1. Introduction

This report presents an analysis of “Swim” data collected by students at the Queensland University of Technology.

The goal of this report is to determine the factors affecting the time to swim one lap of a 25m pool. Following factors were measured during the experiment: lap number, presence of goggles, presence of a shirt, presence of flippers, and the end at which the lap started. Students recorder 24 laps in experimental conditions.

2. Exploratory Data Analysis

There are 5 potential explanatory variables available: Order (lap number), Shirt (0 – No, 1 – Yes), Goggles (0 – No, 1 – Yes), Flippers (0 – No, 1 – Yes), End (0 – start from shallow end, 1 – start from deep end). The last four variables are categorical variables with 0 as a baseline level.

Box-plots shown on Figure 1 illustrate that flippers have a significant impact on the time of swimming one lap, while goggles and shirt have a mild impact and End variable has hardly any impact. Presence of flippers and goggles is associated with a decrease of time, while presence of shirt is associated with an increase of time. Plot shown on Figure 2 shows no correlation between time and order, that is the tiredness of the swimmer probably did not influence the swimming time. Swimmer said that she took a break after approximately 20th lap (because she felt that not taking a break could affect the experiment), but we can see that it did not impact the lap times All these conclusions will be explored further and verified by suitable experiments in the next part of the report.

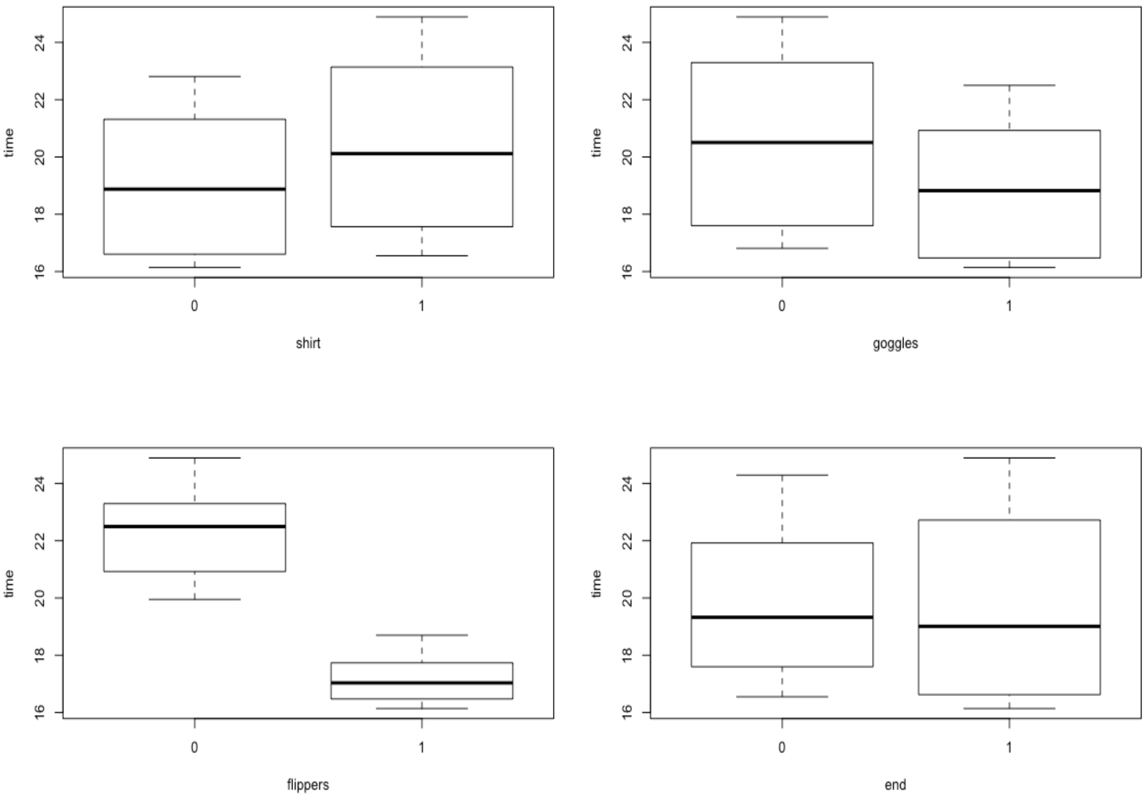

Figure 1. Box plots of time depending on each of the four categorical variables.

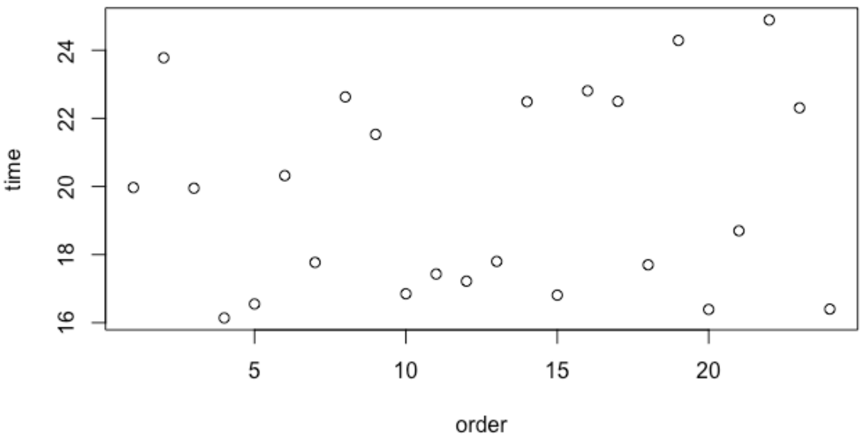

Figure 2. Plot of time against order (number of the lap), variables are independent.

We have also plotted the interaction plots for the categorical variables. Results presented on the Figure 3 reveal a very strong interaction between Goggles and End and between Shirt and End, and a strong interaction between Flippers and End. Lines on the three other plots (for pairs (Goggles, Shirt), (Flippers, Shirt) and (Goggles, Flippers) are close to being parallel so there is only a weak interaction between those variables.

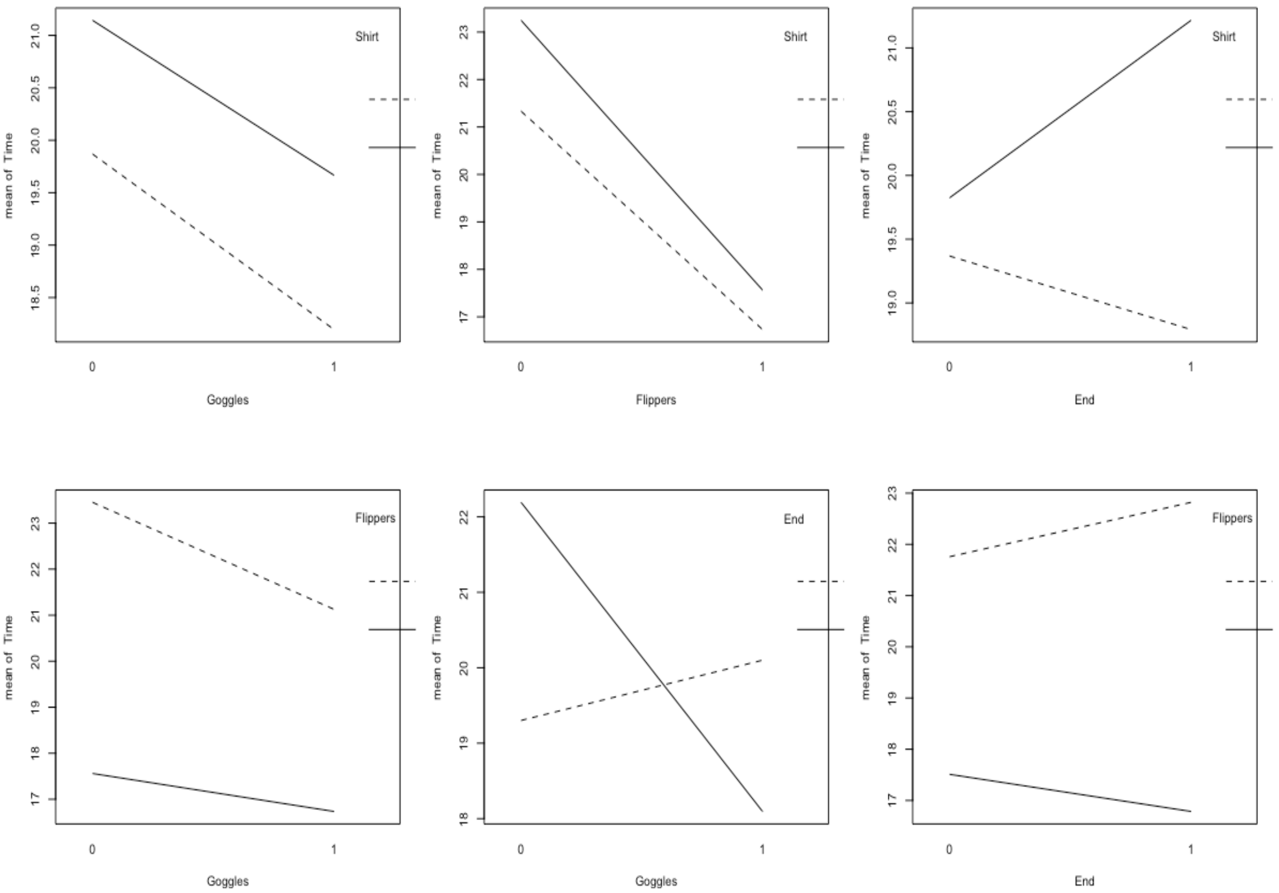

Figure 3. Interaction plots presenting the interactions between each pair of categorical explanatory variables. The x-axis indicates one of the explanatory variables, the two lines indicate the other explanatory variable and the y-axis indicates the mean time taken to swim one lap.

3. Analysis of the Data

3.1 Initial model selection

We start by fitting a normal linear model which includes the main effects and all possible interactions of the potential explanatory variables. However, we notice that data for all 5-way and all 4-way and some 3-way interactions is not available – it is caused by the fact that the number of possible explanatory variables exceeds the numbers of samples (24). The same happens for a model containing 2-way, 3-way and 4-way interactions and for a model containing 2-way and 3-way interactions. Thus, we decide to start the model selection process with a full model containing only 2-way interactions. We assume that the errors are independent and normally distributed with constant variance. The model equation is:

$$E(Time) = \beta_0 + \beta_1 * Order + \beta_2 \mathbb{I}_{Shirt=1} + \beta_3 \mathbb{I}_{End=1} + \beta_4 \mathbb{I}_{Goggles=1} + \beta_5 \mathbb{I}_{Flippers=1} + \beta_6 \mathbb{I}_{Shirt=1,End=1} + \beta_7 \mathbb{I}_{Shirt=1,Goggles=1} + \beta_8 \mathbb{I}_{Shirt=1,Flippers=1} + \beta_9 \mathbb{I}_{End=1,Goggles=1} + \beta_{10} \mathbb{I}_{End=1,Flippers=1} + \beta_{11} \mathbb{I}_{Goggles=1,Flippers=1} + \beta_{12} Order * \mathbb{I}_{Shirt=1} + \beta_{13} Order * \mathbb{I}_{End=1} + \beta_{14} Order * \mathbb{I}_{Googles=1} + \beta_{15} Order * \mathbb{I}_{Flippers=1}$$

Model (*)

Levels 0 of Shirt, End, Goggles and Flippers are used as baselines.

Figure 4 presents residual against fitted values – we can see a random distribution, so the assumptions about standard errors are correct. Plot shown on Figure 5 presents that majority of the points fits well the straight line – it is another confirmation that our assumptions required to use the linear model are correct.

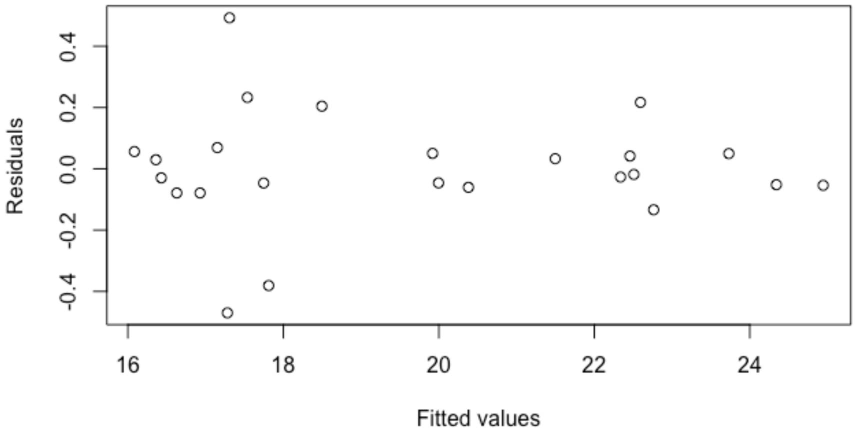

Figure 4. The x-axis indicates fitted values and the y-axis indicates residuals.

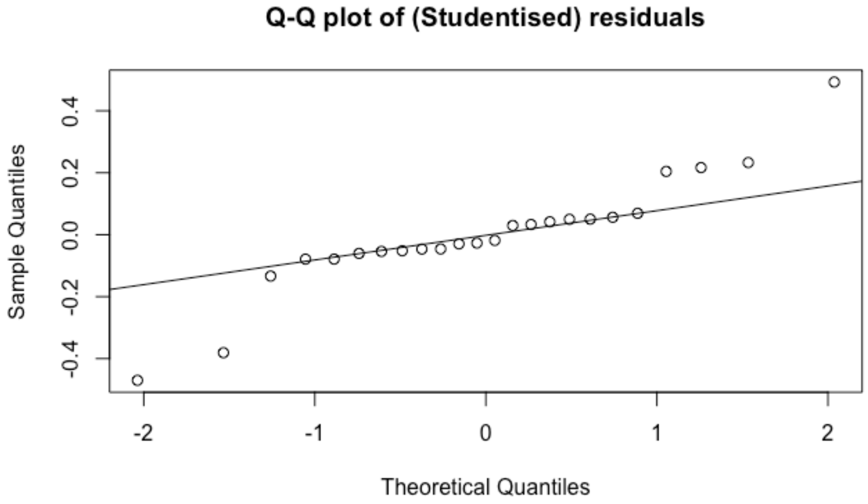

Figure 5. Q-Q plot comparing the studentised residuals to a standard normal distribution.

3.2 Outlier analysis

We start the outlier analysis by plotting the Cook's distance for each data point (presented on Figure 6). We have n = 24 and p = 16, so the bound 8/(n-2p) = -1 will not help us to identify the outliers. However, we notice that points 9 and 10 have a much higher influence than other points. If they also have high leverage, then there are outliers and should be removed. We can see on Figure 7 that those two points indeed have high leverage, so we decide to remove them. Throughout the rest of this report we will be using the dataset without those two points.

Figure 8 presents the plots after the two data points have been removed. Around 95% of all data points has the standardised residual between -2 and 2, exactly as expected. More points with large Cook's distance have appeared but there are no outstanding values, and there are also no points with unusually high leverage (compared to the leverage of other points), so we decide not to remove any other points at this stage.

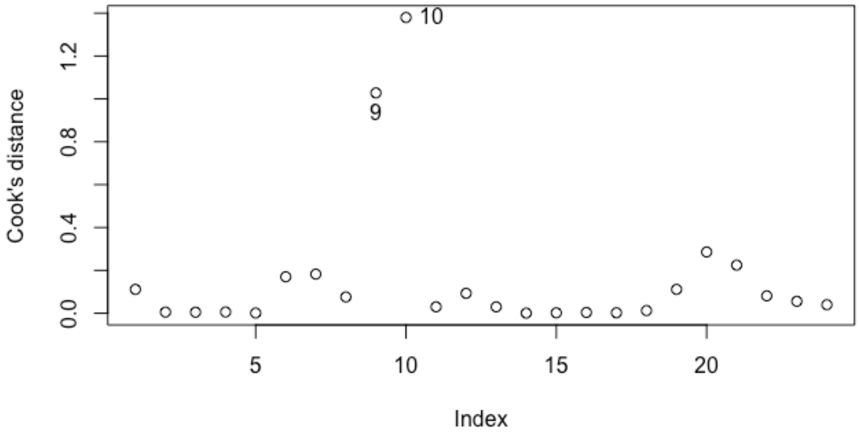

Figure 6. Cook's distance for each data point (labelled by Order variable). Points 9 and 10 have unusually large Cook's distance.

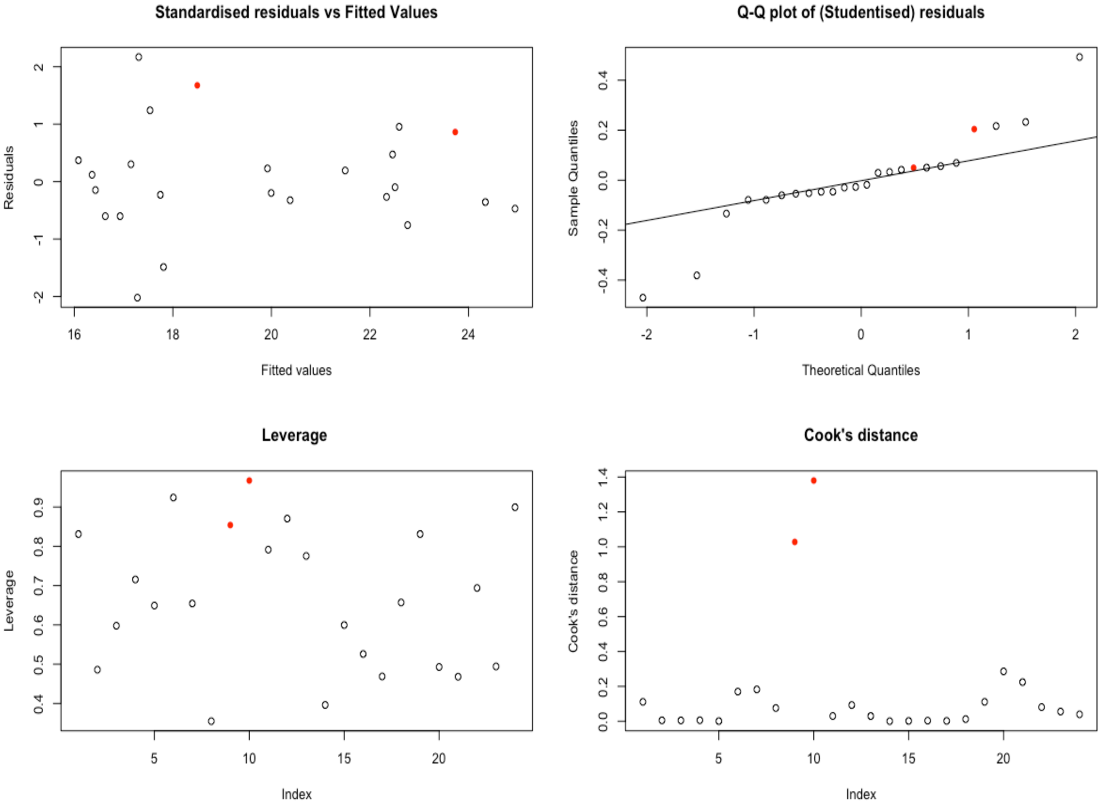

Figure 7. Points 9 and 10 are marked as full red circles. The first plot presents standardised residuals against fitted values. As expected, around 95% of points has the standardised residual between -2 and 2. The second plot presents a Q-Q plot comparing the studentised residuals to a standard normal distribution. The third plot presents the leverage of each point. The fourth plot presents the Cook's distance.

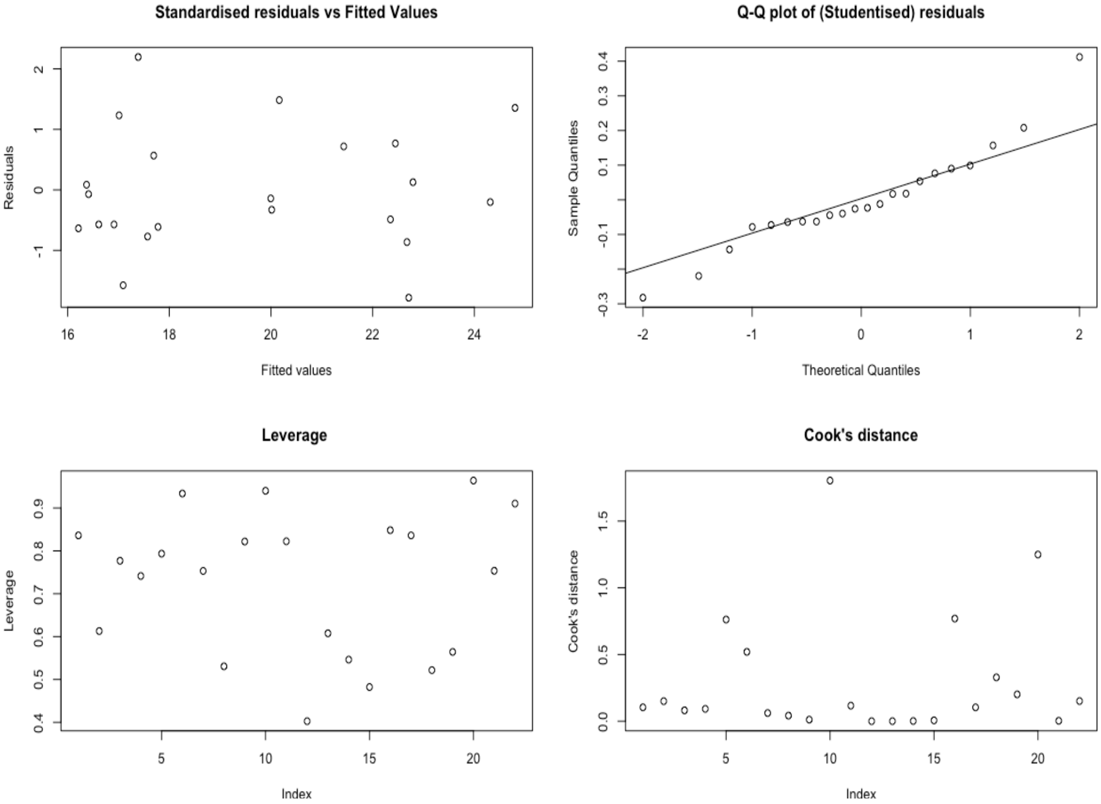

Figure 8. The same plots as on the previous figure, but now data does not include points 9 and 10.

3.3 Model selection

```
                     Estimate    Std. Error   t value    Pr(>|t|)
(Intercept)          23.88823    0.87560      27.282     1.6e-07 ***
Order                -0.06690    0.04101      -1.631     0.153942
Shirt1               -0.54137    0.81209      -0.667     0.529788
Goggles1             -3.90160    0.81014      -4.816     0.002952 **
Flippers1            -4.57851    0.51172      -8.947     0.000109 ***
End1                 -1.33235    0.99524      -1.339     0.229151
Order:Shirt1         0.11775     0.04291      2.744      0.033539 *
Order:Goggles1       0.07609     0.04132      1.842      0.115136
Order:Flippers1      -0.08093    0.04253      -1.903     0.105729
Order:End1           0.08171     0.05851      1.397      0.212038
Shirt1:Goggles1      0.84314     0.61178      1.378      0.217339
Shirt1:Flippers1     -0.86425    0.33520      -2.578     0.041865 *
Shirt1:End1          -0.13093    0.58845      -0.223     0.831302
Goggles1:Flippers1   1.53705     0.40665      3.780      0.009183 **
Goggles1:End1        0.96344     0.72644      1.326      0.233006
Flippers1:End1       -0.40347    0.35117      -1.149     0.294310
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 1. Data about coefficients of the current model (*) – output from the R code attached in the appendix.

We proceed to performing backwards elimination search via ANOVA. Looking at the summary of the current model (Table 1), we note that no term involving variable End is marked as significant. Thus we will test for significance of all terms involving variable End, that is we perform a direct F-test for null model (*) containing all 2-way interactions and alternative model not involving End. ANOVA test returns p-value 0.2549. It is not significant, so we can simplify our model by ignoring End. The new equation for our model (**) is following:

$$E(Time) = \beta_0 + \beta_1 Order + \beta_2 \mathrm{I}_{Shirt=1} + \beta_3 \mathrm{I}_{Goggles=1} + \beta_4 \mathrm{I}_{Flippers=1} + \beta_5 \mathrm{I}_{Shirt=1,Goggles=1} + \beta_6 \mathrm{I}_{Shirt=1,Flippers=1} + \beta_7 \mathrm{I}_{\ Goggles=1,Flippers=1} + \beta_8 Order * \mathrm{I}_{Shirt=1} + + \beta_9 Order * \mathrm{I}_{Googles=1} + \beta_{10} Order * \mathrm{I}_{Flippers=1}$$

```
                     Estimate    Std. Error   t value    Pr(>|t|)
(Intercept)          22.80036    0.47206      48.300     3.67e-14 ***
Order                -0.01257    0.02780      -0.452     0.660037
Shirt1               0.33514     0.55925      0.599      0.561140
Goggles1             -2.87733    0.49334      -5.832     0.000114 ***
Flippers1            -5.12649    0.49996      -10.254    5.75e-07 ***
Order:Shirt1         0.08204     0.02766      2.966      0.012845 *
Order:Goggles1       0.05360     0.03220      1.665      0.124125
Order:Flippers1      -0.03012    0.03519      -0.856     0.410214
Shirt1:Goggles1      0.29428     0.29254      1.006      0.336052
Shirt1:Flippers1     -0.73293    0.35718      -2.052     0.064754 .
Goggles1:Flippers1   1.35897     0.37207      3.652      0.003804 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 2. Data about coefficients of the current model (**) – output from the R code attached in the appendix.

Looking at the summary of the new model (**) we notice Flippers, Goggles and their interaction are all very significant. On the other hand, Shirt and Order are not significant (p-values 0.66 and 0.56 respectively), and only their interaction is significant (p-value 0.01). We will try to remove either one of these variables.

We perform a direct F-test for null model (**) and alternative model not involving "Order". ANOVA test returns p-value 0.04754, which is significant (at 5% significance level), therefore we cannot remove the Order variable.

We perform a direct F-test for null model (**) and alternative model not involving "Shirt". ANOVA test returns p-value 0.0001548, which is very significant, therefore we cannot remove the Shirt variable.

Thus, we cannot remove Shirt or Order. Looking again at the summary of the new model (**) we notice that interaction terms between Shirt, Order and Flippers, Goggles are not significant at all, so we decide to try removing those interaction terms. That is, we want to compare model (**) with model (***) described below:

$$E(Time) = \beta_0 + \beta_1 Order + \beta_2 \mathbb{I}_{Shirt=1} + \beta_3 \mathbb{I}_{Goggles=1} + \beta_4 \mathbb{I}_{Flippers=1} + \beta_5 Order\mathbb{I}_{Shirt=1} + \beta_6 \mathbb{I}_{Goggles=1, Flippers=1}$$

We perform a direct F-test for null model (**) and alternative model (***). ANOVA test returns p-value 0.1211, which is not significant, therefore we decide to simplify our model to model (***).

```
                        Estimate      Std. Error     t value        Pr(>|t|)
(Intercept)             22.13700      0.31031  7     1.338          < 2e-16 ***
Order                   0.02812       0.01681        1.673          0.1150
Shirt1                  0.20185       0.42351        0.477          0.6405
Goggles1                -1.89300      0.26843        -7.052         3.92e-06 ***
Flippers1               -5.49631      0.27095        -20.285        2.57e-12 ***
Order:Shirt1            0.08211       0.03003        2.734          0.0154 *
Goggles1:Flippers1      1.01672       0.36868        2.758          0.0147 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 3. Data about coefficients of the current model (***) – output from the R code attached in the appendix.

We can see that all variables are significant in this model ((even though Order and Shirt are not significant, their interaction is), so we can stop the backwords elimination search via ANOVA. We decide to perform the outlier analysis again and check how well the model fits our data, the results are presented on Figure 9. From the first plot we conclude that residuals and fitted values are indeed independent – the assumptions required for the linear model are still valid. On the second plot we can see Q-Q plot comparing data points with the normal distribution – points fit the straight line very well. On the third plot we can see that there are no points with an exceptionally high leverage – it is a good sign. And similarly, on the fourth plot we see that there are no points with exceptionally high influence. Therefore, there is no need to remove any other data points.

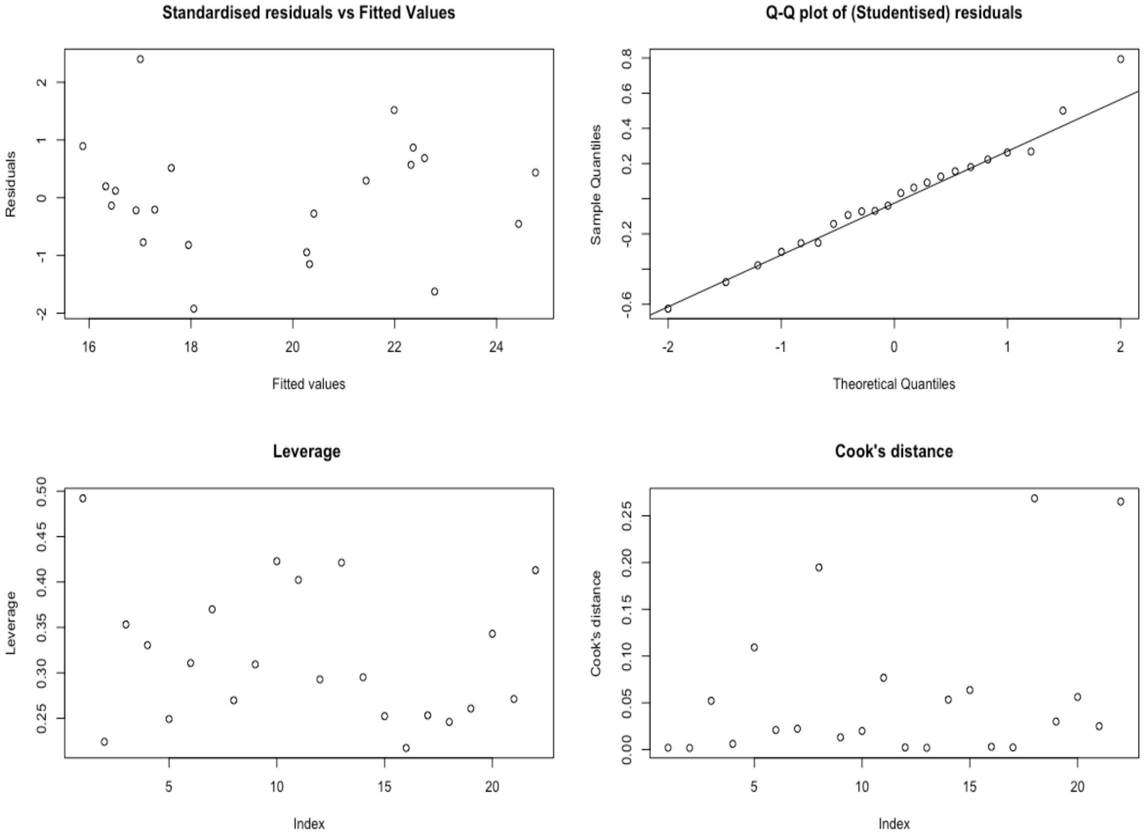

Figure 9. The first plot presents standardised residuals against fitted values. As expected, around 95% of points has the standardised residual between -2 and 2. The second plot presents a Q-Q plot compares the studentised residuals to a standard normal distribution. The third plot presents the leverage of each point. The fourth plot presents the Cook's distance.

4. Conclusions

During the initial outliers analysis two points were identified as outliers and removed. We started the model selection with a full model containing all two-way interactions, but some variables were found not to be significant. Our final model is represented by the equation:

$$E(Time) = \beta_0 + \beta_1 Order + \beta_2 \mathbb{I}_{Shirt=1} + \beta_3 \mathbb{I}_{Goggles=1} + \beta_4 \mathbb{I}_{Flippers=1} + \beta_5 Order \mathbb{I}_{Shirt=1} + \beta_6 \mathbb{I}_{Goggles=1, Flippers=1}$$

Variable End has been removed completely as it was found not to be significant. Only two two-way interactions were included: Order and Shirt, and Goggles and Flippers, because other interactions were found to be insignificant. Even though the swimmer said that she had the impression that she was slower when swimming from the shallow to the deep end, the data did not support this statement. Goggles and Flippers (and the Intercept) were found to be the most significant variables and their estimated coefficients are positive, it means that their presence caused the swimmer to swim faster - this aligns with the information given by the swimmer and with our expectations.

Estimate 22.137 (95% confidence interval: (21.48, 22.80)) of the intercept means that if we ignore the Order, and assume no Shirt, Goggles or Flippers are present then the expected time to swim one lap is equal to 22.137s. The statistical interpretation of the significance of the Goggles*Flippers interaction term is that their effect is not additive: wearing Goggles (and not Flippers) decreases the expected time by 1.89s (95% confidence interval: (-2.47, -1.32)), wearing Flippers (and not Goggles) decreases the expected time by 5.5s (95% confidence interval: (-6.07, -4.92)), while wearing both Goggles and Flippers decreases the expected time only by 6.37s. When the swimmer is wearing a shirt, the expected time increases by 0.08s (95% confidence interval: (0.02, 0.15)) with each consecutive lap. However, neither Shirt or Order has been found to be significant on their own (and the 95% confidence intervals for the two corresponding coefficients contain both positive and negative values). The analysis performed at the end of the previous part confirms that this is a suitable model for this dataset.