Statisical analysis of whistles' data

David Castro, Zaigham Abbas

Wednesday, 12/12/2018

Contents

List of Figures

List of Tables

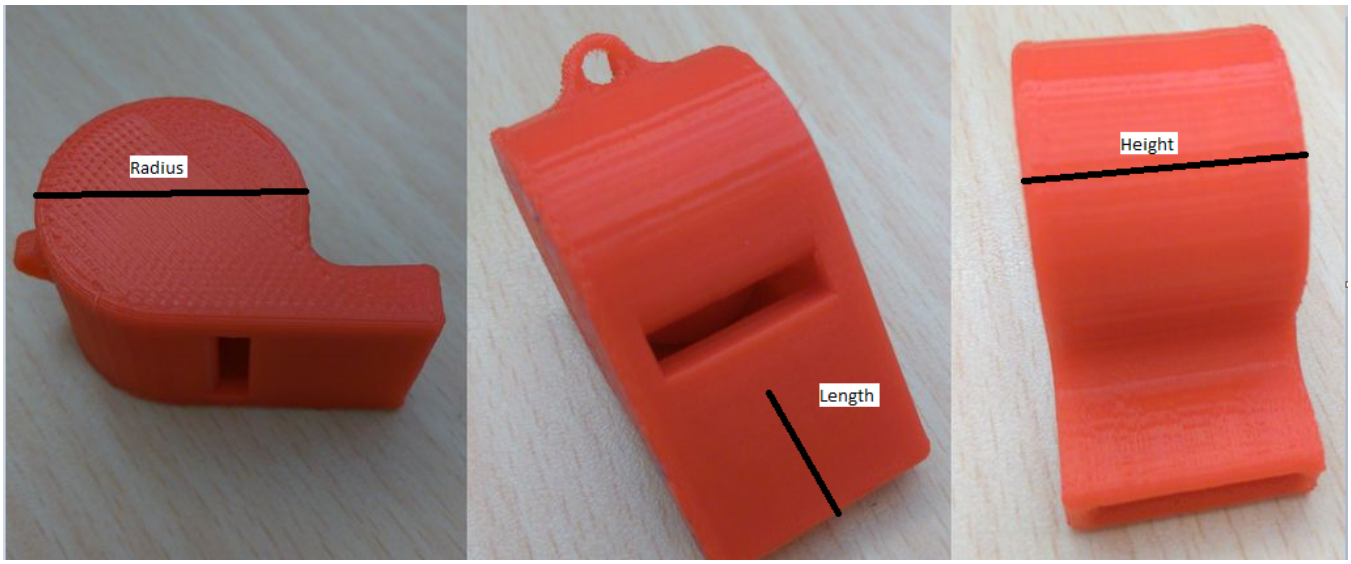

Figure 1: Explanatory variables

Summary

This experiment was conducted to help us determine the effect of the radius (R), length (L), and the height (H) of whistles on its loudness (please refer to figure 1). For our experiment, we used 3 levels for the radius, and 2 levels each for the height and width, making our experiment a $3 * 2^2$ design. Please refer to the table 1 below to find more details on the study design and the encoding of the factor levels. The response variable was measured in Decibels (Db). There were in total 12 treatment combinations of the from RLH (a treatment combination of 3.1.1 implies a radius of 40 mm, a length of 25 mm, and a height of 17 mm). A cell means model, followed by Tukey multiple comparisons, was used to help us find the best treatment combinations. In total 84 observations were collected to detect a difference of 25 Db at a power of 0.8 and at a significance level of 95 %.

Table 1: Factor levels (in mm) and level codes

Levels	R	L	H
1	20	25	17
2	30	45	28
3	40	-	-

The best treatment combination for the loudest whistle was 3.2.2.

Experiment procedures

Pilot Experiment

Before we were able to carry our full experiment, a pilot experiment had to be commenced in order to test our methods and see the possible errors that are expected to appear during the full experiment. The end goal of this pilot experiment was to figure out which statistic between mean, maximum, variance would be ideal for analysing the loudness of the whistle. This was followed by some checks to see if there was a whistle effect (if 2 whistles of the same dimensions would have significantly different responses) or a human effect (if there was some difference in our response variable if 2 people blew the same type of whistle).

In the pilot experiment, 24 whistles of the same dimensions (R = 30mm, L = 35mm, H = 20mm) were printed and 2 of them were given to 12 human subjects each. Each subject blew into the whistle for around 3 seconds and the results were recorded. A mobile application named "decibel X" was used to record the response variable at each blow. It was quickly discovered that there was an issue in the recording: the beginning and the end of the recording had a drastic increase and decrease in the response respectively. Since we were only interested in the loudness, only the recording around the maximum was considered for further analysis and the rest was truncated out.

From the table 2 below, you can see that the maximum was associated with lesser variance than the mean. Hence, we came to the conclusion that it would be best to use the maximums. In the further report, we will be using both maximum and loudness interchangeably to refer to the output response of the whistles.

Table 2: Correlations between mean, max, and variance

	Mean	Max.	V
Mean	1.00	0.99	-0.37
Max.	0.99	1.00	-0.27
V	-0.37	-0.27	1.00

The Figure 2 shows the maximums distributed over the subjects and whistles. As you can see, there was a high amount of variance associated with the subjects. Hence, the human effect was introducing significant variance and having multiple subjects blow whistles was not ideal. Additionally, there was no pattern to the distribution of whistles 1 and 2 at each subject level. Hence, we can conclude that the whistle effect was not significant.

After this, we calculated the variance of the maximums. It was 79.96 Db^2 with a 90% confidence

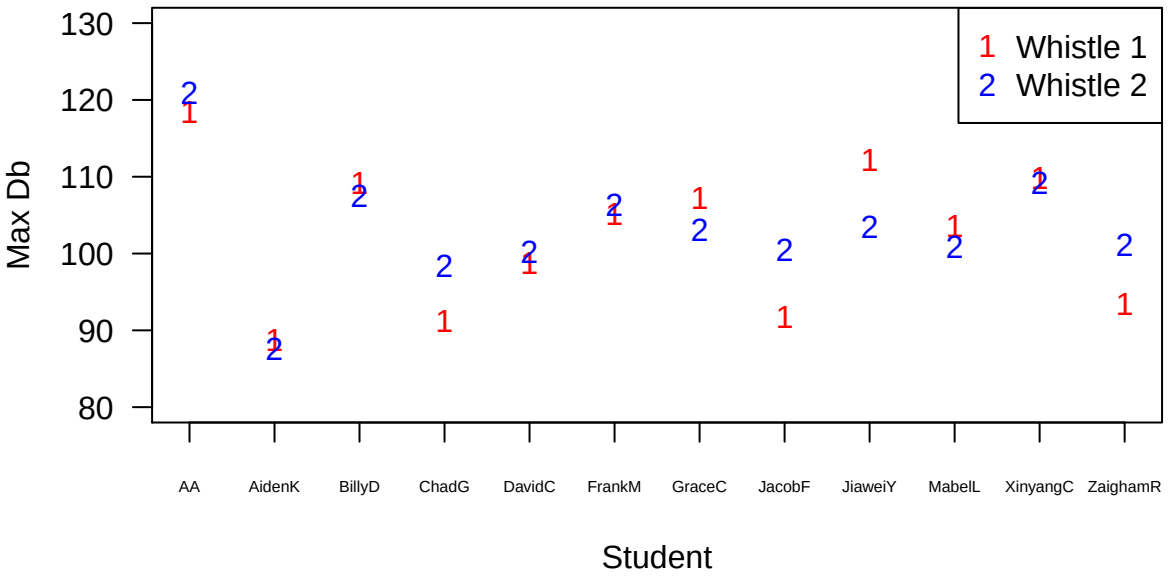

Figure 2: Maxiums vs subjects and whistles

bound at 113.01 Db^2. We used this variance to do a power test at a power of 0.8 and at a significance level of 95 % for various differences in loudness. Table 3 represent this information. The rows represent the differences and the columns represent the total number of treatment combinations. The entries inside the table represent the total number of recordings to be made for that particular experimental design. We needed to begin with a mimimum of 8 treatment combinations because we were trying to analyse 3 different factors with atleast 2 levels each (2^3=8).

Table 3: Sample Sizes for different v and Delta with 90% upper confidence bound

	8	12	18
15	128	216	378
20	80	132	216
25	56	84	144
30	40	72	108

It was believed that having 12 treatment combinations to detect a difference of 25 Db was the most cost and time effective approach. Hence, a 2^2*3 experiment was designed. R was given 3 levels, while L and H were given 2 levels each.

Full-Scale Experiment

Each of the 12 whistle treatment combinations was 3D-printed once and blown 7 times and recorded with the same app. All the whistles were printed in the same 3D-printer with the same kind of plastic in order to avoid machine related variance. The maximum from the recording was extracted and used for further analysis. In order to control for the human effect, only one subject blew into the whistles.

There were issues with the way our data was collected. Firstly, the recordings were not done in a random fashion for the different factor levels. Instead, all of the recordings for each treatment combination were conducted together; The recording sequences for different treatment combinations weren't mixed. For example, all of the 7 recordings for the treatment combination 1.1.1 were collected consecutively instead of collecting once for 1.1.1 and then randomly choosing a treatment combination to do the next recording. Secondly, we believe that there was a lot of human error involved. A human can't consistently blow into 84 whistles in a row. The 1st blow will be very different from the last one.

Statistical Methods

We at first constructed a 3-way complete anova model of the form:

$$Y_{ijkt} = \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk} + \epsilon_{ijkt}$$

$$i = 1, \ldots, 3, j = 1, \ldots, 2, k = 1, \ldots, 2$$

$$\epsilon_{ijkt} \text{ are mutually independent, and } \epsilon_{ijkt} \sim N(0, \sigma^2)$$

The results from this model are as follows:

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## R            2   78.2    39.1   2.572  0.08336 .
## L            1  113.9   113.9   7.487  0.00782 **
## H            1   80.8    80.8   5.315  0.02403 *
## R:L          2   71.7    35.8   2.356  0.10206
## R:H          2   10.2     5.1   0.336  0.71597
## L:H          1    0.0     0.0   0.001  0.97776
## R:L:H        2  745.3   372.6  24.501 7.65e-09 ***
## Residuals   72 1095.0    15.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the model, the 3-way interaction (R:H:L) was significant along with H and L only at a significance level of 95%. Because of the significance of the 3-way interaction, further analysis to help us find the best treatment combination couldn't be carried out with this model. Therefore, we decided to fit a cell-means model of the form below and do Tukey multiple comparisons to help us find the best treatment combination:

$$Y_{ijkt} = \tau_{ijk} + \epsilon_{ijkt}$$

$$i = 1, \ldots, 3, j = 1, \ldots, 2, k = 1, \ldots, 2$$

$$\epsilon_{ijkt} \text{ are mutually independent, and } \epsilon_{ijkt} \sim N(0, \sigma^2)$$

At the same time, there was a problem with the model assumptions. Please refer to Figure 4. The "qq-plot" seems fine,hence we know that it passes the normality check. From the "Residuals vs treatments" and "Residuals vs perdicted" graphs, we know that it passes the fit of the model and the outliers checks. However the "Residuals vs order" plot kind of contricts on the right side. Additionally, the ratio of the maximum and minimum variance among the residuals divided over the treatment combinations was around 102.7. Hence, the model failed the independance of the error terms and equal variance assumptions. We tried various adjustments with our data (log, square root, cubic root, and Tukey-adjustment), but couldn't resolve the issues. Therefore, we decided to continue on with our current model.

Results

The results from the cell means model are as follows:

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## RLH         11   1100  100.01   6.576 1.81e-07 ***
## Residuals   72   1095   15.21
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the cell-means anova, we can conclude that at least one of the RLH levels had a significantly different impact on the loudness as compared to the others at a significance level of 99.9%. We will use Tukey pairwise multiple-comparisons to help us find the best treatment combination at a significance level of 95% (Scheffe was another choice, but we preferred tighter confidence bounds). Figure 3 visualizes the results from the multiple comparisons. The blue bars are confidence intervals for the treatment combinations, and the red arrows are for the comparisons among them. If an

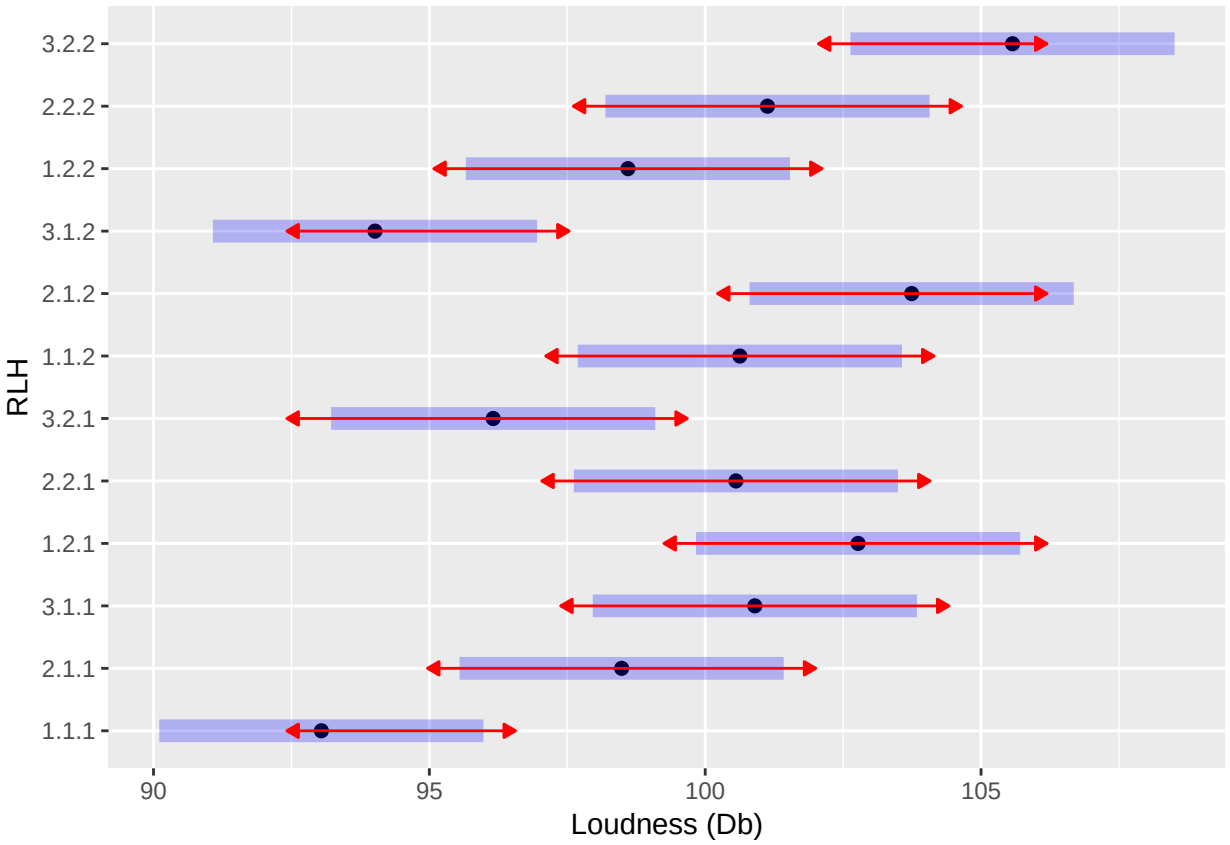

Figure 3: Multiple comparisons' plot

arrow from one level overlaps an arrow from another, the difference is not significant. A detailed chart with Tukey's pairwise comparison is in the appendix (Table 5).

From figure 3, it can be seen that the treatment combination 1.1.1 has the lowest average, while 3.2.2 has the highest average. There are also a lot of overlaps between various intervals. Table 4 summarises these results in another way. In this table, treatment combinations whose maximum loudness are not significantly different from each other are put in a single group. Please note that a treatment combination might fall into more than one group. A lower group number implies that its members generally have a lower loudness than the groups above it (hence those in group 1 have a lower loudness than those in group 5). However, treatment combinations may lie in more than one group. Interestingly, treatment combinations 2.2.2 and 1.2.1 all lie in the same groups (i.e. 345) as well as 2.2.1, 1.1.2, and 3.1.1 (i.e. 2345).

Table 4: Treatment combination groups

Treatment Combination	Group
1.1.1	1
3.1.2	12
3.2.1	123
2.1.1	1234
1.2.2	12345
2.2.1	2345
1.1.2	2345
3.1.1	2345
2.2.2	345
1.2.1	345
2.1.2	45
3.2.2	5

Conclusions

From the multiple comparisons' table in the appendix (table 5), it can be seen that the treatment combination 1.1.1 was the worst: its loudness was significantly different than most of the other whistles and was less than them. Hence, 1.1.1 shouldn't be relied upon to deliver the loudest sound. From table 4 it can be seen that it is in the same group as 3.1.2, 3.2.1, 2.1.1, and 1.2.2. Hence, we will consider these treatment combinations to be insufficient as well. Furthermore, as mentioned before, group 5 had the loudest whistles (i.e. treatment combinations 1.2.2, 2.2.1, 1.1.2, 3.1.1, 2.2.2, 1.2.1, 2.1.2, 3.2.2). Statistically, we aren't allowed to differentiate between the whistles in this group because they are not significantly different. Hence, if we want to stick to the strictest statistical procedures then the whistles with the dimensions in group 5 will give the maximum loudness.

Please keep in mind that in our analysis we didn't control for human temporal error and that contributed to a lot of variance (i.e. we couldn't control how one might be blowing whistles at a given time). Furthermore, there was a lot of overlap within the groups from table 4 (for example the treatment combination 1.2.2 lies in both groups 1 and 5). So in real life terms, we believe that such ambiguity in our results and errors in our data collection allowed us to play around with the rules a bit.

The treatment combination 3.2.2 lies in only group 5 (the maximum group). Other elements of group 5 also lie in other groups which are categorically supposed to have less loudness than group 5. Being

included in other groups raises questions on the credibility of these other whistles at delievering the loudest sound. However, 3.2.2 isn't hampered by such techniaclities. Additionally from group 5, 3.2.2's response is significantly higher than 4 other treatment combinations' responses outside of group 5; no other element of group 5 has as high of such a score.

Therefore, after the experimental analysis, we will recommend treatment combination 3.2.2 (a radius = 40mm, length = 45mm, and height = 28mm) for the maximum loudness.

Future Analysis

The main problem was with the data collection procedures. Please follow our recommendations for any future data collections. Firstly, as we mentioned in the beginning, we didn't collect the data in a random fashion. A random number generator could be used to help randomise the collection of the data. Secondly, if an air pump could be utilised that could help provide constitency to blows provided to the whistles, that would be great. Such a mechanism will allow us to control for human temporal errors.

Appendix

Table 5: Pairwise comparison's by Tukey's method

Contrast	Lower CI Limit	Upper CI Limit	Significance
1.1.1 - 2.1.1	-12.484	1.598	
1.1.1 - 3.1.1	-14.898	-0.816	***
1.1.1 - 1.2.1	-16.770	-2.687	***
1.1.1 - 2.2.1	-14.556	-0.473	***
1.1.1 - 3.2.1	-10.156	3.927	
1.1.1 - 1.1.2	-14.627	-0.544	***
1.1.1 - 2.1.2	-17.741	-3.659	***
1.1.1 - 3.1.2	-8.013	6.070	
1.1.1 - 1.2.2	-12.598	1.484	
1.1.1 - 2.2.2	-15.127	-1.044	***
1.1.1 - 3.2.2	-19.570	-5.487	***
2.1.1 - 3.1.1	-9.456	4.627	
2.1.1 - 1.2.1	-11.327	2.756	
2.1.1 - 2.2.1	-9.113	4.970	
2.1.1 - 3.2.1	-4.713	9.370	
2.1.1 - 1.1.2	-9.184	4.898	
2.1.1 - 2.1.2	-12.298	1.784	
2.1.1 - 3.1.2	-2.570	11.513	
2.1.1 - 1.2.2	-7.156	6.927	
2.1.1 - 2.2.2	-9.684	4.398	
2.1.1 - 3.2.2	-14.127	-0.044	***
3.1.1 - 1.2.1	-8.913	5.170	
3.1.1 - 2.2.1	-6.698	7.384	
3.1.1 - 3.2.1	-2.298	11.784	
3.1.1 - 1.1.2	-6.770	7.313	
3.1.1 - 2.1.2	-9.884	4.198	
3.1.1 - 3.1.2	-0.156	13.927	
3.1.1 - 1.2.2	-4.741	9.341	
3.1.1 - 2.2.2	-7.270	6.813	
3.1.1 - 3.2.2	-11.713	2.370	
1.2.1 - 2.2.1	-4.827	9.256	
1.2.1 - 3.2.1	-0.427	13.656	
1.2.1 - 1.1.2	-4.898	9.184	

Contrast	Lower CI Limit	Upper CI Limit	Significance
1.2.1 - 2.1.2	-8.013	6.070	
1.2.1 - 3.1.2	1.716	15.798	***
1.2.1 - 1.2.2	-2.870	11.213	
1.2.1 - 2.2.2	-5.398	8.684	
1.2.1 - 3.2.2	-9.841	4.241	
2.2.1 - 3.2.1	-2.641	11.441	
2.2.1 - 1.1.2	-7.113	6.970	
2.2.1 - 2.1.2	-10.227	3.856	
2.2.1 - 3.1.2	-0.498	13.584	
2.2.1 - 1.2.2	-5.084	8.998	
2.2.1 - 2.2.2	-7.613	6.470	
2.2.1 - 3.2.2	-12.056	2.027	
3.2.1 - 1.1.2	-11.513	2.570	
3.2.1 - 2.1.2	-14.627	-0.544	***
3.2.1 - 3.1.2	-4.898	9.184	
3.2.1 - 1.2.2	-9.484	4.598	
3.2.1 - 2.2.2	-12.013	2.070	
3.2.1 - 3.2.2	-16.456	-2.373	***
1.1.2 - 2.1.2	-10.156	3.927	
1.1.2 - 3.1.2	-0.427	13.656	
1.1.2 - 1.2.2	-5.013	9.070	
1.1.2 - 2.2.2	-7.541	6.541	
1.1.2 - 3.2.2	-11.984	2.098	
2.1.2 - 3.1.2	2.687	16.770	***
2.1.2 - 1.2.2	-1.898	12.184	
2.1.2 - 2.2.2	-4.427	9.656	
2.1.2 - 3.2.2	-8.870	5.213	
3.1.2 - 1.2.2	-11.627	2.456	
3.1.2 - 2.2.2	-14.156	-0.073	***
3.1.2 - 3.2.2	-18.598	-4.516	***
1.2.2 - 2.2.2	-9.570	4.513	
1.2.2 - 3.2.2	-14.013	0.070	
2.2.2 - 3.2.2	-11.484	2.598	

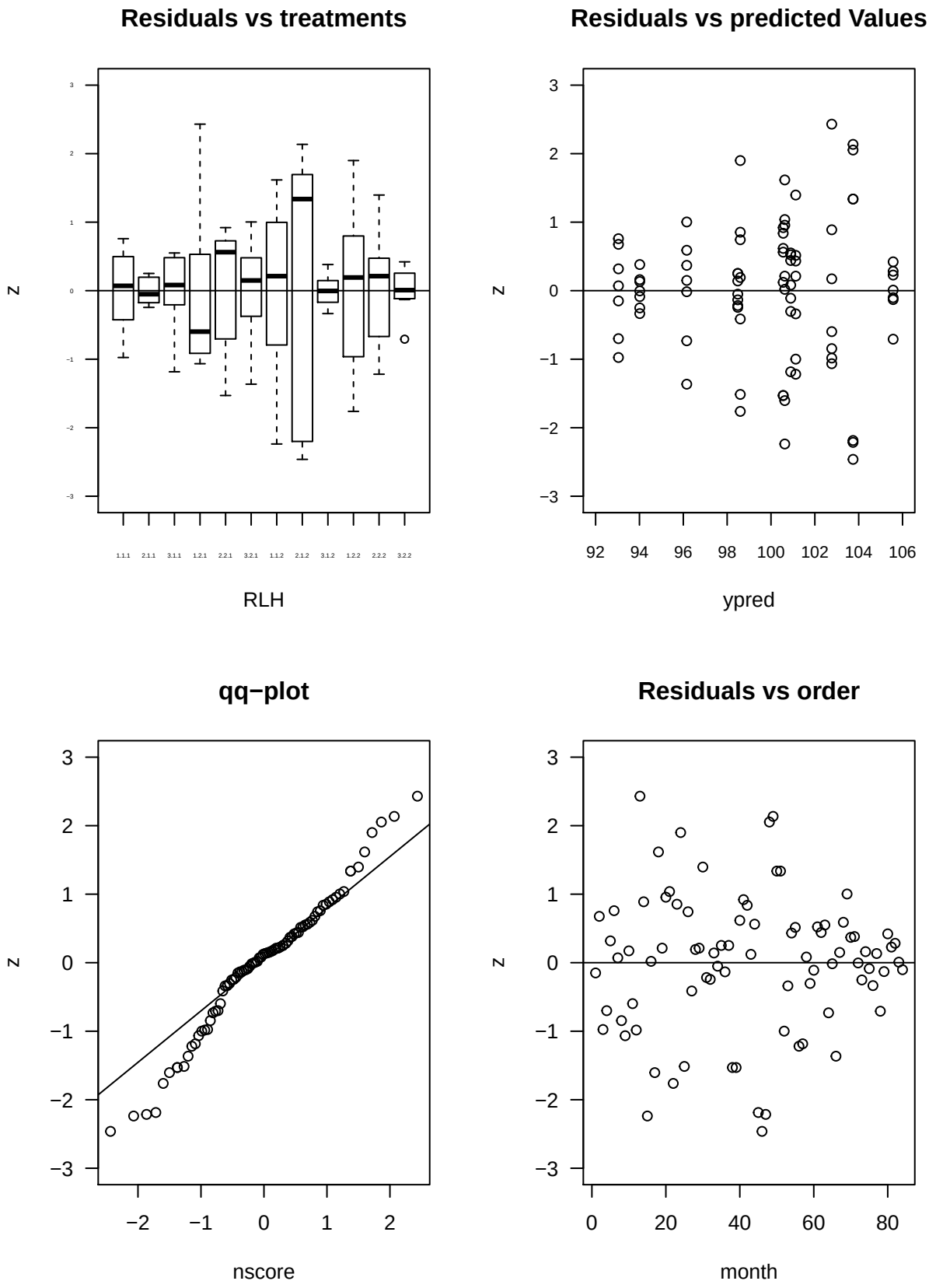

Figure 4: Model assumptions checks

Table 6: Treatment combinations with mean, 95 % confidence levels

TC	Mean	Lower CL	Upper CL	Group
1.1.1	93.04	90.10	95.98	1
3.1.2	94.01	91.08	96.95	12
3.2.1	96.16	93.22	99.10	123
2.1.1	98.49	95.55	101.42	1234
1.2.2	98.60	95.66	101.54	12345
2.2.1	100.56	97.62	103.50	2345
1.1.2	100.63	97.69	103.57	2345
3.1.1	100.90	97.96	103.84	2345
2.2.2	101.13	98.19	104.07	345
1.2.1	102.77	99.83	105.71	345
2.1.2	103.74	100.80	106.68	45
3.2.2	105.57	102.63	108.51	5