STAT 425

Term Project

Experimental Study: Computer booting efficiency within TFDL floors

By:
Fabian Pradipto
Parnashi Patel
Steven La

Table of Contents

Disclaimer: Due to recent circumstances, we were not able to proceed with our original idea exactly because of restricted access to the schools libraries so we have limited our project to be just within TFDL.

1. Motivation

For our project we have decided to test the mean start up times for logging into a computer from when the computer is on stand-by till logged in and ready to use in the TFDL library on the first 3 floors, these would be our factors. Our analysis is to see if these start up times may vary and if one floor could be faster/slower or the same as another which is tested by a set of ten randomly chosen computers on each floor.

We chose the idea because we think that it could be useful information to know whether some floors might have faster/slower or even the same start up times which could affect a student who is looking for a computer to use within the library. Perhaps students may find it useful when they are in a rush to complete a task that requires a computer like printing, submitting work, or even emails. They would most likely prefer to use a computer that will boot up faster versus a computer that might take longer.

For this analysis we think that the busier the computer lab is the slower the boot times for the computers will be because it has a higher usage rate. This is why we are predicting that floor 1 will be the slowest because it has the highest amount of traffic and usually the computers are all occupied, then floor 2 and lastly floor 3.

Because we are using a mixed effect model in our analysis we will have statistical hypothesis' from 3 effects.

Interaction Effect

$H_0 : \sigma^2_{\tau\beta} = 0$

$H_1 : \sigma^2_{\tau\beta} > 0$

Day Effect

$H_0 : \sigma^2_{\beta} = 0$

$H_1 : \sigma^2_{\beta} > 0$

Floor Effect

$H_0 : \tau_{Floor1} = \tau_{Floor2} = \tau_{Floor3}$

2. Design

Data Collection

We collected the data in TFDL on the following date, Wednesday, March 18, 2020 and Friday, March 20, 2020. Two of our members will take some data during the same week based on our schedule, we agreed on taking these data on floor 1, 2, and 3 of TFDL. These floors are arguably the busiest floor in terms of student activity and network traffic.

Since we are planning to conduct the experiment as a two-factorial design we will cross factor “Days” and “Floor” so that for each level of “Days” we collected the data, each variant of “Days” will have each levels of factor “Floors” which includes “Floor 1”, “Floor 2”, and “Floor 3”.

Initially we wanted to collect the data during a busier time period, where a lot of students are actually using these computers. However, due to the current public health situation, most of these computers are vacant. Although it didn’t go as planned, these vacant computers allows us to test and collect these data on them in a fairly controlled environment as there is less traffic which results in an easier time to find unused computers.

We try to take the sample randomly by avoiding collecting them in a sequential order and use what is available to use, this includes the printing computers that have a maximum period of usage of 10 minutes. The following are the raw data we collected on the two separate days.

Raw Data (March 18, 2020)

	Floor 1	**Floor 2**	**Floor 3**	**Days**
Comp 1	16.85	13.30	86.04	Wednesday
Comp 2	16.97	12.71	90.87	Wednesday
Comp 3	17.83	13.70	83.82	Wednesday
Comp 4	13.93	33.73	87.79	Wednesday
Comp 5	89.92	12.73	82.78	Wednesday
Comp 6	15.93	13.97	84.72	Wednesday
Comp 7	13.19	12.02	83.03	Wednesday
Comp 8	13.97	86.57	83.56	Wednesday
Comp 9	14.52	12.16	87.10	Wednesday
Comp 10	17.90	12.90	86.18	Wednesday

Raw Data (March 20, 2020)

	Floor 1	**Floor 2**	**Floor 3**	**Days**
Comp 1	14.14	14.07	11.12	Friday
Comp 2	15.64	12.09	82.02	Friday
Comp 3	15.64	13.24	82.30	Friday
Comp 4	15.79	12.37	82.77	Friday
Comp 5	14.24	85.87	85.64	Friday
Comp 6	13.90	12.02	84.60	Friday
Comp 7	15.89	12.72	84.91	Friday
Comp 8	15.67	13.14	82.36	Friday
Comp 9	15.09	12.29	82.12	Friday
Comp 10	16.97	12.84	83.04	Friday

Data Storage

Although the raw data is intuitively laid out, we will need to perform some data wrangling to allow the data analyzing process to be more efficient. We will be creating individual columns for each of the identifying factors that includes “Days” and “Floor” to match the corresponding computer boot-up time. We will not be needing to add the computer variable “C1”, “C2”, etc, as it was randomly selected and the “C1” in the first floor is not linked with “C1” from the second and third floor. We will use the following code to import our data into R.

```
time = c(16.85, 16.97, 17.83, 13.93, 89.92, 15.93, 13.19, 13.97, 14.52, 17.90,
14.14, 15.64, 15.64, 15.79, 14.24, 13.90, 15.89, 15.67, 15.09, 16.97, 13.30,
12.71, 13.70, 33.73, 12.73, 13.97, 12.02, 86.57, 12.16, 12.90, 14.07, 12.09,
13.24, 12.37, 85.87, 12.02, 12.72, 13.14, 12.29, 12.84, 86.04, 90.87, 83.82,
87.79, 82.78, 84.72, 83.03, 83.56, 87.10, 86.18, 11.12, 82.02, 82.30, 82.77,
85.64, 84.60, 84.91, 82.36, 82.12, 83.04)
floor = c(rep("F1", 20), rep("F2", 20), rep("F3", 20))
day = c(rep("Wed", 10), rep("Fri", 10))
day = c(rep(day, 1))
project.df = data.frame(floor, day, time)
project.df
```

This is our output from the R code above.

```
floor days   time
F1     Wed    16.85
F1     Wed    16.97
F1     Wed    17.83
...................
F3     Fri    82.36
F3     Fri    82.12
F3     Fri    83.04
```

Design of the data

We will be using the two factorial design when analyzing our data, as we are interested in observing the possible interaction between “Days” and “Floor” with regards to the time it takes for the computer to boot-up. We will be employing the mixed-effects model in our two factorial analysis where:

- “Days” factor is random
- “Floor” factor is fixed

Since our model is of the mixed-effects, we can express them in the following way

$X_{ij} = \mu + \tau_i + \beta_{jRandom} + (\tau\beta)_{ijRandom} + e_{ijl}$

- τ_i - Floor effect with k levels where $k = 3$
- $\beta_{jRandom}$ - Day effect with b levels where $b = 2$
- $X_{ij} = \mu + \tau_i + \beta_{jRandom} + (\tau\beta)_{ijRandom} + e_{ijl}$- Interaction effect between “Floor” and “Day”

The “Day” factor is random because we randomly sampled two days out of the possible seven in a week, to make an inference about the “Day” population and see whether the resulting computer boot-up times differ between the different days..

The “Floor” factor on the other hand is fixed because we purposefully limit our parameter of interest to only involve “Floor 1”, “Floor 2”, and “Floor 3” of TFDL. In this case, the inference about the influence of the different floors on the resulting computer boot-up time can only be exclusively applied to these 3 floors and does not tell us anything about the other floors that we did not take a sample of.

The interaction effect between “Floor” and “Day” is an advantage for the two-factorial design, as it is meant to observe whether the resulting computer boot-up time at one level of “Day” differs for other levels of “Floor” and vice-versa. For example we are interested to see whether the differences is significant enough when using a computer in “Floor 3” on a “Wednesday” compared to using them on a “Friday”.

3. Analysis

Visual Diagnostics

To analyze a mixed-effect model in a two-factorial design, we have to satisfy a certain conditions, the conditions is as follows:

1. The residual terms, are Normally distributed with a mean of 0 and variance σ^2_{common}
2. The variation in the response variable for all k -levels of Factor A is homogeneous with σ^2_{common}
3. The variation in the response variable for all b -levels of Factor B is homogeneous with σ^2_{common}

We can check whether or not these conditions are satisfied through plotting some visual diagnostics and/or applying some test(s) to our data.

We will first create our Anova table using the following code.

```
project.aov = aov(time ~ floor + day + floor:day, data=project.df)
summary(project.aov)
```

Output

```
Call:
   aov(formula = time ~ floor + day + floor:day, data = project.df)

Terms:
                   floor      day floor:day
Sum of Squares  49050.98   641.51    141.12
Deg. of Freedom        2        1         2
                Residuals
Sum of Squares   19539.00
Deg. of Freedom        54

Residual standard error: 19.02192
Estimated effects may be unbalanced
            Df Sum Sq Mean Sq F value   Pr(>F)
floor        2  49051   24525  67.781 1.89e-15
day          1    642     642   1.773    0.189
floor:day    2    141      71   0.195    0.823
Residuals   54  19539     362

floor       ***
day
floor:day
Residuals
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

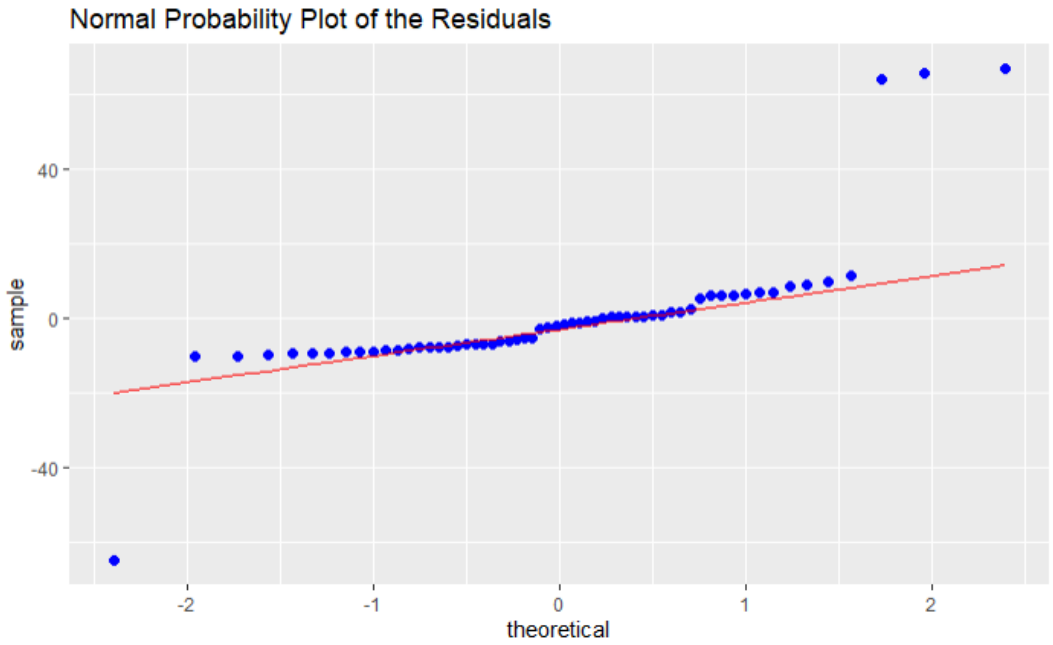
Condition 1:

To check whether or not the residuals are normally distributed, we will find our residual terms and load it into our data.frame using the following code.

```
eijlproj = residuals(project.aov)
fittedproj = fitted(project.aov)
projdiag.df = cbind(project.df, eijlproj, fittedproj)
```

And we will use the following code to graph

```
ggplot(data=projdiag.df, aes(sample=eijlproj)) + stat_qq(size=2, col="blue") +
stat_qqline(col="red") + ggtitle("Normal Probability Plot of the Residuals")
```



From the graph above, we can see that there are significant deviations near the end points but most of the pattern still lies along the line, We are going to run the Shapiro-Wilks to confirm if the normality condition is satisfied.

```
shapiro.test(projdiag.df$eijlproj)
W = 0.63444, p-value = 5.847e-11
```

We can see that the p value of $5.847e-11 < 0.05$, which indicates that the normality condition is not satisfied. This could be a problem as our design presumes this condition, however it is quite likely that the 4 outliers that deviate significantly from the line is what makes these residuals not normally distributed despite most of the points being on/near the line.

Condition 2 and Condition 3:

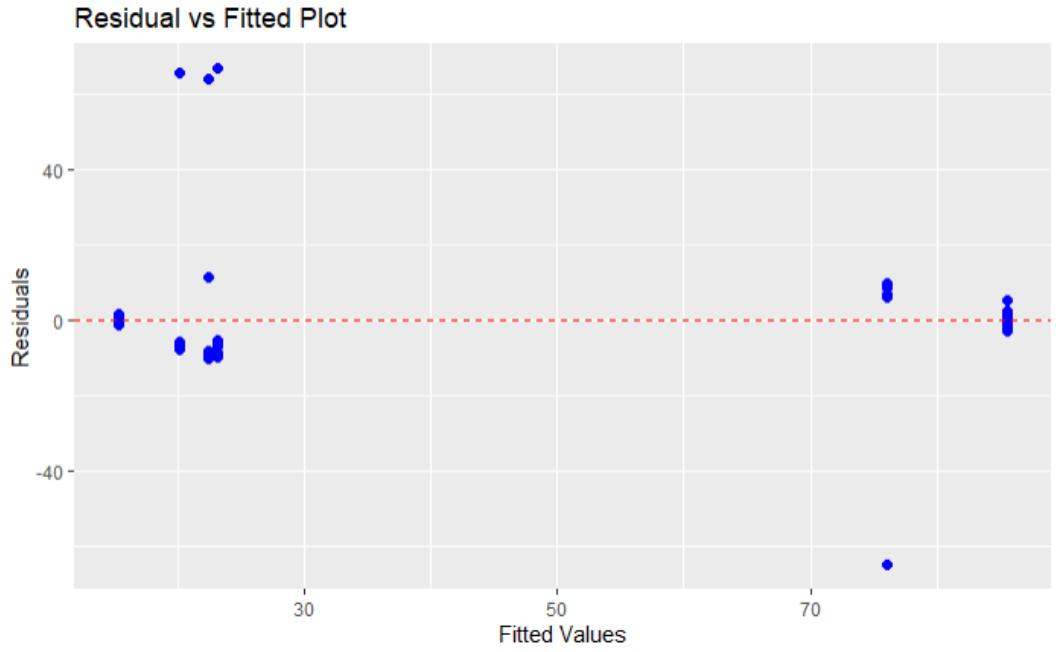
To verify that the data satisfy condition 2 and 3, we will be plotting the following graphs

- Residuals to fitted plot
- Residuals to means of the boot-up time by “floor”
- Residuals to means of the boot-up time by “day”

Residuals to fitted

To plot the residuals to fitted plot we will be using the code below

```
ggplot(data=projdiag.df, aes(x = fittedproj, y = eijlproj)) + geom_point(size=2,
col="blue") + xlab("Fitted Values") + ylab("Residuals") +
geom_hline(yintercept=0, linetype="dashed", col="red") + ggtitle("Residual
Plot")
```



We can see from our plot above that the data are not evenly scattered and the pattern of the data is somewhat lacking below the line on the left-side of the plot and is lacking above the line on the right-side of the plot. However, we did not see an obvious wedge that may indicate unequal variances; so, we will assume that the variances are equal for the sake of the experiment.

Residuals to “floor” levels mean and Residuals to “day” levels mean

We are hoping for a rectangular “band” pattern which indicates that the variance are equal within the means. In order to plot both plots, we will first find the mean values of the computer boot-up time by the floor using the favstats function.

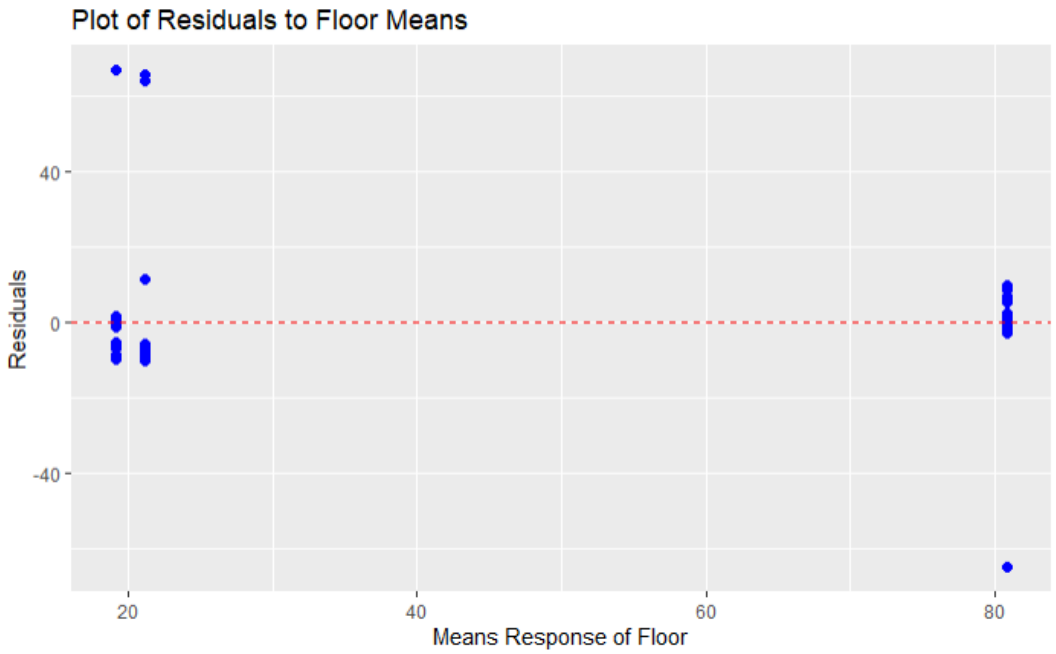
```
>favstats(~time|floor, data=project.df)

Output:
 floor   min      Q1 median      Q3   max    mean       sd  n missing
1    F1 13.19 14.2150 15.655 16.8800 89.92 19.1990 16.70136 20       0
2    F2 12.02 12.3500 12.870 13.7675 86.57 21.2220 22.71987 20       0
3    F3 11.12 82.6675 83.690 85.7400 90.87 80.8385 16.56597 20       0
>favstats(~time|day, data=project.df)

Output:
 day   min      Q1 median     Q3   max     mean       sd  n missing
1 Fri 11.12 13.1650  15.64 82.255 85.87 37.15000 33.41699 30       0
2 Wed 12.02 13.7575  17.40 84.495 90.87 43.68967 35.40258 30       0
```

We will then bind our mean value of “Floor” and “Day” to a new data frame

```
floor.means = c(rep(19.1990, 20), rep(21.2220, 20), rep(80.8385, 20))
day.means = rep(c(rep(37.15000, 10), rep(43.68967, 10)), 3)
floordiag2.df = cbind(projdiag.df, floor.means, day.means)
```

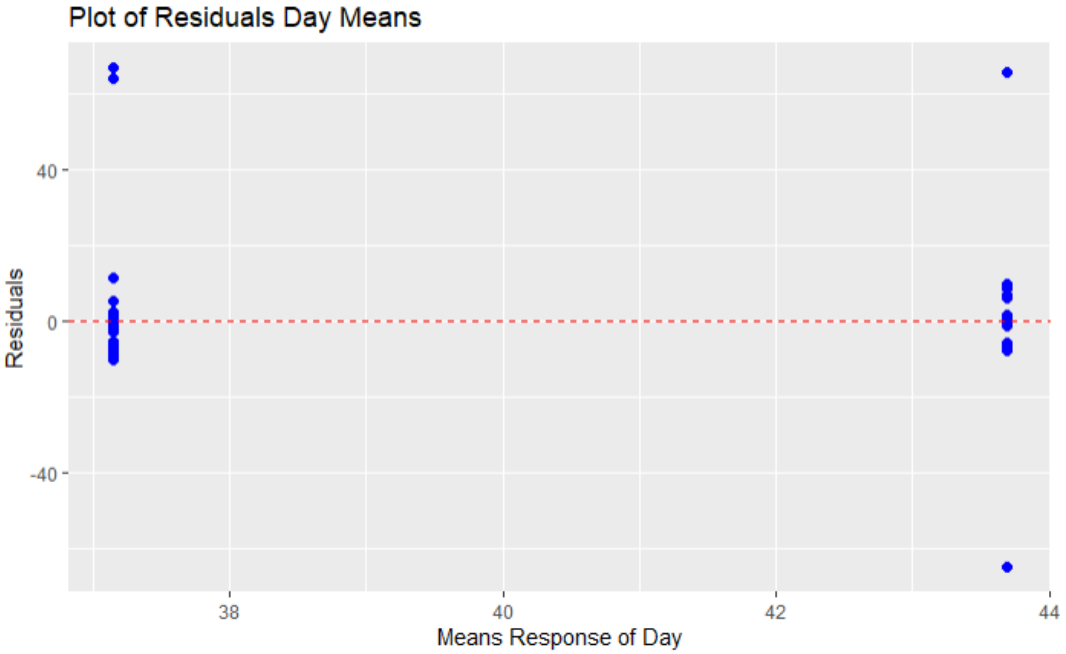
The following code is used to plot residual to floor means

```
ggplot(data=floordiag2.df, aes(x = floor.means, y = eijlproj)) +
geom_point(size=2, col="blue") + xlab("Means Response of Floor") +
ylab("Residuals") + ggtitle("Plot of Residuals to Floor Means") +
geom_hline(yintercept=0, linetype="dashed", col="red")
```



The following code is used to plot residual to day means

```
ggplot(data=floordiag2.df, aes(x = day.means, y = eijlproj)) +
geom_point(size=2, col="blue") + xlab("Means Response of Day") +
ylab("Residuals") + ggtitle("Plot of Residuals Day Means") +
geom_hline(yintercept=0, linetype="dashed", col="red")
```



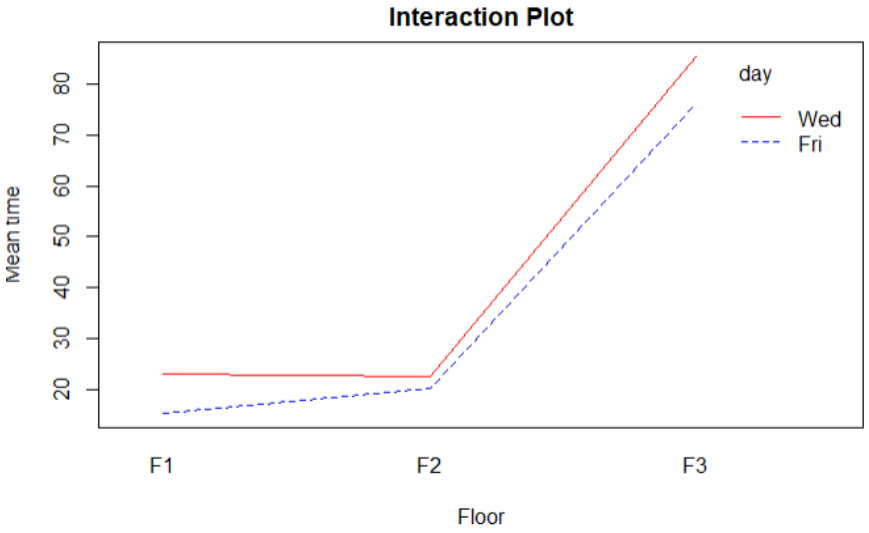
We can see that both plots of the residuals to floor means and residuals to day means that the data points mostly lie near to each other in a consistent pattern, which indicates that both the variance for the k-levels and b-levels are approximately the same. Adding the residual to the fitted plot and the two residuals to mean plot, we can conclude that condition 2 and 3 is satisfied.

Hypothesis Testing

Interaction effect:

Before stating our statistical hypotheses, we are going to plot the interaction plot, to get an idea of how both factors are related to each other, if they are at all.

```
with(project.df, {interaction.plot(x.factor = floor, trace.factor=day,
response=time, fun=mean, col=c("blue", "red", "purple"), xlab="Floor",
ylab="Mean time", main="Interaction Plot")})
```



Our plot above suggests that there does not exist an interaction between "Floor" and "Day" as the line did not cross. We can further confirm our suspicion by using the following statistical hypotheses

$H_0 : \sigma^2_{\tau\beta} = 0$ (There is no variation in the computer boot-up times in all the different combinations of "Floor" and "Day")

$H_1 : \sigma^2_{\tau\beta} > 0$ (There is a variation in the computer boot-up times in all the different combinations of "Floor" and "Day")

$$F_{obs} = \frac{MSAB}{MSE} = \frac{71}{362} = 0.1961$$

We will then find our p-value using R

```
1-pf(0.1961, df1 = 2, df2 = 54)
```

$$P(F_{2,54} > 0.1961) = 0.8225$$

Since $0.8225 > 0.05$, we fail to reject the null hypothesis and infer that there is no variation between the three-different floors (fixed factor) and the day (random factor) on the time to boot the computer in TFDL library. We can see that our interaction plot also supported that there will be no interaction between "Day" and "Floor".

Day effect:

$H_0 : \sigma_\beta^2 = 0$ (there is no variation in the computer boot-up time amongst all levels of "Days")

$H_1 : \sigma_\beta^2 > 0$ (there is a variation in the computer boot-up time amongst all levels of "Days")

$$F_{obs} = \frac{MSB}{MSAB} = \frac{642}{71} = 9.0423$$

We will then find our p-value using R

```
1-pf(9.0423, df1 = 1, df2 = 2)
```

$$P(F_{1,2} > 9.0423) = 0.09508$$

Since $0.09508 > 0.05$, we fail to reject the null hypothesis and conclude that there does not exist a day effect or that there is no variation between in the computer boot-up time amongst "Friday" and "Wednesday".

Floor effect:

$H_0 : \tau_{Floor1} = \tau_{Floor2} = \tau_{Floor3}$ (The mean computer boot-up time is equal amongst all the "Floor")

$H_1 : H_0 \text{is false}$ (The mean computer boot-up time is not equal amongst all the "Floor")

$$F_{obs} = \frac{MSA}{MSAB} = \frac{24525}{71} = 345.4225$$

We will then find our p-value using R

```
1-pf(345.4225, df1 = 2, df2 = 2)
```

$$P(F_{2,2} > 345.4225) = 0.0028$$

Since $0.0028 < 0.05$, we reject the null hypothesis and conclude that there does exist a "Floor" effect. This means that there exist a difference in the mean computer boot-up time amongst "Floor 1", "Floor 2", and "Floor 3".

Since our result is statistically significant, we can then run a multiple comparison method to figure out which "Floor" factor differs from one another.

Multiple-Comparison Test

Based on our previous hypothesis test, we have found that the only result that was found to be significant was the “Floor” effect. We will be applying the Tukey’s HSD test to our data in order to figure out which “Floor” has a significantly different mean computer boot-up time.

The following are the code for to run Tukey’s HSD
Input:

```
floorvarietyTukey = TukeyHSD(aov(time ~ floor + day + floor:day,
data=project.df), ordered=T)
floorvarietyTukey$floor
```

Output:

```
        diff       lwr      upr        p adj
F2-F1  2.0230 -12.47368 16.51968 9.396282e-01
F3-F1 61.6395  47.14282 76.13618 5.463408e-13
F3-F2 59.6165  45.11982 74.11318 7.436274e-13
```

We can see from our result above that our finding states that the mean computer boot-up time is not significantly different between $(\mu_{F2} - \mu_{F1})$, but is significantly different between $(\mu_{F3} - \mu_{F1})$ and $(\mu_{F2} - \mu_{F2})$. Hence we can summarize our findings as follows.

Finding:
$\mu_{F3} > (\mu_{F2} = \mu_{F1})$

Then we know that the third floor of TFDL has the slowest mean computer boot-up time out of the non-quiet floors.

4. Conclusion

At the start of the experiment, we suspected that the computers in “Floor 1” would take the longest time to boot. However, through a series of tests and comparisons; our findings have corrected us and show that the “Floor 3” computers take the longest time to boot. Keeping in mind that TFDL was fairly empty and the active users were low in all 3 floors when the data was collected, the result shown by the Tukey HSD is something that is worth bringing up to TFDL’s IT department. As it shows that the problem may lie within the system itself instead of any underlying external factors.

After a change in our design experiment, we suspected that the “Day” effect would not exist. This is simply because of a lack of usage of these computers in TFDL due to the COVID-19 outbreak resulting in TFDL being fairly empty throughout the week. In our original design; we suspected that using the computers on a weekend or a common deadline date such as Friday would create a significant enough effect on the computer-boot up times due to a higher traffic. With the new design, testing for the “Day” effect seems to have confirmed our suspicion.

In terms of the interaction effect between “Day” and “Factor”, we had no specific suspicion that we are confident enough to mention in the beginning.We tested for interaction effect because we were interested in seeing how the factors of data behave with each other. After testing for it, we suspect that an interaction would be more likely to happen if the computers were actually used and that certain “Floor” might have a lower usage on certain “Day”.

It is both surprising and reassuring to see and compare our suspicion and findings at the end of the experiment. If we could do It again, we would probably resample the data so that we can pass the visual diagnostics test in a more convincing way. We were also curious about the results if we collected the data during a normal school year where in-class teaching is the main method of course delivery.

Citations

Montgomery, Douglas C. *Design and Analysis of Experiments, 10th edition.* Wiley, 2019.

Stallard, Jim. "*Partially Completed Lecture Notes - March 4.*" D2L, 4 Mar. 2020, d2l.ucalgary.ca.

Stallard, Jim. "*Partially Completed Lecture Notes - March 6.*" D2L, 6 Mar. 2020, d2l.ucalgary.ca.

Stallard, Jim. "*Partially Completed Lecture Notes - March 9.*" D2L, 9 Mar. 2020, d2l.ucalgary.ca.

Stallard, Jim. "*Partially Completed Lecture Notes - March 11.*" D2L, 11 Mar. 2020, d2l.ucalgary.ca.