STATS 201/8 Assignment 4

Anish Hota ahot228

Due Date: 3pm Thursday 26th September

```
## Welcome to emmeans.
## Caution: You lose important information if you filter this package's results.
## See '? untidy'
```

1 Question 1 [16 Marks]

An agricultural researcher was investigating different methods of improving potato crop yields. A large field was available for an experiment. It was split up into 10 by 10 metre blocks, with two metre pathways separating them. The blocks were randomly allocated to one of three treatment groups, Treatment 1, Treatment 2 or Control. The two treatment groups had one of two different mixtures of soil enhancers added to the soil before planting. Otherwise, the three groups were just grown using standard growing practices. After the plants had fully grown, the plants were all harvested and the yield for each plot was recorded.

The dataset is stored in Potato.csv and includes variables:

Variable	Description
Yield	the yield of the block of potato plants (in kg),
Group	the treatment applied to the block of potato plants, coded as: (Control, T1 for Treatment 1, and T2 for Treatment 2)

We want to know if there is evidence that treating the potatoes by adding soil enhancers increased yield and if one of the treatments was more effective than the other. Quantify any differences.

Instructions:

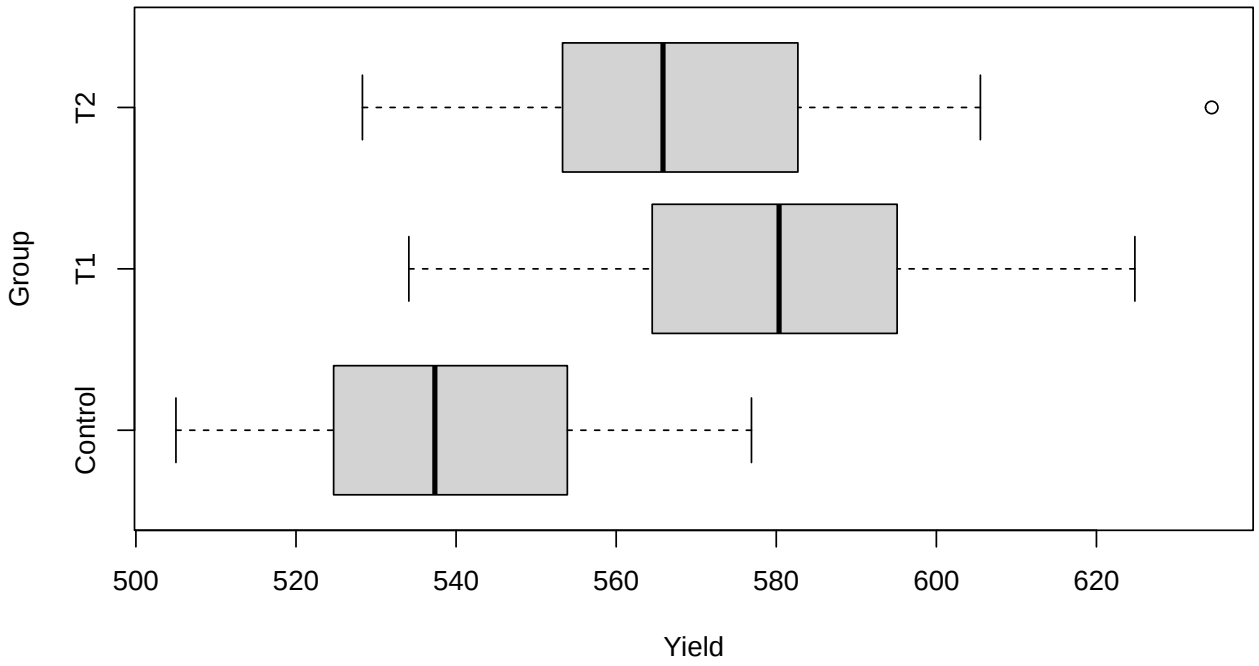
- Make sure you change your name and UPI/ID number at the top of the assignment.
- Comment on the plot/summary statistics of the data.
- Fit an appropriate model to the data. Check the model assumptions.
- Write appropriate **Methods and Assumption Checks**.
- Write an appropriate **Executive Summary**.

1.1 Question of interest/goal of the study:

We want to know if there is evidence that treating the potatoes by adding soil enhancers increased yield and if one of the treatments was more effective than the other.

1.2 Read in and inspect the data:

```
Potato.df <- read.csv("Potato.csv", stringsAsFactors=TRUE)
boxplot(Yield ~ Group,horizontal=TRUE, Potato.df)
```




```
summaryStats(Yield ~ Group, Potato.df)
```

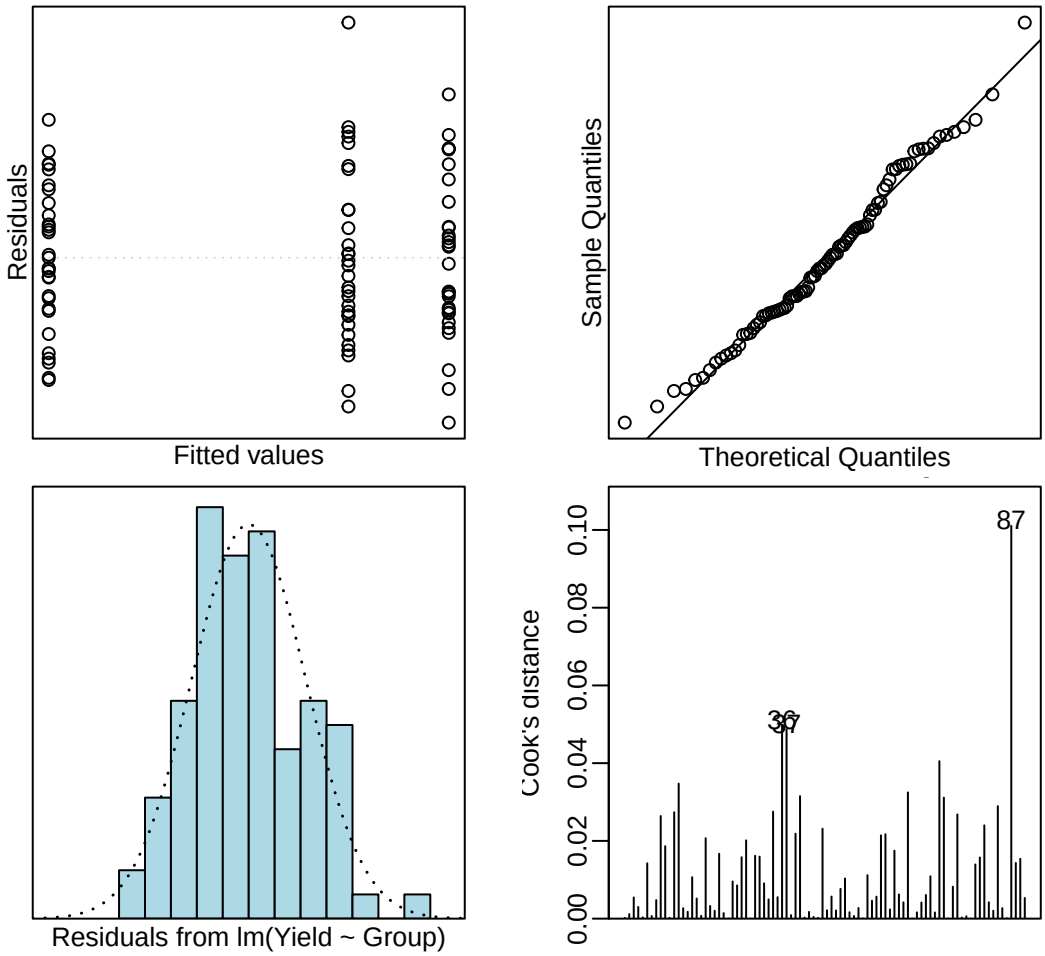
```
##         Sample Size     Mean Median  Std Dev Midspread
## Control          30 538.8067 537.35 19.81454    27.625
## T1               30 579.6833 580.35 22.18453    28.675
## T2               30 569.4333 565.85 24.60162    29.325
```

1.3 Comment on the plots and summary statistics:

We can see that a treatment does help in the yield of potatoes. There is a higher amount of yield in Treatment 1 than there is in Treatment 2 and the control is lower than both treatments. We can also see by the summary table, however, that there is only a sample size of 30, which is quite small so it would be difficult to draw conclusions from this data about the impacts of the treatments.

1.4 Fit an appropriate model and check assumptions:

```
potato.fit <- lm(Yield ~ Group, Potato.df)
modcheck(potato.fit)
```




```
summary(potato.fit)
```

```
##
## Call:
## lm(formula = Yield ~ Group, data = Potato.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.583 -15.121  -1.908  14.637  64.967
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  538.807      4.069 132.422  < 2e-16 ***
## GroupT1       40.877      5.754   7.104 3.15e-10 ***
## GroupT2       30.627      5.754   5.322 7.88e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.29 on 87 degrees of freedom
## Multiple R-squared:  0.3858, Adjusted R-squared:  0.3717
## F-statistic: 27.32 on 2 and 87 DF,  p-value: 6.196e-10
```

```r
confint(potato.fit)
```

```
##                 2.5 %    97.5 %
## (Intercept) 530.71936 546.89397
## GroupT1      29.43949  52.31385
## GroupT2      19.18949  42.06385
```

1.5 Methods and assumption checks:

The residual plot seems to have no problems in our model. The data seems to be normalized. The plots of land were randomly allocated and measurements seemed to be done independently. All model assumptions are satisfied.

Model Equation: $Yield_i = \beta_0 + \beta_1 * Group_i + \epsilon_i$ where $\epsilon_i \sim iid\ N(0, \sigma^2)$

Our model explains 39% of the variability of the data, which is not much but considering a small sample sie this is expected.

1.6 Executive summary:

We wanted to know if there is evidence that treating the potatoes by adding soil enhancers increased yield and if one of the treatments was more effective than the other.

We have very strong evidence that there is a difference between using treatment 1 and using treatment 2. Treatment 1 produces between 29.44 and 52.31 kg more than Treatment 2. (p-value = 3.15e-10)

We have also very strong evidence that there is a difference between not using treatment and using treatment. This can be seen as treatment 2 has between 19.19 and 42.06 kg more than the control variable. (p-value = 7.88e-07)

---

2 Question 2 [20 Marks]

A software company is developing a new computer game. The manager wants to see what effect the speed setting of the game has on the length of time players survive in the game and how this differs depending on players experience level. A large group of play testers is available, most of whom already have some experience playing the new game, and each is allocated to play a version of the game with varying speed settings.

We wish to determine (1) how the speed settings affect players survival times, (2) how survival times differ between players with more and less experience and (3) whether the effects of changing speed settings is different for players with more and less experience. We also wish to quantify any significant effects.

The data is stored is Game.csv and contains the following variables:

Variable	Description
Time	The survival time for the player (in minutes),
Speed	The speed setting at which the software was running, coded as: (Low, Med or Rapid - with Med standing for medium speed),
Experience	The experience level the player, coded as:(Less or More).

Instructions:

- Comment on the interaction plots of the data..

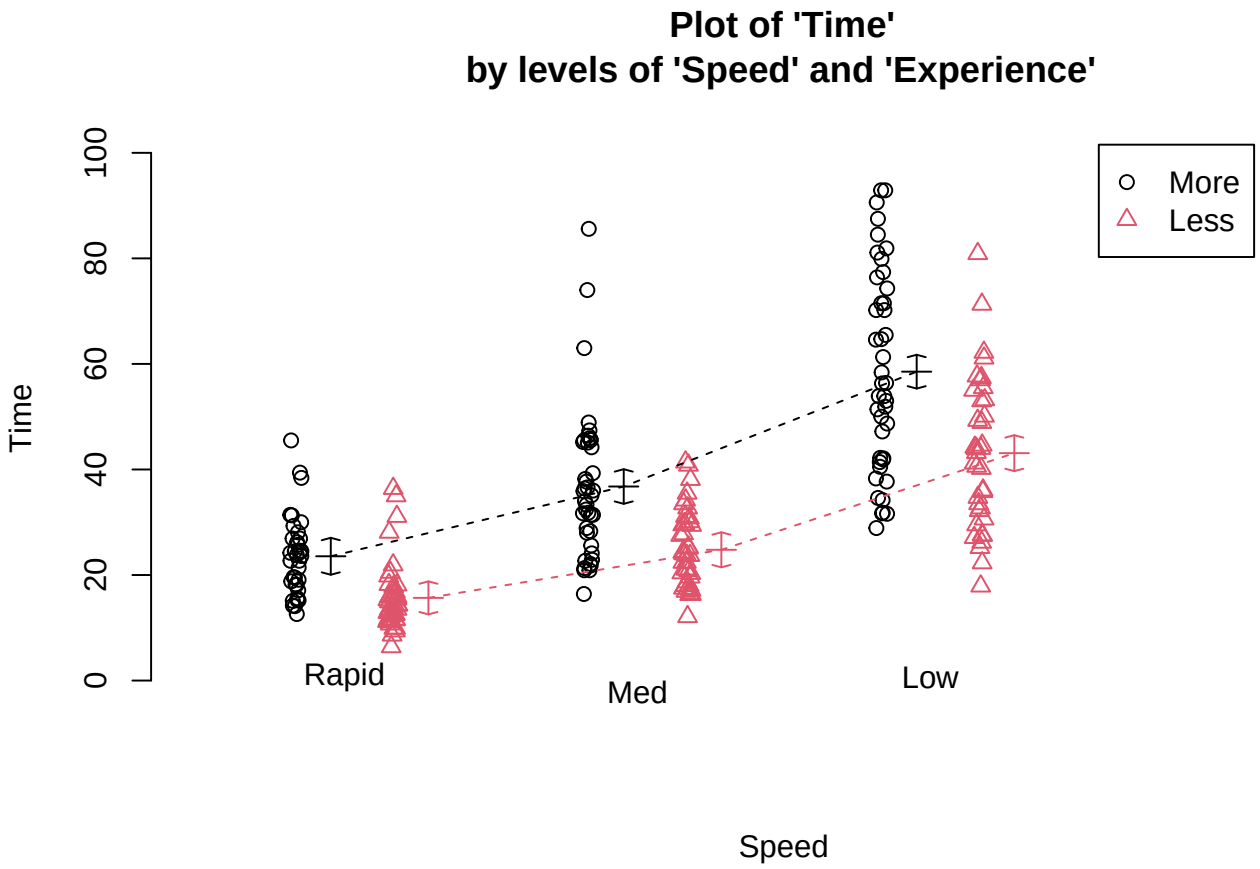
- Fit an appropriate linear model to the data. If necessary, change the model until you are satisfied that you have found the simplest adequate model. Check the model assumptions. Generate the inference output you need from the final model.
- Briefly discuss why using a log transformation of TotalKg is better in this analysis?
- Fit an appropriate model to the data. Check the model assumptions.
- Write appropriate **Methods and Assumption Checks**.
- Write an appropriate **Executive Summary**. Ensure you address the questions of interest.

2.1 Question of interest/goal of the study:

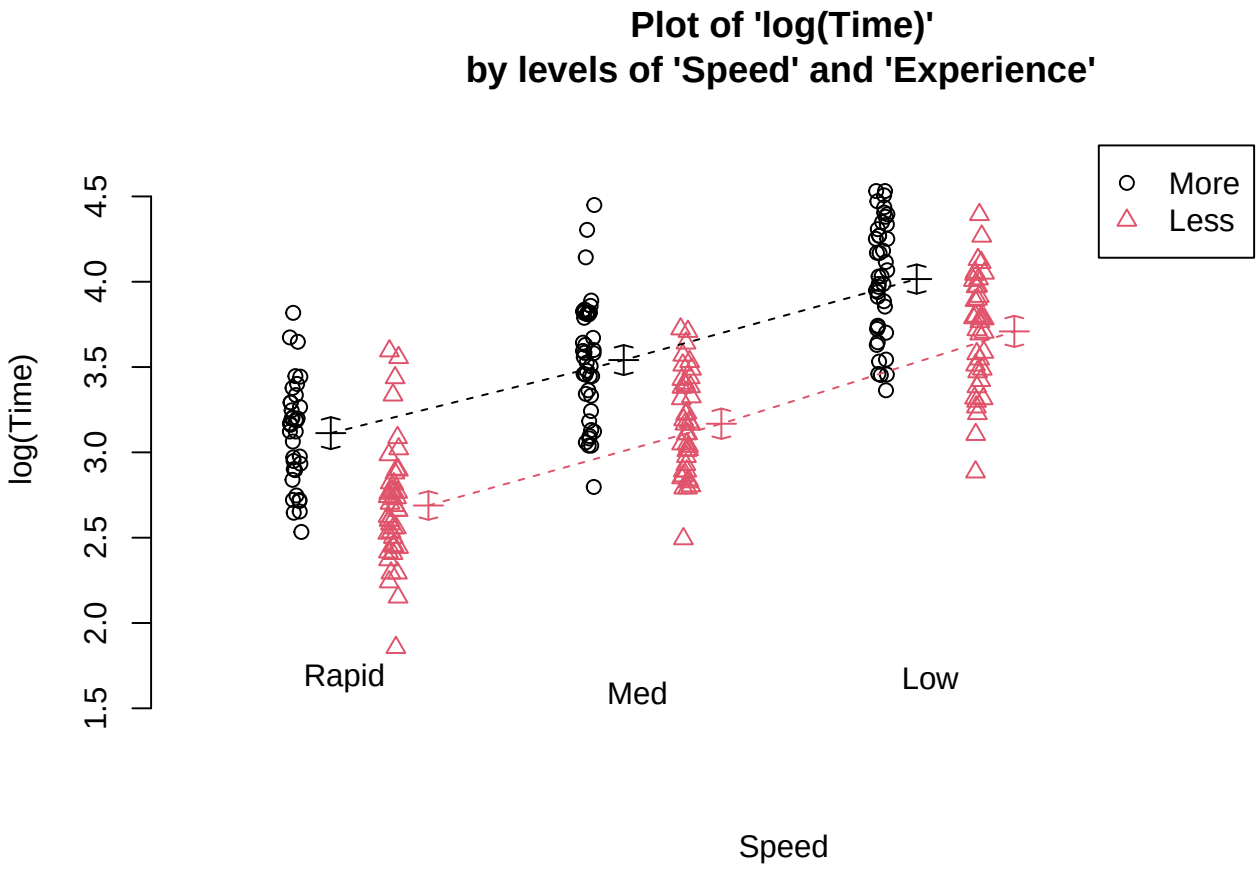
The questions we want answered are: how the speed settings affect players survival times, how survival times differ between players with more and less experience and whether the effects of changing speed settings is different for players with more and less experience?

2.2 Read in and inspect the data:

```
Game.df <- read.csv("Game.csv")
interactionPlots(Time ~ Speed + Experience, data = Game.df)
```




```
interactionPlots(log(Time) ~ Speed + Experience, data = Game.df)
```

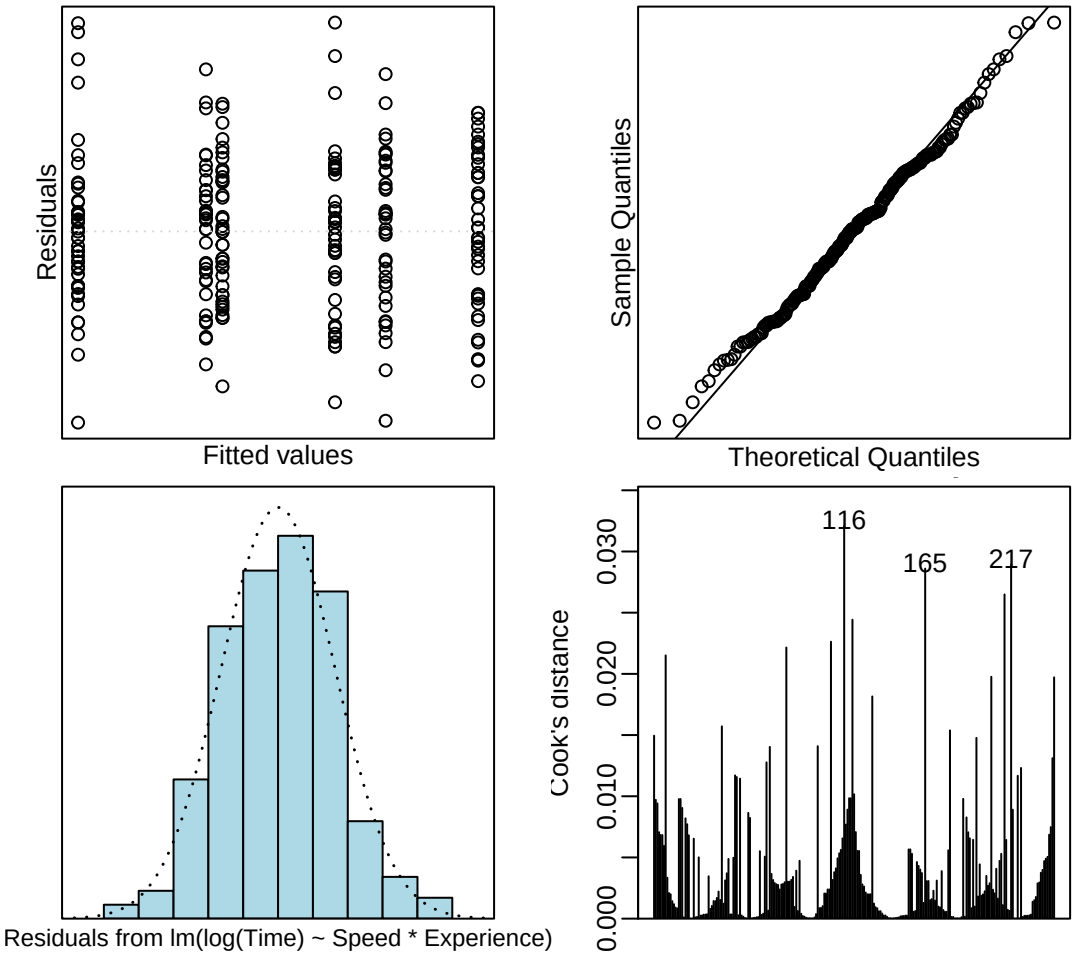


2.3 Comment on the plots:

The data seems skewed and not equally varied before a log was added to Time. There seems to be a generally higher time with people with more experience than less experience.There seems to be an increasing relationship between time and decrease in speed.

2.4 Fit model, simplify as necessary, and generate inference output:

```
game.fit <- lm(log(Time) ~ Speed * Experience, data = Game.df)
modcheck(game.fit)
```




```
summary(game.fit)
```

```
##
## Call:
## lm(formula = log(Time) ~ Speed * Experience, data = Game.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.83231 -0.23760  0.01026  0.23621  0.90857
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 3.70896    0.05377  68.980  < 2e-16 ***
## SpeedMed                   -0.54117    0.07508  -7.208 7.57e-12 ***
## SpeedRapid                 -1.02036    0.07302 -13.973  < 2e-16 ***
## ExperienceMore              0.30669    0.07380   4.156 4.53e-05 ***
## SpeedMed:ExperienceMore     0.06663    0.10427   0.639    0.523
## SpeedRapid:ExperienceMore   0.11749    0.10459   1.123    0.262
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3315 on 237 degrees of freedom
## Multiple R-squared:  0.6502, Adjusted R-squared:  0.6428
## F-statistic: 88.12 on 5 and 237 DF,  p-value: < 2.2e-16
```

```r
confint(game.fit)
```

```
##                                  2.5 %     97.5 %
## (Intercept)                 3.60303744  3.8148872
## SpeedMed                   -0.68908392 -0.3932518
## SpeedRapid                 -1.16421247 -0.8764986
## ExperienceMore              0.16131339  0.4520745
## SpeedMed:ExperienceMore    -0.13878386  0.2720382
## SpeedRapid:ExperienceMore -0.08855014  0.3235356
```

2.5 Why is using a log transformation of Time better in this analysis?

The data was skewed and there wasn't equality of variance, and after the log transformation on time the data became more spread out and varied equally so the log transformation was neccessary.

2.6 Methods and assumption checks:

There was a log transformation needed as ther there was no equality of variance but after a log transformation was made there is equal variance, the data seems normalised and the residuals seems good. So all assumptions seemed to be satisfied in the final model.

Model of the equation:

$log(Time_i) = \beta_0 + \beta_1 * Speed_i + \beta_2 * Experience_i + \beta_3 * Speed_i * Experience_i + \epsilon_i$ where $\epsilon_i \sim iid\ N(0, \sigma^2)$

Our data explains 64.8% of the variability of the data.

2.7 Executive summary:

The questions we want answered were: how the speed settings affect players survival times, how survival times differ between players with more and less experience and whether the effects of changing speed settings is different for players with more and less experience?

We have very strong evidence that there is a difference in survival times depending on the speed with the difference between low and medium speed being between a log of 0.40 and 0.61 minutes and the difference between medium and rapid speed being between a log of 0.86 and 1.07 minutes. This said as the speed goes down the more time is being taken. (p-value = 2e-16)

We have strong evidence that more experience takes more time than less experience, more experience is about 0.284 and 0.452 minutes higher than less experience. (p-value = 8.09e-16)

There is no evidence that there is a difference between speed and its effect and on more or less experience this can by seen by high p-values of 0.523 and 0.262 for the difference between the three speeds.