Stat 602

Group 3: Sungrim Lee, Kyungjin Cho

**Final Report**

**1. Introduction**

For the final project, our group has chosen the experiment which analyzes the difference of driving time it takes from Hilldale to Capitol, Madison, WI. As students in UW-Madison experiencing the traffic every weekday, we were genuinely curious about the time it takes to drive to work in Madison. Since a lot of people with personal vehicles in Madison reside in Hilldale, we set the starting point as Hilldale. Then, we set the destination point as the State Capitol due to the fact that a lot of firms are located in this specific area. Our question of interest would be “Will there be a significant difference in the time it takes to drive to the State Capitol from Hilldale depending on the specific time on the weekday?”

**2. Experimental Design & Justification for Chosen Experimental Design**

Our group will set the driving time taken to Wisconsin State Capitol from Hilldale as a response variable (in minutes), and a treatment variable will be a time slot separated by three different time slots (9 AM, 6 PM, 10 PM). However, on weekdays, the driving time taken can be affected by the day of week. Hence, in order to remove these controllable nuisance factors, we will block such factor by measuring the driving time over three days (Monday, Wednesday, Friday, but not on weekends) in a week. To make this experiment feasible, Randomized Complete Block Design is used. Below is a table that explains the model starting point (Origin) and the end point (Destination). It also contains the reference image of map from Target Hilldale to the State Capitol.

| Origin | 750 Hilldale Wy, Madison, WI 53705  Target Parking Lot |
| --- | --- |
| Destination | 2 E Main St, Madison, WI 53703  Wisconsin State Capitol |
|  | |

*<Table 2.1 - Measurement Origin, Destination and Reference Map>*

We decided to use Randomized Complete Block Design since our nuisance factor is controllable, as mentioned above. We block this factor by driving on different weekdays. This way, we are able to eliminate the effect on our treatment of interest. The table below represents our experimental design.

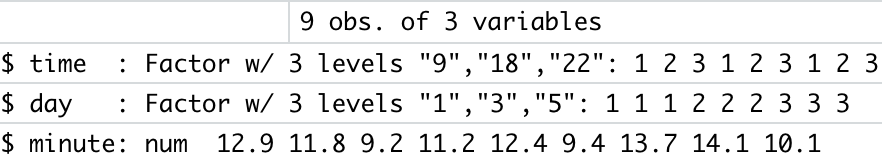
| Day (Block) | | |
| --- | --- | --- |
| Mon | Wed | Fri |
| 9 AM | 10 PM | 6 PM |
| 6 PM | 9 AM | 10 PM |
| 10 PM | 6 PM | 9 AM |

*<Table 2.2 - RCBD Experimental Design>*

**3. Measurement and Model**

Measurements were repeated under almost the same conditions. The same person in our group drove the same car along with the same route from the same parking lot in front of Target, Hilldale to the State Capitol, not speeded at all. Fortunately, there wasn’t any snowing or raining days while driving, which if it was, might have affected the measurements.

The dataset consists of three columns: time, day, and minutes. ‘Time’ is a feature of interest where it indicates the three different times (24-hour basis), and ‘day’ is a potential nuisance factor in which each number means a categorical value of three days (1: Monday, 3: Wednesday, 5: Friday). Driving time is measured in minutes and rounded up to one decimal place. From the observations, we have Randomized Complete Block Design with 3 treatments, 3 blocks, and 1 observation.



*<Table 3.1 - Variables decomposition>*

**Model : yij = μ + τi + βj + eij**

*where μ is the grand mean, τ is the treatment effect, β is the block effect*

Based on this design, our model is constructed like the above. Since τ and β indicate deviations from the grand mean, the summations of all elements from τ and β are 0. Our effect of interest is whether different times of the day affect the driving time taken to the State Capitol from Hilldale. Therefore, our hypotheses is:

**H0 : τ1 = τ2 = ···= τa = 0**

**HA : τi is not equal to 0 for any i**

**4. Variance Decomposition**

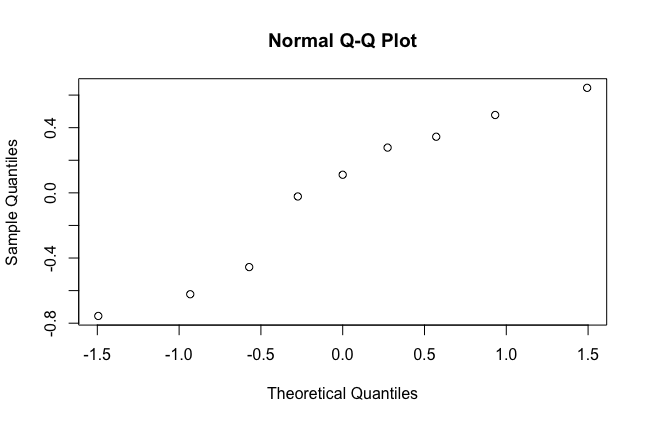
To test the hypotheses, variance decomposition is used to calculate F-statistics. The total variability of RCBD is partitioned into:

SSTot = SStrt + SSblock + SSE , where SSTot = ∑∑(yij − ..)2,

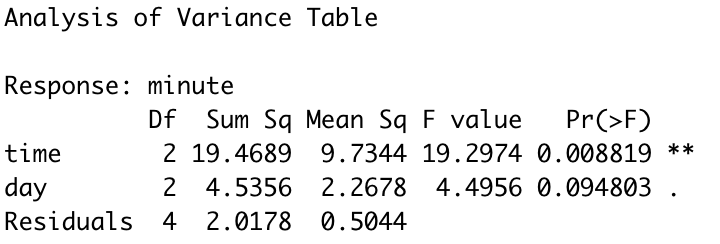
SStrt = b∑(i. − ..)2, SSblock = a∑(.j − ..)2,

SSE = ∑∑(yij − i. − .j + ..)2

There are a\*b observations in total, so the total degrees of freedom is a\*b − 1, and the degrees of freedom for treatment and the block are a − 1, and b − 1, respectively. The degrees of freedom for error term is calculated by (a − 1)\*(b − 1) so that the total degrees of freedom is equal to the summation of treatment, block, and error terms. Utilizing these values, the ANOVA table can be established as shown below. We also assume that the error terms follow a normal distribution N(0, σ2) to implement and interpret ANOVA. Violation of this assumption can be checked by a QQ-plot. The QQ-plot of our model is shown below and it seems that the error points fit to a relatively straight linear line, which means that the model does not violate a normality assumption.

**

*<Graph 4.1 - QQ plot>*

**

*<Table 4.1 - ANOVA table of the Model>*

*(R code details can be found in the last section)*

Since the p-value for time variable is 0.008819 which is smaller than 0.05, we reject the null hypothesis and conclude that the driving time taken from the Capitol to Hilldale is significantly different from the time of the day one drives a car.

**5. Relative Efficiency**

We did not regard the p-value for the blocking factor ‘day’, due to several reasons. First, we already suspect that the blocks are different. We think that the day of the week will affect our experiment since there tends to be a higher volume of traffic on Friday due to the workers going home on time. However, we still checked if our blocking factor was a good choice. The block is considered to be efficient and effective at reducing noise when MSblock > MSE. Since our MSblock = 2.2678, and MSE = 0.5044, the block factor is considered to be effective and informative.

Furthermore, to formally check how efficient the selected blocking scheme is, the relative efficiency can be computed. Relative Efficiency is a measure of the ratio between two estimators’ variances in general. Yet, RCBD has different characteristics, such as degrees of freedom and MSE from CRD’s. Therefore, its variance ratio should be adjusted by a precision factor to calculate the relative efficiency of an RCBD to CRD as shown below:

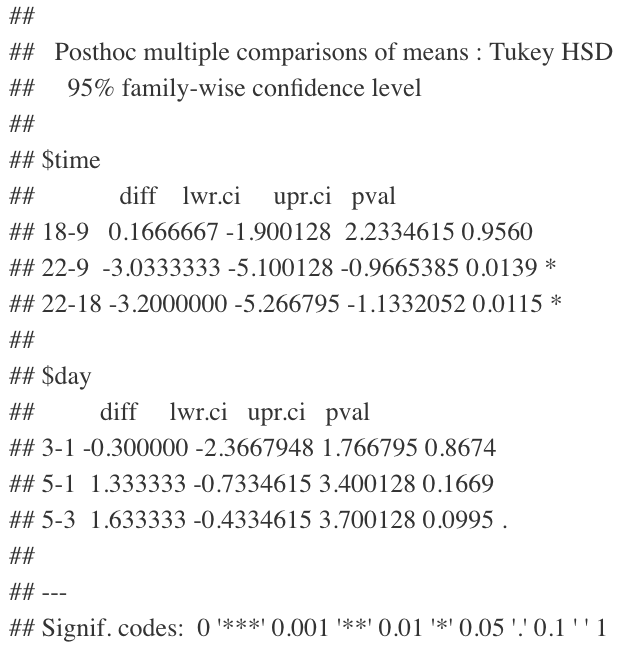
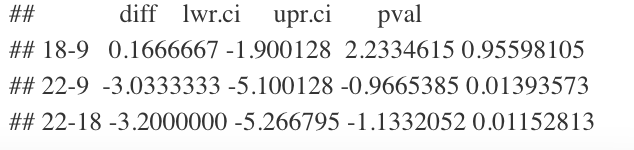
RERCBD:CRD = pf, where pf = and

RERCBD:CRD > 1 means that the RCBD is a more efficient design than CRD. RERCBD:CRD > 1.25 indicates an effective blocking scheme in general. Estimate of is 0.5044, which is directly from MSE. Moreover, the estimate of can be calculated by taking the weighted average of MSE and MSblock. This is weighted by degrees of freedom, which yields 0.94525 from . Thus, the estimate of RERCBD:CRD is equal to 1.874 multiplied by pf (0.918), which yields 1.72. Hence, we conclude that the RCBD blocking the day of week factor is an efficient blocking scheme compared to CRD.

**6. Post ANOVA**

We then implemented a post ANOVA test to further investigate the significance of each pairs. Among three post hoc tests, our group chose to implement Tukey’s HSD. Tukey’s method is a better estimator when all the groups have the same sample size because it’s Honestly Significant Difference controls the familywise error rate to be . Our model has the same sample size of all groups so we thought that Tukey’s HSD would be the best interpreter.

Our group used R to implement the Tukey’s HSD test and below is the result summary of Tukey’s test:

** **

*<Table 6.1 - Full model of Tukey’s Post hoc test(right), Time variable Post hoc test(left) >*

In this model, we are going to focus on the time variable’s post ANOVA test. The table above on the left is the full model analysis performed with Tukey’s post hoc test, and on the right solely interprets the time variable of Tukey’s test. According to the table, among the three pairwise comparisons, two of them were highly significant with a p-value less than 0.05. Those two include 22-9 and 22-18. Therefore, we conclude that the pairs 22-9 and 22-8 are significantly different at =0.05 using Tukey’s method. This result was also observable in the data itself, since we realized that the driving at 10:00 PM usually took less than the time we drove at 9 AM and 6 PM.

Although two of the three pairwise comparisons were significant, our group did not see much difference between the driving time of 9 AM and 6 PM. We interpret that between those two time slots, there are not much differences in driving time it takes from Target to Capitol. We expect that this happens due to the rush hour, where the traffic level is at its highest of people driving to work and returning to home.

**7. Conclusion**

Our group performed a Randomized Complete Block Design to test the effect of a difference in the time of the day one drives a car from Hilldale to the Capitol. With the significant p-value of our ANOVA analysis, we conclude that there exists a significant difference in the time of the day one drives a car to work. We also found that our blocking scheme is efficient and effective both informally and formally by checking MSE and relative efficiency of the RCBD to CRD, respectively.

We then implemented a further post hoc analysis on the variable ‘time’ in order to test the pairwise comparisons. Our group’s interest was “which time groups were significantly different from one another?”. We analyzed the post hoc test using Tukey’s method and came up with the conclusion that two out of three pairs were highly significant. Those two pairs were the ones we expected to be different, which was 10PM - 9AM, and 10PM - 6PM. Logically, we expected that those two pairs will be significantly different due to the high volume of traffic during the rush hours of 9 AM and 6 PM.

For the future improvement, we can add other days of week like weekends to compare the driving time on weekdays with weekends. Also, more than one person can test drive to see whether different drivers affect the driving time by blocking the day of week and time slot for each day. To test this, Latin-Square Design should be used; however, the Latin-Square design should be treated carefully. It is because our current model has ‘p’ less than five and having ‘p’ less than five is not so practical to use for the Latin-Square design.

**7. R code**

library(DescTools)

#reading the csv data

df<-read.csv('/Users/irenecho/Desktop/drivingtime.csv')

#setting time and day variable as a factor

df$time <- as.factor(df$time)

df$day <- as.factor(df$day)

#creating the ANOVA table

mod <- lm(minute ~ time + day, data = df)

anova(mod)

#perform a Post Hoc test on 'time' variable

PostANOVA <- PostHocTest(aov(mod), method = "hsd")

PostANOVA$time