 0.04), pch = 1 + 15*i, col = 1 + i)
text((1:n)[i], cooks.distance(lmc)[i], row.names(swim)[i], srt = 90, adj = 1.1)
abline(8/(n - 2*p), 0, lty = 2)

swim2 = swim[-which(i),]
lmc2 = lm(log(time) ~ (log(dist)+course)*sex+stroke+course, data = swim2)
n2 = n - num_outliers
i2 <- cooks.distance(lmc2) > (8/(n2 - 2*p))
num_outliers2 = sum(i2)

#pairs(subset(swim2, select = -event),lower.panel = NULL, pch = 1 + 9*i2, col = 1 + i2)
#plot(lmc2)

par(mfrow = c(2, 2))

qqnorm(rstudent(lmc2), main = NULL, pch = 1 + 15*i2, col = 1 + i2)
qqline(rstudent(lmc2))

plot(fitted.values(lmc2), rstudent(lmc2), pch = 1 + 15*i2, col = 1 + i2)
abline(h = c(-2,2), col = "red")

plot(hatvalues(lmc2), ylim = c(0, 0.1), pch = 1 + 15*i2, col = 1 + i2)
abline(2*p/n2, 0, lty = 2)

plot(cooks.distance(lmc2), ylim = c(0, 0.04), pch = 1 + 15*i2, col = 1 + i2)
abline(8/(n2 - 2*p), 0, lty = 2)

summary(lmc2)

#Interpretation

d1 = c( "Freestyle" , "Butterfly" , "Backstroke" , "Medley" , "Breaststroke" )
```

```r
d2 = exp(c( -0.093726 , -0.012176 , 0 , 0.019691 , 0.118649 ))
dat = as.data.frame(t(data.frame("exp_beta"=d2)), header = TRUE)
colnames(dat) = d1
pander(dat)

#5 Prediction

levels(swim$stroke)
namen = c("RaceA","RaceB","RaceC","RaceD")
distn = c(400,50,400,100)
stroken = c("Freestyle","Backstroke","Butterfly","Medley")
sexn = c("F","F","F","F")
coursen = c("Long","Long","Long","Long")
newframe = data.frame("name"=namen, "dist"=distn, "stroke"=stroken, "sex"=sexn,
                      "course"=coursen)
pred_intervals = exp(predict(lmc2, newframe, interval="prediction", level = 0.95))
newframe$lwr = pred_intervals[,"lwr"]
newframe$time = pred_intervals[,"fit"]
newframe$upr = pred_intervals[,"upr"]
ggplot(newframe, aes(x=name, colour = stroke, shape = as.factor(dist))) +
  geom_point(aes(y=time)) +
  geom_ribbon(aes(ymin = lwr, ymax = upr), fill = "blue", alpha = 0.2)
pander(newframe)
```