Practical for Michaelmas Term Week 4

P576

November 7 2020

1 Exploratory Analysis

The data contains the times from the swimming races of the 2016 Olympics and the 2016 World Championship. Upon a small exploration, we can see that the data has 6 variables:

1. **event:** is composed of 16 factors namely, 50,100,200,400 metre races for Freestyle; 50, 100,200 metre races for back and breast strokes, butterfly, and medley. A bar chart is provided below to show the distribution of different types of events. The event has no additional information on top of the *dist* and *stroke* columns. So, we do not use this variable in the further analysis as it causes multicollinearity in the linear regression models.

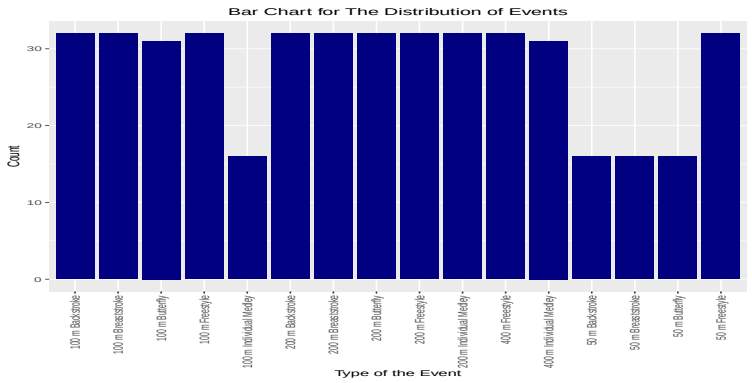

Figure 1: Plots for Time

2. **dist:** is an integer-valued variable. However, we convert the integer values into of 4 factors , namely, 50,100, 200,400 metres to summarize. We can see that there 80 observations with 50 metres, 143 with 100, 160 with 200, and 63 with 400 metres.

3. **stroke:** is composed of 5 factors namely, Freestyle, Backstroke, Breaststroke, Butterfly, and Medley. There are 80 backstroke, and breaststroke competitions, 79 butterfly and medley competitions, and 128 freestyle competitions.

4. **sex:** is composed of 2 factors M and F. There are 222 females, and 224 males in the sample.

5. **course:** is composed of 2 factors namely short and long. There are 191 long, and 255 short courses in the sample.

6. **time:** is a continuous variable and the unit is seconds. Minimum value is 21.1, 25% percentile is 50.81, median is 84.56, mean is 99.95, 75% percentile is 126.81, and the maximum value is 278.06 seconds. From the relatively large discrepancy between the median and the mean, we can suspect that there are outliers on the upper values.

As can be understood from the summary and the Figure 1, the variables are relatively equally distributed across different categories. When it comes to the only continuous variable time in the data set, a box plot and a scatter plot summarizing the variable look like:

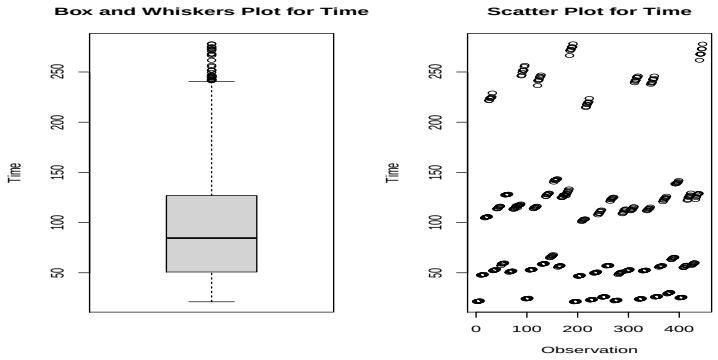

Figure 2: Plots for Time variable

Plots for Time variable already lay bare some information about the variable. There are many observations that fall over the 75% percentile. This allows us to visualize and clarify our suspicion about many outliers in time. From the scatter plot, we can guess that the data has levels of magnitude where observations tend to clutter. This will be clearer with Figure 4.

When we plot the variables in pairs, we see that it is not very informative as all of the variables except one is categorical:

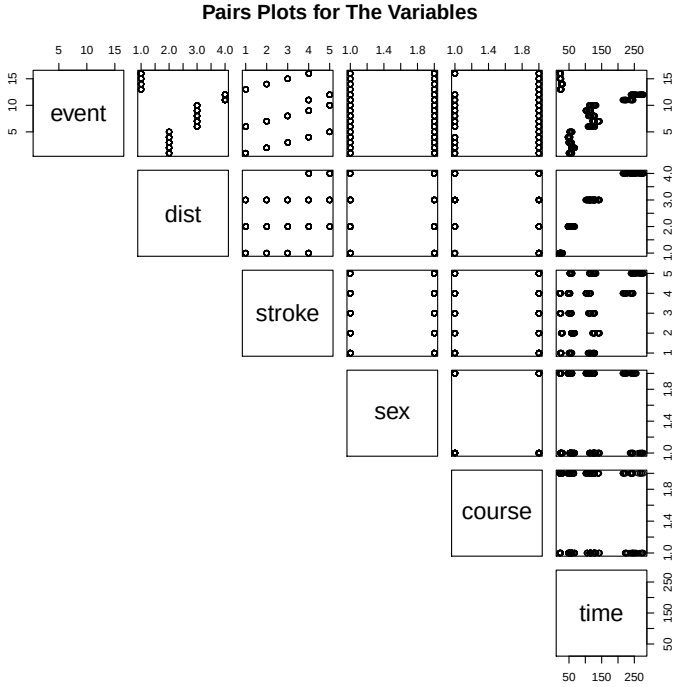

Figure 3: Pair plot

However, it is easy to see that there is a relation between event and time, and dist and time. Now to

confirm our guess using the Figure 2, we plot time variable against the other categorical variables to see if we can decipher the reasons behind the clusters.

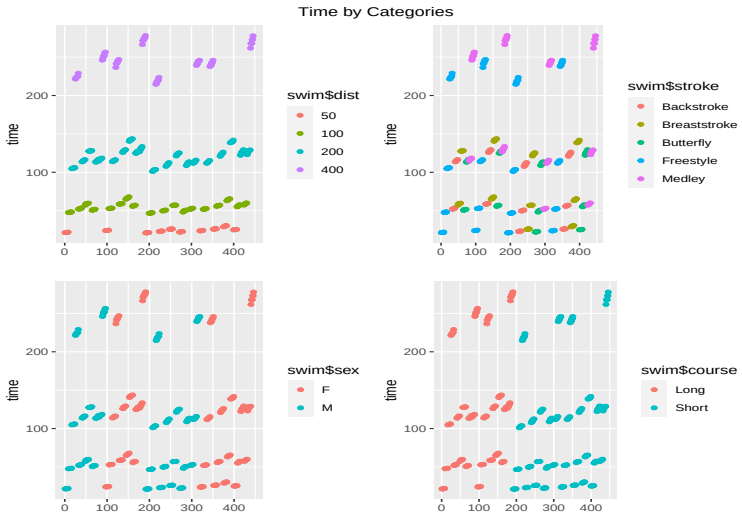

Figure 4: Time variable cluster analysis

With Figure 4, it is now clear that all these four variables have an impact on time. Distance seems to make the biggest difference. Freestyle seems to be the fastest stroke, followed by backstroke, butterfly, and medley respectively, and the slowest stroke seems to be breaststroke. Females seem to be slower than males and long courses seem to take more time than the short courses. All of these observations are compatible with common sense.

2 Statistical Modelling

2.1 Model Selection

Guided by these observations, we take the logs of time and dist variables and fit a linear regression using all the explanatory variables and their interactions. We then use the *dredge* function in the $MuMIn$ package in R to do a model selection based on the Akaike Information Criterion. The resulting equation is given by the Equation 1:

$$log(time_i) = \alpha_i + \beta log(d_i) + \sum_{j=1}^{4} \gamma_j s_{ij} + \rho sex_i + \kappa c_i + \delta_i + e_i \tag{1}$$

where δ_i stands for the interaction terms:

$$\delta_i = \sum_{j=1}^{4} \epsilon_j s_{ij} d_i + \theta log(d_i) sex_i + \eta sex_i s_i + \phi c_i s_i + \gamma c_i sex_i \tag{2}$$

where alpha is the intercept, d stands for distance, s stands for stroke, sex stands for sex, and c stands for course. As can be understood, the interaction term between $course$ and $log(dist)$ is dropped as a result of the AIC elimination. The results are given below:

Table 1: Regression Results

	Dependent variable:	
	log(time)	
	(1)	(2)
courseShort	-0.039^{***}	-0.034^{***}
	(0.012)	(0.003)
log(dist)	1.116^{***}	1.117^{***}
	(0.004)	(0.003)
sexM	-0.154^{***}	-0.154^{***}
	(0.011)	(0.011)
strokeBreaststroke	0.140^{***}	0.140^{***}
	(0.021)	(0.021)
strokeButterfly	-0.187^{***}	-0.187^{***}
	(0.021)	(0.021)
strokeFreestyle	-0.087^{***}	-0.085^{***}
	(0.017)	(0.017)
strokeMedley	0.134^{***}	0.133^{***}
	(0.023)	(0.023)
courseShort:log(dist)	0.001	
	(0.002)	
courseShort:sexM	-0.012^{***}	-0.012^{***}
	(0.003)	(0.003)
courseShort:strokeBreaststroke	0.010^{**}	0.010^{**}
	(0.004)	(0.004)
courseShort:strokeButterfly	0.023^{***}	0.023^{***}
	(0.004)	(0.004)
courseShort:strokeFreestyle	0.022^{***}	0.022^{***}
	(0.004)	(0.004)
courseShort:strokeMedley	0.012^{**}	0.013^{***}
	(0.005)	(0.004)
log(dist):sexM	0.010^{***}	0.010^{***}
	(0.002)	(0.002)
log(dist):strokeBreaststroke	-0.005	-0.005
	(0.004)	(0.004)
log(dist):strokeButterfly	0.034^{***}	0.034^{***}
	(0.004)	(0.004)
log(dist):strokeFreestyle	-0.005	-0.005
	(0.003)	(0.003)
log(dist):strokeMedley	-0.023^{***}	-0.023^{***}
	(0.004)	(0.004)
sexM:strokeBreaststroke	-0.007^{*}	-0.007^{*}
	(0.004)	(0.004)
sexM:strokeButterfly	0.002	0.002
	(0.004)	(0.004)
sexM:strokeFreestyle	0.010^{***}	0.010^{***}
	(0.004)	(0.004)
sexM:strokeMedley	0.004	0.004
	(0.004)	(0.004)
Constant	-1.066^{***}	-1.070^{***}
	(0.018)	(0.016)
Observations	446	446
R^2	1.000	1.000
Adjusted R^2	1.000	1.000
Residual Std. Error	0.013 (df = 423)	0.013 (df = 424)
F Statistic	$66{,}439.040^{***}$ (df = 22; 423)	$69{,}732.060^{***}$ (df = 21; 424)
Note:		*p<0.1; **p<0.05; ***p<0.01

The interpretation of Table 1 will be provided in Section 3. Further, we use Box-Cox analysis to determine if a re-scaling is necessary. The result is given in Figure 5

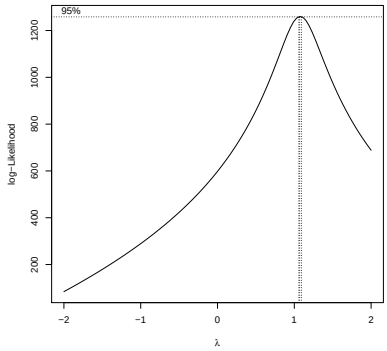

Figure 5: Box-Cox Analysis

The optimal exponent for the dependent variable seems to be 1.1. We decide not to re-scale the variable as this is a very small adjustment and the model fit seems to be decent as can be seen in Section 2.2.

2.2 Outlier Analysis

We first take a look at the residuals versus the fitted values.

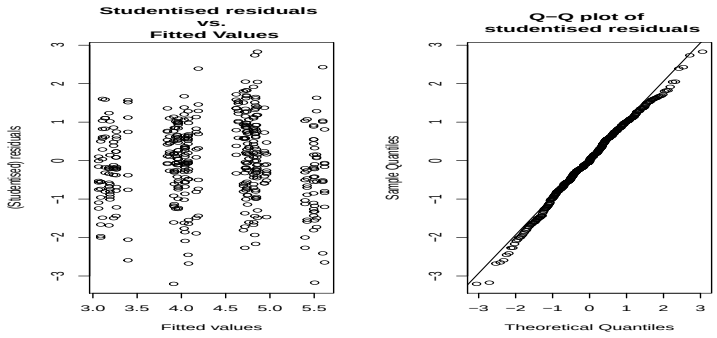

Figure 6: Analysis of The Residuals

There seems to be no distinguishable pattern in the plot of studentised residuals versus the fitted values. This confirms one of the key assumptions of our model: residuals and the fitted values are independent. Further looking at the Q-Q plot, we see that residuals line up with the percentiles of a normal distribution. This verifies the other key assumption of our linear modelling, that is the residuals follow a normal distribution. These two graphs are evidences that our fit is quite good.

As can be seen from Figure 7, there are outliers at the extremes. However, the fit is not bad on either end as for example even the minimum value seems to be in line with the studentised Q-Q plot. There are a few outliers as can be seen from the studentised residuals graph. However, these outlier values are still close enough to some of the clusters and some outliers are expected in athletic competition as talent varies vastly among athletes. The leverage values seem to be all below the critical threshold.

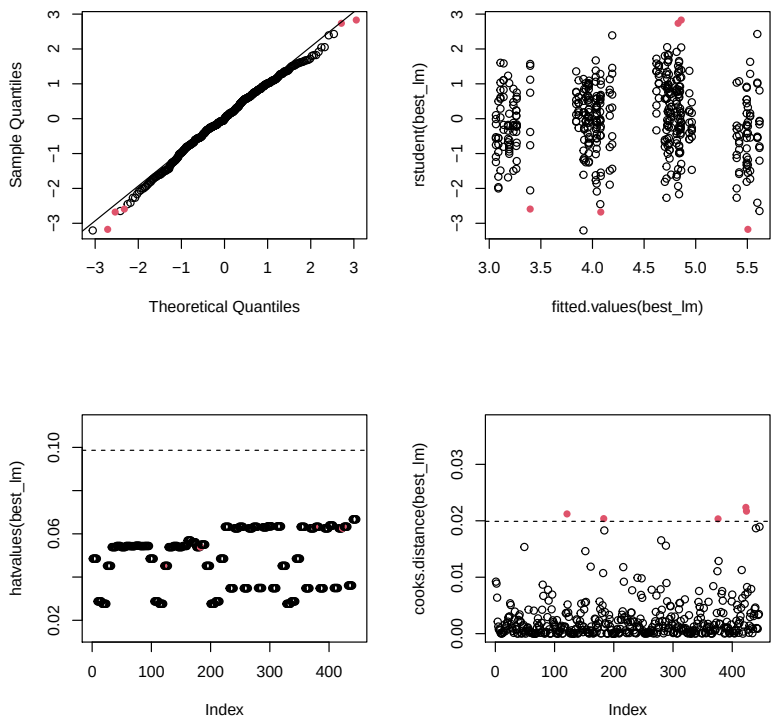

Figure 7: Model Diagnosis

As a result of the visualizations, we decide not to drop any of the outliers as none of the outliers seem to be significantly farther from the other observations. Also, literature suggests that in athletic competitions, talent creates vast differences in time across athletes. So, we believe all of the observations are reasonable.

3 Model Interpretation

In Table 1, the intercept stands for the $log(time)$ seconds it takes for *a female, in a long course backstroke competition of 0 meters*. The italic text is our ***baseline***. This is not a realistic number because of 0 distance. The intercept only gains meaning through other variables in such a model.

Now we derive some equalities to talk about the impact of the terms in Equation 1.

$$log(y) = \beta log(x) \iff \frac{\delta y}{y} = \beta \frac{\delta x}{x} \tag{3}$$

$$log(y) = \beta x \iff \frac{\delta y}{y} = \gamma \delta x \tag{4}$$

From Equation 3, we can see that β stands for the percent change in y in reaction to a percent change in x. From Equation 4, we can see that γ stands for the percent change in y in reaction to a unit change in x. Since all of our terms in Equation 1 are in either one of these forms. We can now interpret the impact of each specific variable. We use the second column in Table 1 which stands for the best linear model after AIC elimination.

1. **dist:** The coefficient on $log(dist)$ is 1.1, this means that for every percentage change in distance the time goes up by 1.1% seconds in the *baseline* category. This is significant at the 1% level. Since there are interaction terms involving $log(dist)$, we can see that for male athletes, the time goes up by 1.11%

for every 1% change in distance, which is also significant at 1% level. The time also goes up by 1.11%, 1.12% , 1.12%, and 1.11% seconds respectively for breast,butterfly, freestyle, and medley strokes as opposed to backstroke for females for a 1% increase in distance. The coefficients for interaction terms involving butterfly and medley are significant at the 1% level. The other coefficients are insignificant.

2. **sex:** The coefficient on sex is -.154. This means that given all other *baseline* characteristics, time goes down by -.15% seconds for males as opposed to females. The interaction term between course and sex suggests that the differential impact of course on percentage change in time for males is -.01%. The interaction term for $log(dist)$ and sex was explained in Bullet point 1.

3. **course:** The coefficient on course suggests that it takes .03% less time to finish a Long course for someone with baseline characteristics as opposed to a short course. This effect is significant at 1% level. The differential impact of different strokes are -.01%, .01%, .02%, .02%, .01% for breaststroke, butterfly, freestyle, and medley respectively as opposed to backstroke. The differential impact of medley, butterfly and freestyle are significant at the 1% level, the differential impact of breaststroke is significant at the 5% level.

4. **stroke:** All the coefficients related to the interactions of stroke are explained in other bullet points. In the baseline, the percentage time changes .14%, -.19%, -.09%, and .13% for breaststroke, butterfly, freestyle, and medley respectively as opposed to backstroke. So, we can see that the fastest stroke is butterfly among the different styles.

These findings are in line with the observations we made following Figure 4.

4 Model Predictions

The predictions are given in Table 2

Table 2: Predictions

	dist	stroke	sex	course	dist_factor	pred	lower	upper
1	400.00	Freestyle	F	Long	400	245.98	239.70	252.42
2	50.00	Backstroke	F	Long	50	27.11	26.40	27.84
3	400.00	Butterfly	F	Long	400	280.62	273.24	288.20
4	100.00	Medley	F	Long	100	60.41	58.82	62.04

As a sanity check, we try to verify the predictions through interpolation using the results from Table 1, and Figure 4.

```
> # Sanity check
> mean(swim[dist==400 & stroke=="Freestyle" & sex=="F" & course=="Long",time])
[1] 243.0788
> (mean(swim[dist== 100 & stroke == "Backstroke" & sex== "F" & course=="Long", time])/2
+ mean(swim[dist== 50 & stroke == "Freestyle" & sex== "F" & course=="Long", time]))/2
[1] 26.83094
> (mean(swim[dist== 200 & stroke == "Butterfly" & sex== "F" & course=="Long", time])*2
+ mean(swim[dist== 400 & stroke == "Medley" & sex== "F" & course=="Long", time]))/2
[1] 262.9737
> (mean(swim[dist== 200 & stroke == "Medley" & sex== "F" & course=="Long", time])/2
+ mean(swim[dist== 100 & stroke == "Butterfly" & sex== "F" & course=="Long", time]))/2
[1] 60.79625
```

These values are chosen to be exactly the same as the explanatory variables in the prediction if such values exist. Otherwise, using Figure 4, we choose the closest approximation e.g. there is no data for a 400 meter butterfly race. We observe that butterfly and medley are close enough given the data. Then, we extrapolate a time for a 400 meter butterfly race using 400 meter medley race, and multiplying the 200 meter butterfly race time by 2. We then take the mean of these two variables.

All predictions fall between the prediction intervals except the prediction for 400 meter butterfly race. This is tolerable as 400 meter is a far cry from 200 meter races and medley turns out to be not a great approximation for butterfly stroke.

5 Conclusion

In this analysis, we found that all the explanatory variables are significantly important for predicting race times. We analyzed the differential impact of these variables across different characteristics in Section 2.2. The fit of the model was good, which we checked using multiple diagnosis methods. None of the outliers stood out enough to be dropped from our analysis. The goodness of fit was further confirmed in the Section 4. We both used our model to make predictions, and also verified the predictions using interpolation from the existing data points. We believe that the models could have been improved if we had a variable that represent talent, or at least a proxy of it as we believe that talent is a significant factor in athletic races. We believe that such a variable would further improve the fit of the models on the extremes.

6 Code

```
library(data.table)
library(dplyr)
library(ggplot2)
library(gridExtra)
library(leaps)
library(MuMIn)
library(MASS)
library(grid)
library(stargazer)
library(xtable)

options(na.action = "na.fail")

swim <- read.csv("C:/Users/kaany/OneDrive/Desktop/Practicals/
MT4/swim.csv", stringsAsFactors=T)

swim = as.data.table(swim)
swim[,dist_factor:= as.factor(dist)]
# 1.1
str(swim)
head(swim)
tail(swim)
par(mfrow= c(1,2))


summary(swim[,-c("time","event")])

pdf("Pairs_Plots_for_The_Variables.pdf")
pairs(swim, lower.panel = NULL, main="Pairs_Plots_for_The_Variables")
dev.off()

pdf("event_bar_chart.pdf")
ggplot(swim, aes(x=event)) +
  geom_bar(fill = "navy") +
   ggtitle("Bar_Chart_for_The_Distribution_of_Events") +
   labs(y="Count", x = "Type_of_the_Event") +
   theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5),
         plot.title = element_text(hjust = 0.5))
dev.off()

pdf("dist_of_time.pdf")
par(mfrow= c(1,2))
boxplot(swim$time, ylim=c(min(swim$time),max(swim$time)), ylab="Time",
         main = "Box_and_Whiskers_Plot_for_Time")
plot(swim$time, xlab="Observation", ylab="Time",
      main="Scatter_Plot_for_Time")
dev.off()

pdf("Time_by_Categories.pdf")
g1 = qplot(y=time, data = swim, colour = swim$dist)
g2 = qplot(y=time, data = swim, colour = swim$sex)
g3 = qplot(y=time, data = swim, colour = swim$stroke)
g4 = qplot(y=time, data = swim, colour = swim$course)
```

```
my_plots <- list(g1, g2, g3,g4)
my_layout <- rbind(c(1, 3), c(2, 4))
grid.arrange(grobs = my_plots, layout_matrix = my_layout,
             top = textGrob("Time by Categories"))
dev.off()

# swim[, dist:= as.integer(dist)]
# 1.2
# lm1 <- lm(log(time) ~ (dist_factor + stroke + sex + course)^2, data = swim)
lm1 <- lm(log(time) ~ (course +log(dist) +  sex + stroke)^2, data = swim)

b = dredge(lm1)
best_lm = get.models(b, 1)[[1]]
# qqnorm(rstudent(best_lm), main = "Q-Q plot of \n studentised residuals")
# qqline(rstudent(best_lm))

s <- stargazer(lm1, best_lm, title="Regression Results",
align=TRUE, column.sep.width = "0.4pt",
font.size="tiny")
fileConn<-file("stargazer.txt")
writeLines(s, fileConn)
close(fileConn)

  # best_lm = lm(log(time) ~ course + log(dist) + sex + stroke +
#       course:sex + course:stroke + log(dist):sex + log(dist):stroke +
#       sex:stroke + 1, data = swim)

summary(best_lm)

par(mfrow=c(1,1))
pdf("Box Cox for the best model.pdf")
boxcox(lm1, lambda=seq(-2,2,1/10))
dev.off()
res = boxcox(lm1, lambda=seq(-2,2,1/10), plotit=F)
dev.off()
lambda = res$x[which(res$y==max(res$y))]



pdf("residuals vs fitted values.pdf")
par(mfrow=c(1,2))
plot(resid(best_lm) ~ fitted(best_lm), data = swim,
     xlab = "Fitted values", ylab = "Residuals")
plot(rstudent(best_lm) ~ fitted(best_lm), xlab = "Fitted values",
     ylab = "(Studentised) residuals")
dev.off()

pdf("qq plots for the residuals.pdf")
par(mfrow = c(1,2))
plot(rstudent(best_lm) ~ fitted(best_lm), main = "Studentised
residuals \n vs. \n Fitted Values",
xlab = "Fitted values", ylab = "(Studentised) residuals")
qqnorm(rstudent(best_lm), main = "Q-Q plot of \n studentised
residuals")
qqline(rstudent(best_lm))
dev.off()


(n <- dim(swim)[1])
# (p <- dim(swim)[2])
# Hard-coding this to account for the interaction terms in the model
p <- length(coef(best_lm))

(i <- cooks.distance(best_lm) > (8/(n - 2*p)))

pdf("Pairs plot with red outliers.pdf")
pairs(swim, lower.panel = NULL, pch = 1 + 15*i, col = 1 + i)
dev.off()

pdf("Model diagnosis with red dots.pdf")
par(mfrow = c(2, 2))
```

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

plot(fitted.values(best_lm), rstudent(best_lm), pch = 1 + 15*i, col = 1 + i)
# text(fitted.values(best_lm), rstudent(best_lm), abbreviate(row.names(sw)), adj = -0.2)

# Leverage values
plot(hatvalues(best_lm), ylim = c(min(hatvalues(best_lm))*1/2,
max(hatvalues(best_lm))*5/3), pch = 1 + 15*i, col = 1 + i)
# text(hatvalues(best_lm), row.names(sw), srt = 90, adj =
-0.1)
abline(2*p/n, 0, lty = 2)

plot(cooks.distance(best_lm), ylim =
c(min(cooks.distance(best_lm)) * 1/2,
max(cooks.distance(best_lm)) * 5/3), pch = 1 + 15*i, col = 1 +
i)
# text(cooks.distance(best_lm), row.names(swim), srt = 90, adj = 1.1)
abline(8/(n - 2*p), 0, lty = 2)
dev.off()

# matrix(c(1,2,3,4, 1,2,3,4), nrow=2)

data <- data.table(dist = c(400,50,400,100),
                   stroke= c("Freestyle","Backstroke", "Butterfly","Medley"),
                   sex= rep("F",4),
                   course= rep("Long", 4))
data[,dist_factor:=as.factor(dist)]

# data[,event:= paste0(dist, " m ", stroke)]
# data[,time:=NA]
# data[,time_bc := time^lambda]
pred_table = cbind(data,exp(predict(best_lm,as.data.frame(data), interval="predict")))

xtable(pred_table)
> # Sanity check
> mean(swim[dist==400 & stroke=="Freestyle" & sex=="F" & course=="Long",time])
> (mean(swim[dist== 100 & stroke == "Backstroke" & sex== "F" & course=="Long", time])/2
+ mean(swim[dist== 50 & stroke == "Freestyle" & sex== "F" & course=="Long", time]))/2
> (mean(swim[dist== 200 & stroke == "Butterfly" & sex== "F" & course=="Long", time])*2
+ mean(swim[dist== 400 & stroke == "Medley" & sex== "F" & course=="Long", time]))/2
> (mean(swim[dist== 200 & stroke == "Medley" & sex== "F" & course=="Long", time])/2
+ mean(swim[dist== 100 & stroke == "Butterfly" & sex== "F" & course=="Long", time]))/2
```