Linear Models, Marked Practical

P118

November 9, 2021

1 Summary

In this report we examine trends in competitor times in swimming races using data from the finals of individual events at the 2016 Olympics and the finals of the 2016 World Championship, exploring the dependence of swim times on other information about the event. We use a Box-Cox transformation to model a normal linear relationship between time and the explanatory variables and discuss the suitability and interpretability of this model compared to other possible normal linear models for the data. We then use this fitted model to determine the significance and effect of variables in the model. Finally, we use the model to predict times for additional races.

2 Introduction

We are examining data consisting of competitor times for swimming races from the individual events at the 2016 Olympics and the 2016 World Championships. Table 1 summarises the information about variables recorded for each time in the dataset. Note that since `event` combines the information contained in `dist`, `stroke`, `sex` and `course`, it contains no additional information and may be ignored in our analysis. `time` is a continuous variable while `stroke`, `sex` and `course` are unordered factors. `dist` variable represents a numerical value, however, it is restricted to only 4 values and so may be treated either as a discrete or continuous variable.

Table 1: Summary of variables in the dataset.

Variable	**Description**	**Levels**
`event`	Name of event	
`dist`	Length of event in meters	50, 100, 200, 400
`stroke`	Stroke swum in the event	Freestyle, Backstroke, Breastroke, Butterfly, Medley
`sex`	Gender of event participants	W (women), M (men)
`course`	Indication of 25m (short) or 50m (long) pool	Short, Long
`time`	Time taken for one swimmer in the final, in seconds	(Continuous variable)

Figure 1 visualises the distribution of the competitor times. We see that there are four distinct peaks that correspond to the four distance categories in the data. Each successive peak appears to have a greater spread.

Plotting against each variable against every other, as in Figure 2, we note that for the most part there does not appear to be a clear relationship between any of the explanatory variables except for distance and stroke where there are some distances for which there is no competition for a particular stroke. There are clear relationships between distance, stroke, sex, course and time, which we shall explore in closer detail.

Distance appears to explain a large part of the variation in times. Plotting time against distance (Figure 3) suggests a potential linear relationship between `dist` and `time` if we consider `dist` as a continuous variable, and a non-linear relationship if we consider it as a categorical variable.

There are subtle differences in the distribution of times for each level of the other explanatory variables (Figure 4).

A final consideration is the potential normality of the data. The histogram of the competitor times clearly shows that the data is not normal, however, within each distance category, variation appears to be close to normal but with deviation in the tails, as the points follow closely the quantile lines

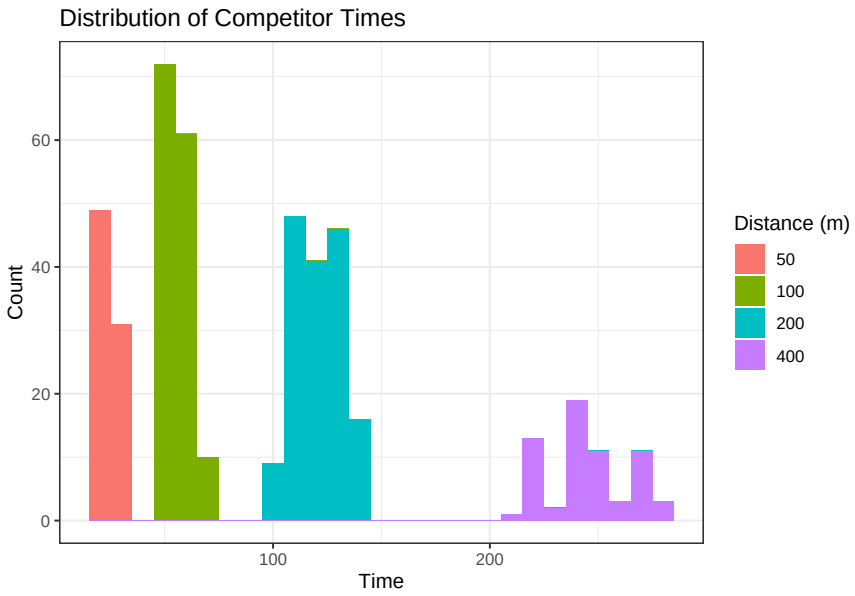

Figure 1: Histogram showing the distribution of competitor times. There are four peaks with increasing spread corresponding to the four distance categories in the data.

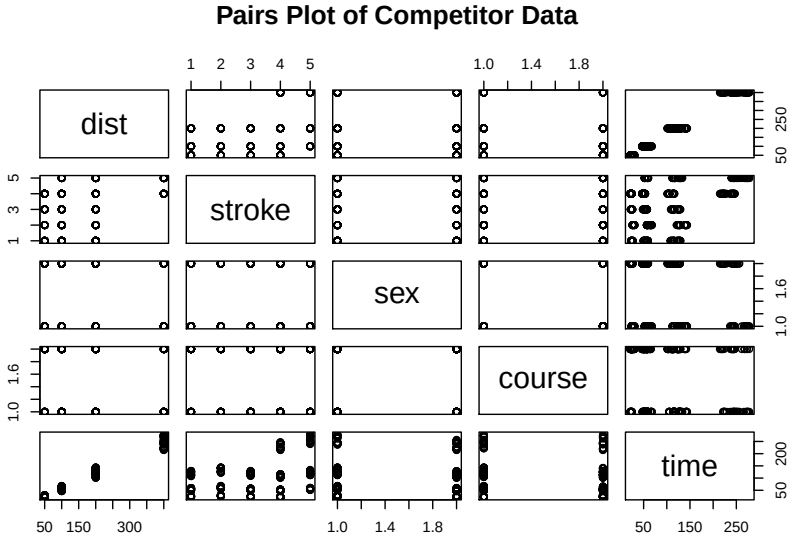

Figure 2: Pairs plot demonstrating the relationship between all variables in the data. There are clear relationships between time and other variables, but mostly unclear associations between other variables.

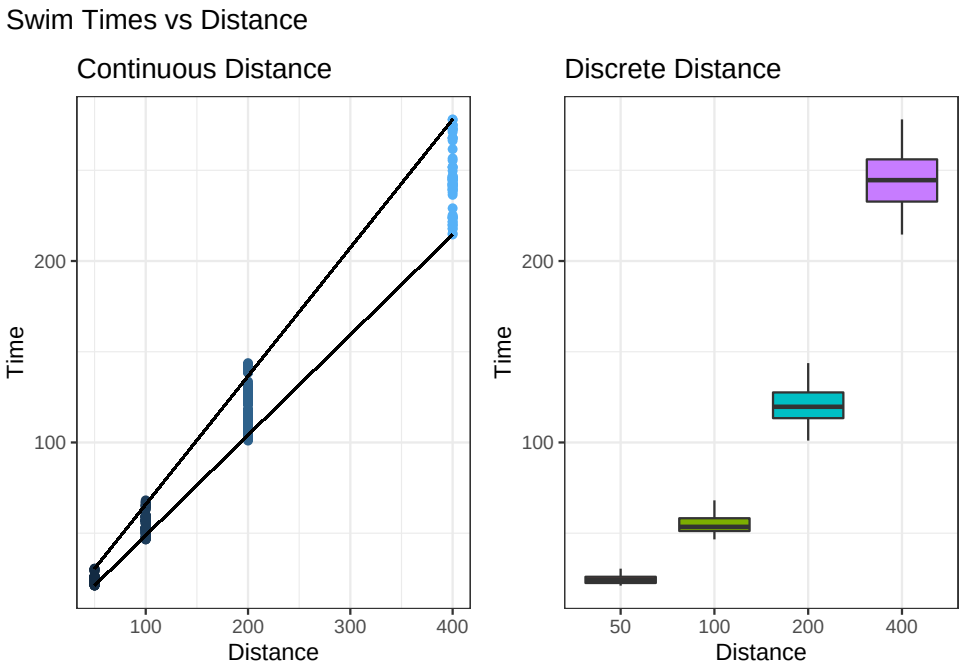

Figure 3: Plots showing the relationship between distance and time. When considered as a continuous variable `dist` suggests a potential linear relationship between the variables; when considered as a categorical variable there is a non-linear relationship.

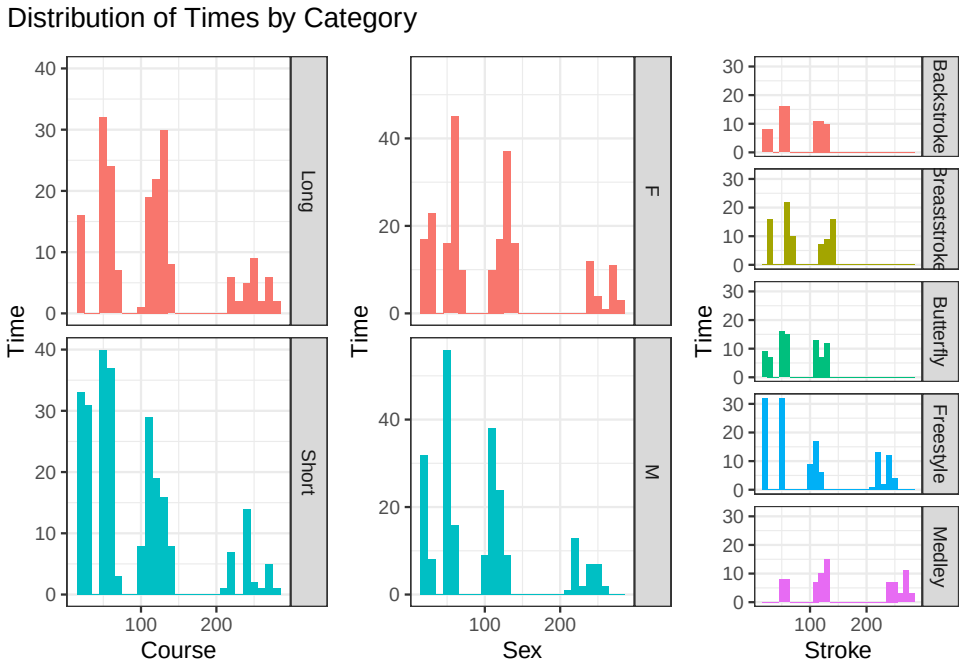

Figure 4: Histograms showing the distributional differences for different factor levels of the categorical explanatory variables in the data. There are subtle differences in distribution at each level.

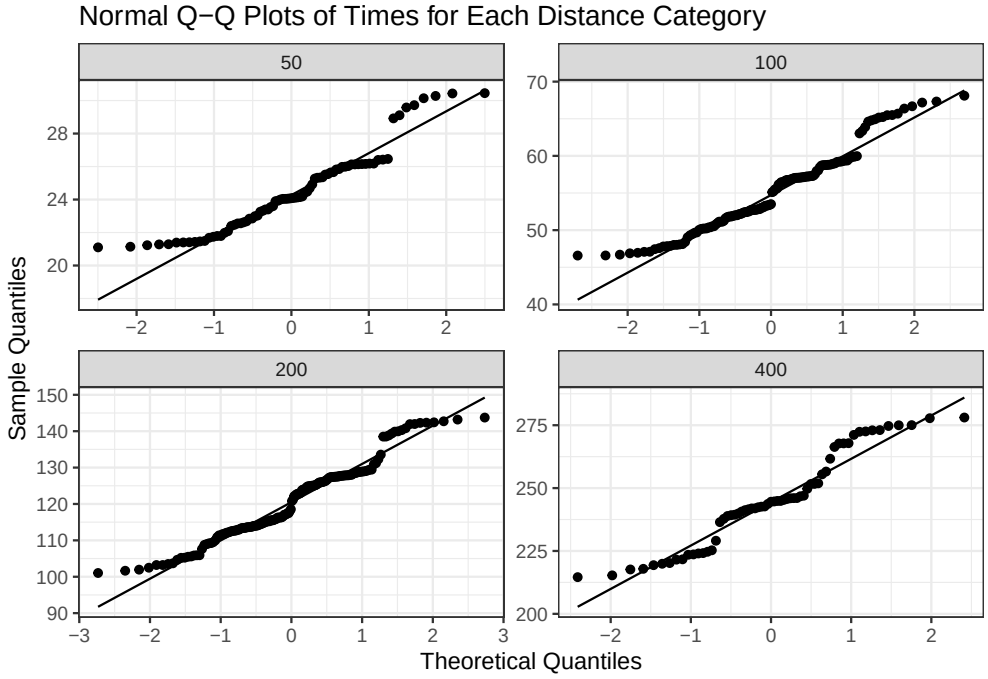

Figure 5: Normal Q-Q plots for each distance category. At each level, the data appears to be approximately normally distributed.

(Figure 5), with residual variation still present, and thus a normal linear model would seem to be a suitable model choice in this context.

3 Methods and Results

3.1 Model Selection

We observed a clear increase in variance with distance, and even fitting the least parsimonious linear model with interaction terms:

$$\texttt{time} \sim \texttt{dist} * \texttt{stroke} * \texttt{sex} * \texttt{course}, \tag{1}$$

the residual variation, as shown in Figure 6, shows a clear association with fitted values which indicates that a normal linear model of this type is not appropriate for the data.

By examining the situation in more detail, we can arrive at a more plausible model. Consider that for a body moving at constant velocity the relationship between the displacement d, velocity v and time elapsed t is given by $t = \frac{d}{v}$. Thus, if a swimmer is moving at a constant speed, there is a linear relationship between displacement and time taken. In this context, the velocity of the swimmer might be related to the characteristics of the swimmer and the race.

Suppose that for every observation i, the velocity is normally distributed such that

$$v_i = \mathbf{x}_i^T \beta + \epsilon_i, \text{ with } \epsilon_i \sim \mathcal{N}(0, \sigma^2), \tag{2}$$

where $\mathbf{x}_i \in \mathbb{R}^p$ is a vector of explanatory variables, $\beta \in \mathbb{R}^p$ is a vector of coefficients and $\epsilon_i \in \mathbb{R}$ is the random error with variance σ^2. Then the time would be modelled by

$$t = \frac{1}{\mathbf{x}_i^T \beta + \epsilon_i} d \iff \frac{1}{t} = (\mathbf{x}_i^T \beta + \epsilon_i) \cdot \frac{1}{d}. \tag{3}$$

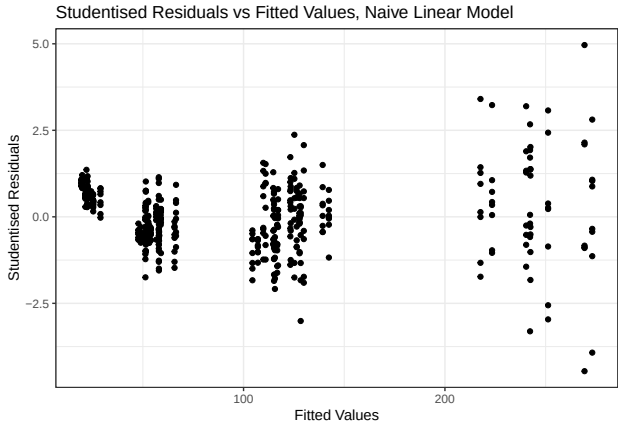

Figure 6: A plot of residuals against fitted values for a linear model with all interaction terms. Even with the inclusion of all terms, there is still a clear increase in residuals with the increase of time.

In this model, $\mathbb{V}\mathbf{ar}[\frac{1}{t}] = \frac{\sigma^2}{d^2}$. Thus for larger values of d the variance in $\frac{1}{t}$ becomes smaller, which seems consistent with our observation that with increasing d and t, since we previously observed a linear relationship, there is also increasing variance.

Alternatively, we could model the reciprocal of the velocity v_i as normally distributed such that

$$\frac{1}{v_i} = \mathbf{x}_i^T\beta + \epsilon_i, \text{ with } \epsilon_i \sim \mathcal{N}(0, \sigma^2), \tag{4}$$

$$\implies t = (\mathbf{x}_i^T\beta + \epsilon_i)d. \tag{5}$$

This means that $\mathbb{V}\mathbf{ar}[t] = d^2\sigma^2$, which again is plausible given the increasing variance with t and d.

Both (3) and (5) are weighted normal linear models. We therefore fit the following models to the data:

$$\frac{1}{\texttt{time}} \sim (\texttt{stroke} + \texttt{sex} + \texttt{course}) * \frac{1}{\texttt{distance}}, \;\; \texttt{weight}_i = \texttt{distance}_i; \tag{6}$$

$$\texttt{time} \sim (\texttt{stroke} + \texttt{sex} + \texttt{course}) * \texttt{distance}, \;\; \texttt{weight}_i = \frac{1}{\texttt{distance}_i}; \tag{7}$$

the variance for observation i is given by $(\sigma/\texttt{weight}_i)^2$. Looking at plots of residuals against fitted values in Figure 7, we see that that there is far less of a noticeable pattern in the residuals than model (1), but that there are still four clusters with distinct shapes suggesting that we still do not have homoskedasticity.

The final model that we could consider is considering distance as a discrete variable, which necessitates a transformation applied to the response variable since there is a non constant variance in the distance. Using the model in (1) (but with distance a factor), we can use a profile likelihood to determine a suitable λ for a suitable Box-Cox transformation given by:

$$t^{(\lambda)} = \begin{cases} \frac{t^\lambda - 1}{\lambda}, \text{ if } \lambda \neq 0 \\ \log(t), \text{ if } \lambda = 0. \end{cases} \tag{8}$$

Figure 8, shows a graph the profile likelihood of λ and a 95% confidence interval for its value. For interpretability, we shall choose $\lambda = 0$ for the response variable transformation using the graph.

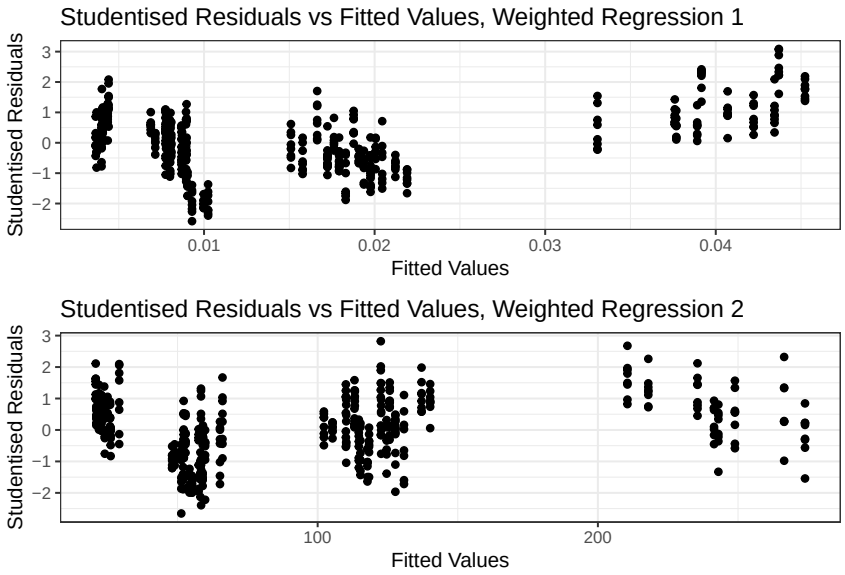

Figure 7: Residuals plotted against fitted values for both weighted least squares models. In both models, we see that there appears to be less dependence on the fitted value, but that the dependency has not completely been eliminated.

Table 2: RSS values for the proposed models structures, only regressing against |distl.

Normal Linear (1)	**Weighted 1** (3)	**Weighted 2** (5)	**Box-Cox Transformation** (9)
42740	64380	42740	42350

Fitting the following model:

$$\log(\texttt{time}) \sim \texttt{dist} * \texttt{stroke} * \texttt{sex} * \texttt{course} \tag{9}$$

Plotting residuals against fitted values (Figure 9), we see that residuals seem to be independent of fitted values.

It remains to compare the models proposed to determine which is the most suitable for the data. In Table 2 we see the residual sum of squares (RSS) values for the four proposed models structures, only regressing against distance so that parsimony is comparable and since distance provides the greatest contribution towards variance in the response. In terms of interpretability, the transformation suggested by the Box-Cox method is sensical since swimmers tiring in longer races, amongst other factors, would mean that it is more appropriate to model the distance categorically and in a non-linear fashion. We see that Box-Cox transformation has the smallest RSS which would indicate a potentially superior fit. The Box-Cox transformation eliminates the pattern in the studentised residuals, has residuals that appear to conform most closely to a normal distribution (Figure 10) so we shall use this model, with $\lambda = 0$, to further analyse the data. Also, note that the predictions required only require the specification of distances at existing factor levels, so the factorial treatment of `dist` is not an issue here.

3.2 Variable Selection

Since there are somewhat numerous possible combinations of inclusion and exclusion of variables and interactions from the model, we shall use automatic model selection using the Akaike

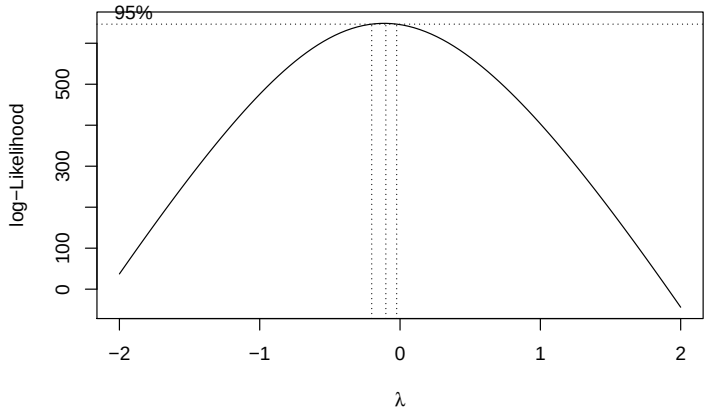

Figure 8: A graph of the profile likelihood of λ, the Box-Cox transform parameter, for a normal linear model of the data. The nearest interpretable value of λ to the maximum likelihood estimate and 95% interval is 0.

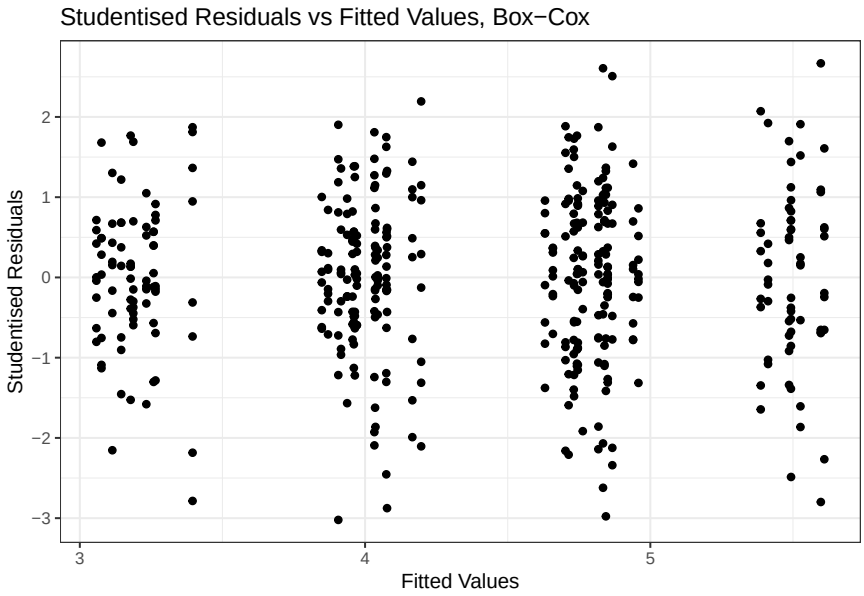

Figure 9: Residuals plotted against fitted values for the Box-Cox transformation model. We see that the dependence on the fitted values seems to have been removed with the transformation.

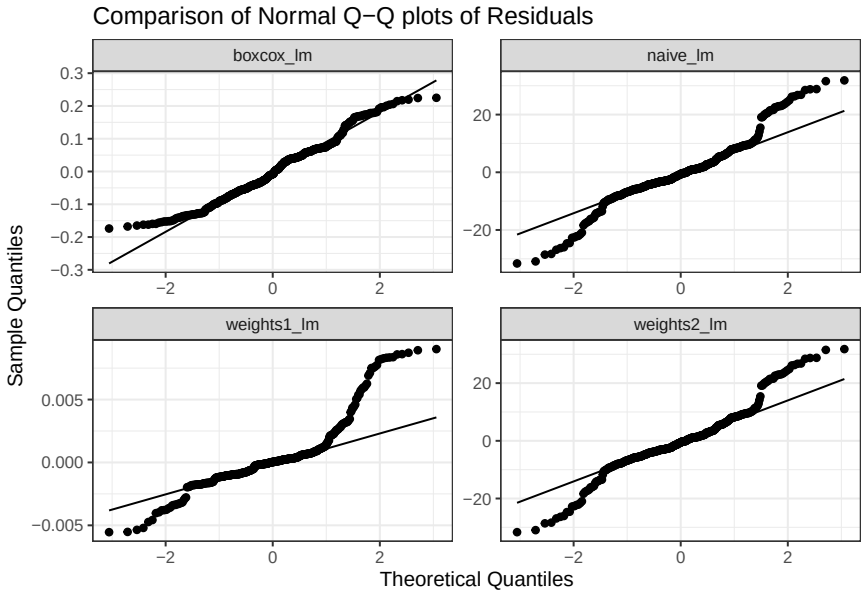

Figure 10: Comparison of Normal Q-Q plots of the residuals for the four proposed models. The Box-Cox transformation produces residuals that are the closest to normally distributed.

information criterion to search for possible candidates for models that balance fit with parsimony. Using a minimal model of

$$\log(\texttt{time}) \sim 1, \tag{10}$$

and a maximal model of

$$\log(\texttt{time}) \sim \texttt{dist} * \texttt{stroke} * \texttt{sex} * \texttt{course}, \tag{11}$$

we use forward and backward selection to search for a suitable model. Both methods returned the model

$$\begin{aligned} \log(\texttt{time}) \sim & \texttt{dist} + \texttt{stroke} + \texttt{sex} + \texttt{course} \\ & + \texttt{dist} * \texttt{stroke} + \texttt{dist} * \texttt{sex} \\ & + \texttt{stroke} * \texttt{sex} + \texttt{stroke} * \texttt{course} \\ & + \texttt{sex} * \texttt{course}. \end{aligned} \tag{12}$$

There model is fairly interpretable having only second order interactions.

Looking at the ANOVA tables (available in Section 5.2) for the model in (12) and for the saturated model (9), we see that the third order interactions do not seem to explain a significant amount of variance and that the interaction $\texttt{dist} * \texttt{course}$ is also not significant.

3.3 Outlier Detection

For this section, we shall be using the model found by automatic selection to try to identify outliers.

The largest leverage in the model was 0.0858; typically, we treat points with a leverage greater than $\frac{2p}{n}$ as unusual, however, in this fit, this quantity is given by 0.152, so there appear to be no points of concern with regards to leverage.

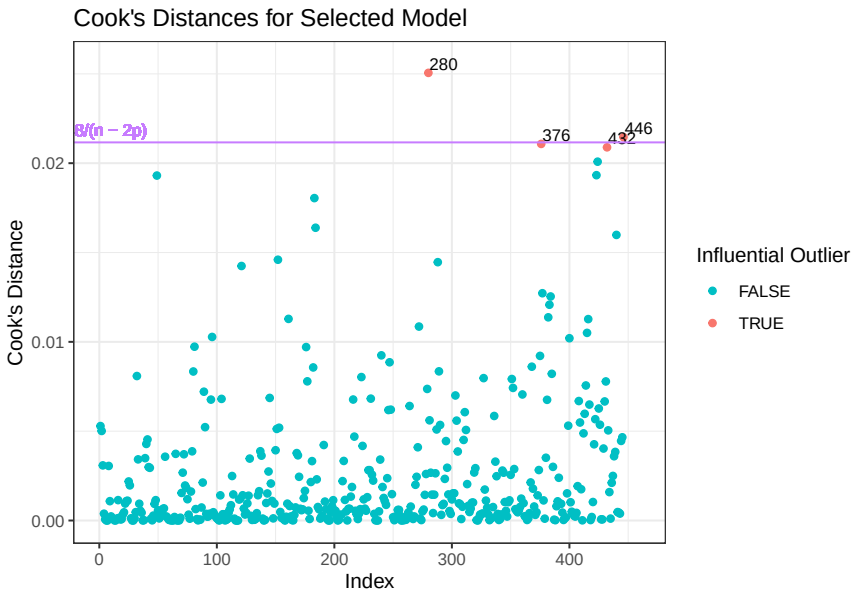

Figure 11: A plot of Cook's distances for the selected model. Points approximately above the $\frac{8}{n-2p}$ threshold are potentially problematic.

Cook's distances $C_k \gtrsim \frac{8}{n-2p}$ are potential causes for concern and may require further examination. Entries indexed 280 and 446 are greater than the threshold and thus influential outliers, and 376 and 446 are slightly below the bound so could also potentially be problematic data points (Figure 11). Looking at Figure 12, we can see that points with large Cook's distances look to be the points that deviate the most from the central part of the distribution for each factor level of distance. The deviation does not seem to be excessively large compared to the other points or unusual to warrant exclusion from the dataset, thus we shall include these points for the remainder of the analysis.

3.4 Model Interpretation

Table 3 shows estimated coefficients for the model in (12). The response is $\log(\texttt{time})$, thus $\exp(\hat{\beta}_{\text{intercept}})$, where $\beta_{\text{intercept}}$ is the coefficient of the intercept, can be interpreted as the maximum likelihood estimate time taken for a female swimming a 50m backstroke race on a long course since these are the baseline levels for each of the factors. Multiplying the baseline by $\exp(\sum_{i \in I} \hat{\beta}_i)$, where β_i is the coefficient for the ith term in the model, gives the maximum likelihood estimate for the time taken for the combination of factors specified by the collection of variables and interactions in I. Table 4 shows the factor multiplying the baseline for a particular variable/interaction.

There is evidence that the individual explanatory variables have a significant effect on the time taken from the baseline level. As expected, the model suggests that increasing distance means increasing times, as the coefficients are positive. Looking at the coefficients for stroke, for 50m the model suggests that breaststroke is slower than backstroke and that butterfly and freestyle are slightly faster. The variable "strokeMedley" does not really have an interpretation alone since the medley is only done for distances greater than 50m. There is also evidence to suggest that times by males and from short courses are also smaller.

The significant interactions between distance and stroke suggest that times do not increase uniformly across the strokes but that different strokes have different increases in times with distance,

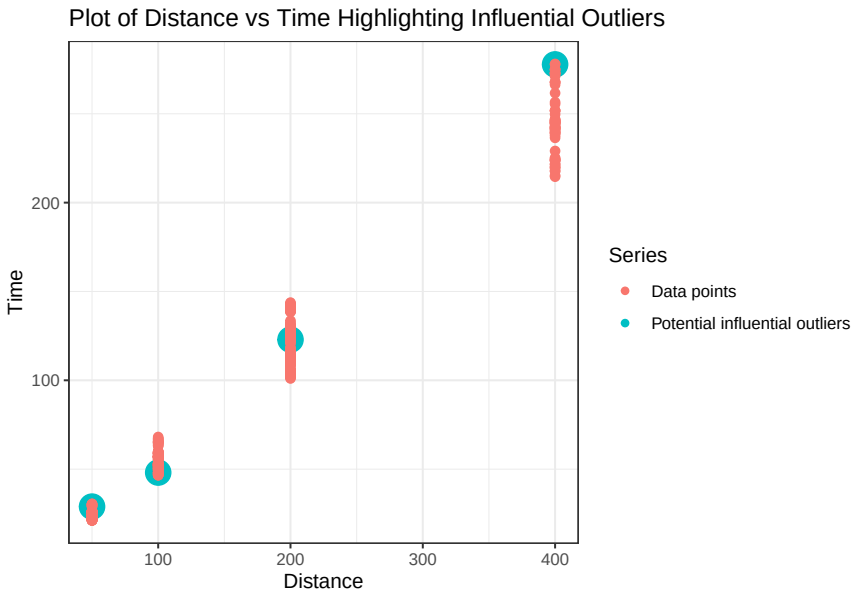

Figure 12: A plot of distance vs time measurements in the dataset with potential influential outliers highlighted. The points do not seem to be out of place for the distribution and so will be included for the remainder of the analysis.

apart from breaststroke which does not appear not to show significant differences from backstroke.

The interaction between distance and sex suggests that times do not increase similarly for each sex and that there is evidence that males have greater increases in time with increasing distance.

The interaction between stroke and sex do not appear particularly significant, but there is some evidence to suggest that after accounting for the overall difference between times for men and women, that men then are slightly slower at freestyle.

Finally, there appears to be significant differences between all the strokes from the backstroke baseline with regard to the differences in time between short and long courses.

3.5 Predictions

In this section we use the selected model to produce prediction intervals. First, we create an interval for predicted $\log(\texttt{time})$, $[\log(\text{PI}_{\text{lower}}), \log(\text{PI}_{\text{upper}})]$ and exponentiate this to produce $[\text{PI}_{\text{lower}}, \text{PI}_{\text{upper}}]$. The predictions and intervals are given in Table 5.

4 Conclusions

We have analysed trends in competitor times in swimming races from the 2016 Olympics and 2016 World cup using a Box-Cox transformation to deal with the heteroscedastic data, concluding that this was a more suitable model type than just a normal linear model or a weighted regression. Using the Akaike information criterion we selected a model that adequately explains the variation in times but is still parsimonious. After looking for outliers, we found that there were no data points that needed to be excluded. Finally, we interpreted the fitted model and used it to produce predictions.

Table 3: Estimated values for coefficients of the selected model from (12). ***, **, * and . indicate significance at the 0.001, 0.01, 0.05 and 0.1 levels respectively using a T-test against null hypothesis $\mathcal{H}_0 : \beta_i = 0$, where β_i is the coefficient for this ith variable/interaction.

Variable/Interaction	$\hat{\beta}$	SE$[\beta]$	T statistic	$\mathbb{P}[\lvert t\rvert > 0]$	Signif.
(Intercept)	3.3	0.00465	710	0	***
dist100	0.768	0.00426	180	0	***
dist200	1.55	0.00426	363	0	***
dist400	2.3	0.00541	425	0	***
strokeBreaststroke	0.113	0.0062	18.2	5.97e-55	***
strokeButterfly	-0.0562	0.00622	-9.04	6.1e-18	***
strokeFreestyle	-0.117	0.0051	-23	4.03e-76	***
strokeMedley	0.0127	0.00415	3.07	0.00232	**
sexM	-0.114	0.00409	-27.9	2.67e-97	***
courseShort	-0.0349	0.00318	-11	7.48e-25	***
dist100*strokeBreaststroke	0.00782	0.00553	1.42	0.158	
dist200*strokeBreaststroke	-0.00303	0.00553	-0.549	0.583	
dist100*strokeButterfly	0.0242	0.00553	4.38	1.52e-05	***
dist200*strokeButterfly	0.0468	0.00553	8.47	4.38e-16	***
dist100*strokeFreestyle	0.0148	0.0049	3.02	0.0027	**
dist200*strokeFreestyle	0.0133	0.0049	2.72	0.00685	**
dist400*strokeFreestyle	0.0115	0.00573	2.01	0.0447	*
dist100*strokeMedley	0.00989	0.0049	2.02	0.0443	*
dist100*sexM	0.00654	0.0034	1.93	0.0547	.
dist200*sexM	0.012	0.0034	3.52	0.000475	***
dist400*sexM	0.0222	0.00444	5	8.67e-07	***
strokeBreaststroke*sexM	-0.00725	0.00374	-1.94	0.053	.
strokeButterfly*sexM	0.00166	0.00375	0.442	0.659	
strokeFreestyle*sexM	0.00907	0.00351	2.58	0.0101	*
strokeMedley*sexM	0.00321	0.00405	0.794	0.428	
strokeBreaststroke*courseShort	0.0124	0.00418	2.97	0.0032	**
strokeButterfly*courseShort	0.0237	0.0042	5.66	2.87e-08	***
strokeFreestyle*courseShort	0.0238	0.00362	6.57	1.52e-10	***
strokeMedley*courseShort	0.0164	0.0042	3.9	0.000113	***
sexM*courseShort	-0.0124	0.00234	-5.31	1.83e-07	***

Table 4: Transformed values for the coefficient estimates.

Variable/Interaction	$\exp(\hat{\beta})$
(Intercept)	27.11
dist100	2.155
dist200	4.711
dist400	9.974
strokeBreaststroke	1.12
strokeButterfly	0.9454
strokeFreestyle	0.8896
strokeMedley	1.013
sexM	0.8923
courseShort	0.9657
dist100*strokeBreaststroke	1.008
dist200*strokeBreaststroke	0.997
dist100*strokeButterfly	1.024
dist200*strokeButterfly	1.048
dist100*strokeFreestyle	1.015
dist200*strokeFreestyle	1.013
dist400*strokeFreestyle	1.012
dist100*strokeMedley	1.01
dist100*sexM	1.007
dist200*sexM	1.012
dist400*sexM	1.022
strokeBreaststroke*sexM	0.9928
strokeButterfly*sexM	1.002
strokeFreestyle*sexM	1.009
strokeMedley*sexM	1.003
strokeBreaststroke*courseShort	1.012
strokeButterfly*courseShort	1.024
strokeFreestyle*courseShort	1.024
strokeMedley*courseShort	1.017
sexM*courseShort	0.9877

Table 5: Prediction values and estimates using the selected model.

Name	`dist`	`stroke`	`sex`	`course`	$\log(t_{\text{predicted}})$	$\log(\text{PI}_{\text{lower}})$	$\log(\text{PI}_{\text{upper}})$
RaceA	400	Freestyle	F	Long	5.495	5.471	5.519
RaceB	50	Backstroke	F	Long	3.303	3.279	3.328
RaceC	400	Butterfly	F	Long	5.544	5.518	5.571
RaceD	100	Medley	F	Long	4.095	4.07	4.119

Name	`dist`	`stroke`	`sex`	`course`	$t_{\text{predicted}}$	PI_{lower}	PI_{upper}
RaceA	400	Freestyle	F	Long	243.5	237.7	249.4
RaceB	50	Backstroke	F	Long	27.21	26.54	27.9
RaceC	400	Butterfly	F	Long	255.8	249	262.8
RaceD	100	Medley	F	Long	60.01	58.53	61.53

5 Appendix

5.1 Suppementary Code

Suplementary code may be found via the following link:

https://github.com/ThaliaSeale/Week-4-Marked-Practical

5.2 Analysis of Variance Tables

5.2.1 ANOVA of Saturated Model

```
## 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
```

5.2.2 ANOVA of Selected Model

```
## 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
```