Assignment 3: Design of experiments

TMA4267 Linear statistical models spring 2020

Silje Anfindsen

20 April, 2020

Introduction

We have all experienced the annoying feeling when the internet speed is slow. Maybe it takes ages to download the 2-minutes-short video on YouTube, or it does not download at all. Then you maybe try to move your computer closer to the router, or ask your sister to turn off the wifi on her phone to give your device a greater part of the internet in the room. So, what is it that makes the internet speed, more specific the download-speed fast or slow? Which factors play a big role in this experiment, and which do not? In this study I want to investigate and identify which factors contribute the most to a slow download speed on my computer.

Selection of factors and levels:

As already discussed, we all have an idea of which factors that may affect the network speed. I have chosen to investigate four factors, some which may seems obvious. I have also chosen two different levels for each factor. High level is denoted by *+1*, and low level as *-1*. I tried to find levels giving the factors enough room to possibly variate enough for me to observe it. The table below displays these factors together with the two levels.

Factor	*Level: +1*	*Level: -1*
A	Busy computer	No program running
B	Blutooth devices nearby	No bluetooth nearby
C	Distance: seperate room	Negligble distance
D	Obstacle around router	No obstacles

A. I am going to run a heavy program, more explicit a numerical scheme running in Python at the same time as I measure the download speed and compare it with no program running on the computer (except the speedometer).

B. I will investigate if the effect of bluetooth devices near the router will affect the internet speed. I will use two phones with bluetooth as well as two headsets connected to bluetooth.

C. Here we will measaure the internet speed for the device located in another room (about 7 m away) from the router vs. the device being as close as possible.

D. The last factor is meant to represent possible objects blocking the signal. Therefore I will put the router inside a pot made of iron and measure the signal compared to no pot blocking it.

I expect the increased distance from the router (factor C) will interact with the obstacle around the router (factor D). I find it quite easy to check that the factors are at the desired level as everything is visible. The most challening factor has to be the busy computer (factor A), as it is hard to define a busy computer. It is

also hard to know when the computer is struggeling the most, and when it is not struggeling at all. In this case I will try to notice when the computer is most overheated in orer to define a busy computer.

Selection of response variable:

In order to investigate the internet speed I will measure use the download speed for the network in my home as response. To do this, I will use my own laptop, a 1.5-year old Macbook Pro as device. The download speed is measured using the public webpage: speedtest.net. The unit is Megabits per second, Mbps. According to an article on allconnect.com, [1] a good internet speed is 25 Mbps, or above. Fast internet speeds that make your internet connection able to support multiple devices at once is in the 100+ Mbps range.

In this case I have chosen to measure the download speed as this is what we usually exploit when being on the internet; watching a video, uploading a webpage etc. The response variable could also be the upload speed if we were interested in how fast we can upload our photo album to icloud for example. In this experiment I will use the 5 GHz frequency, but we could also have measured the internet speed using 2,4 GHz, which often has better coverage but less speed. These are the two usual frequency available in most households.

The speedometer I have found is free and probably not the most advanced and precise instrument. On the other hand, it performs several measurements during some seconds in order to find the average of the download speed.

Choice of design:

My experiment consists of 4 factors and as the costs related to the experimentsg are low, I have the possibility to perform a full 2^4 factorial design. If some main effects are confounded with some 2-level interactions, we say that the resolution of design is 3. The desired resolution is usually as high as possible. For full factorial designs we have noe confounding and therefore the resolution is said to be "infinity".

Blocking can be used to minimize the effect of nuisance factors in the design and thereby prevent them from obscuring the main effects. As I cannot find any clear nuisance factors, that is factor with effect on the response but of no interest, blocking is not used in this experiment.

Before starting the experiments I have tested the webpage-measurement of the download speed. It seems to be a bit unstable, and sometimes it gives results that are very off when I cannot explain why. Therefore I decide to do two repititions during each experiment, and then have two observations of the response per row, which I will find the mean for. Alternativly I could have performed two replicates where the measurement are taken during identical but different experimental runs, often with randomization. Repetitions is easier and as I expect it will contribute to decrease the effect of the unexplainable unstability of the instrument, it is preferable.

Implementation of the experiment:

According to Tyssedal [2], an experiment should always be performed in a randomized order as randomization is our best gurantee to obtain independent observations and reduces the chances for external factors to influence our response. Therefore I have made sure to randomize the runs in R before doing the experiments.

As explained above, the instrument has a tendency to be a bit unstable, this was also the case during some of the level combinations. Therefore, at the end of the experiment I had to repeat and replace the response for two of the level combinations where the result was clearly disturbed. Except for this, the design of experiments was implemented as planned.

A genuine run replicate is a repeated run that is subject to all the the sources of error (unexplained variation in a collection of observations) made at different conditions. In this design I decided to do repitions, but without reseting the experimental setup between reptitions. Therefore the experiments do not reflect the total variablility of the experiment, and cannot be called genuine run replicates.

Analysis of data

We will now setup the design matrix and fill in the observed responses. Recall, the mean of the response is calculated for each row. The design matrix with the 32 response values, with randomization, is printed below.

```
# setting up the 2^4 experiment with randomization
set.seed(123)
plan <- FrF2(nruns=16, nfactors=4, randomize=TRUE)

# A=busy computer
# B=bluetooth
# C=distance
# D=obstacle

# observed response, y ( 32 observations x 2)
y1 <- c(88, 116, 293, 107,  97, 282, 198, 182, 230, 259, 284, 294, 293, 96, 231, 293)
y2 <- c(75,  72, 236,  61, 294,  88,  89, 182, 197, 293, 292, 294, 287, 39, 229, 293)
y <- (y1+y2)/2

# add the responses to the design
plan <- add.response(plan,y)
#plan$y <- as.numeric(as.character(plan$y))
plan
```

```
##     A  B  C  D     y
## 1  -1  1  1  1  81.5
## 2   1  1  1  1  94.0
## 3  -1  1 -1 -1 264.5
## 4   1 -1  1  1  84.0
## 5   1 -1 -1  1 195.5
## 6   1 -1 -1 -1 185.0
## 7   1 -1  1 -1 143.5
## 8  -1 -1  1 -1 182.0
## 9   1  1 -1 -1 213.5
## 10  1  1 -1  1 276.0
## 11 -1  1  1 -1 288.0
## 12 -1 -1 -1 -1 294.0
## 13 -1  1 -1  1 290.0
## 14 -1 -1  1  1  67.5
## 15  1  1  1 -1 230.0
## 16 -1 -1 -1  1 293.0
## class=design, type= full factorial
```

Now, do a R analysis with a linear regression model for the experiment.

```
#linear regression model
lm4 <- lm(y~(.)^4,data=plan)
summary(lm4)
```

```
##
## Call:
## lm.default(formula = y ~ (.)^4, data = plan)
```

```
##
## Residuals:
## ALL 16 residuals are 0: no residual degrees of freedom!
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  198.875         NA      NA       NA
## A1           -21.188         NA      NA       NA
## B1            18.312         NA      NA       NA
## C1           -52.562         NA      NA       NA
## D1           -26.188         NA      NA       NA
## A1:B1          7.375         NA      NA       NA
## A1:C1         12.750         NA      NA       NA
## A1:D1         10.875         NA      NA       NA
## B1:C1          8.750         NA      NA       NA
## B1:D1         -5.625         NA      NA       NA
## C1:D1        -38.375         NA      NA       NA
## A1:B1:C1     -10.312         NA      NA       NA
## A1:B1:D1       2.562         NA      NA       NA
## A1:C1:D1       4.812         NA      NA       NA
## B1:C1:D1     -15.438         NA      NA       NA
## A1:B1:C1:D1   -0.625         NA      NA       NA
##
## Residual standard error: NaN on 0 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:    NaN
## F-statistic:   NaN on 15 and 0 DF,  p-value: NA
```

To obtain the estimates effects we multiply the coefficients above with 2.

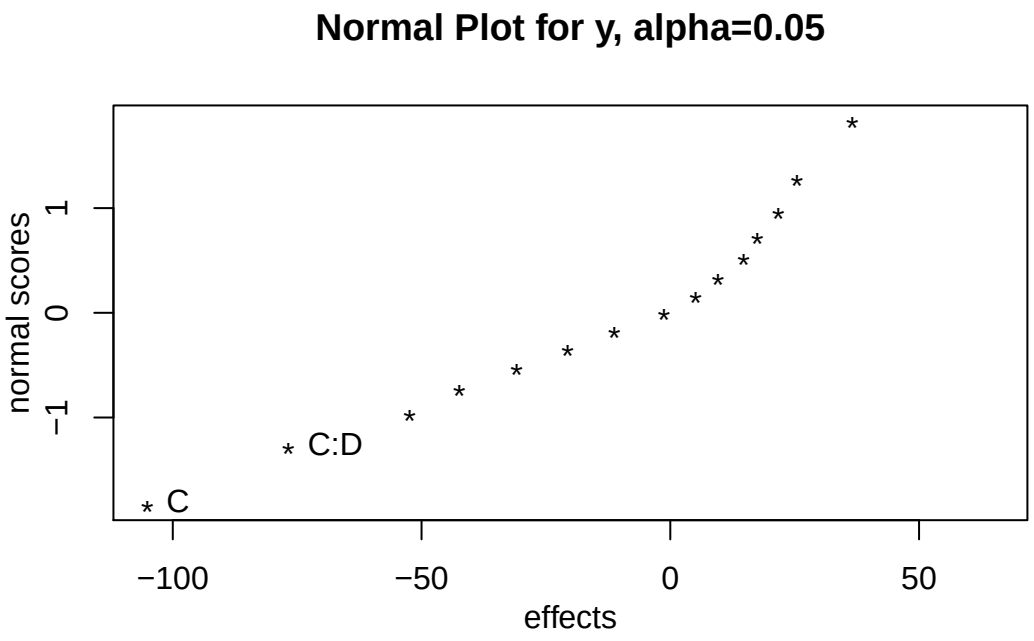
```
#estimated effects
effects <- 2*lm4$coeff
effects
```

```
## (Intercept)          A1          B1          C1          D1       A1:B1
##     397.750     -42.375      36.625    -105.125     -52.375      14.750
##       A1:C1       A1:D1       B1:C1       B1:D1       C1:D1    A1:B1:C1
##      25.500      21.750      17.500     -11.250     -76.750     -20.625
##    A1:B1:D1    A1:C1:D1    B1:C1:D1 A1:B1:C1:D1
##       5.125       9.625     -30.875      -1.250
```

Checking statistical significance

We will now use the normal probability plot of the effects to determine the magnitude, direction, and the importance of the effects. Here we know that effects that are further from zero, are more significant.

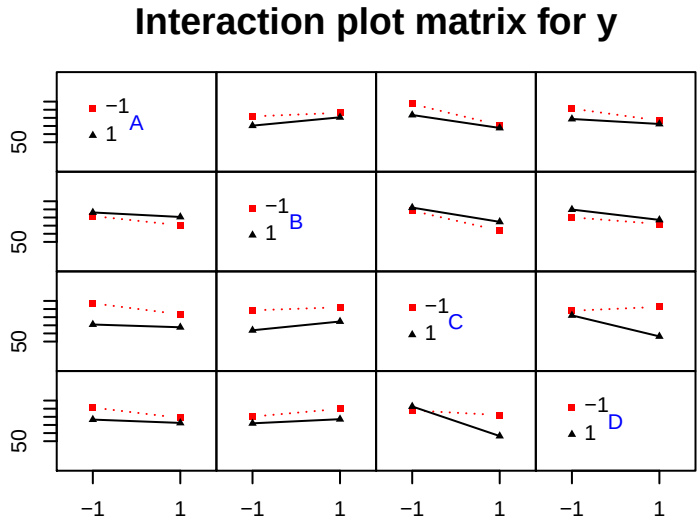
```
DanielPlot(lm4)
```



The normal plot indicates that the main effect *distance* (C) and the interaction between *distance* and *obstacle* (CD) are significant. We can also see the direction of the effect on the response. As the two factors has negative standardized effect, we know that when the distance increases and when the obstable is blocking the signal, the response (*download speed*) will decrease. This is as expected.

From the print-out of the estimated effects we notice that the two factors C and CD clearly are biggest in absolute value, ande therefore indicating a higher significance. Usually, the three-factor and four-factor interactions are small and therefore negligeble. We can illustrate this through a interaction plot.

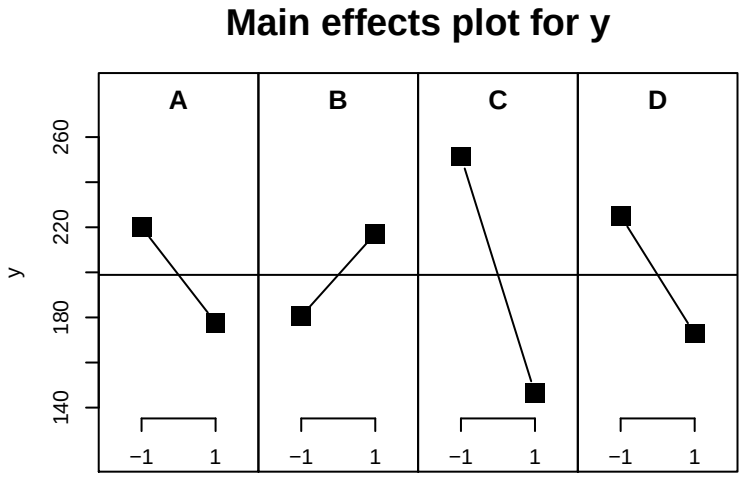
```
IAPlot(lm4)
```



The plot illustrate interatictions between main factors. If there is no interaction between two factors the effect of one factor is the same independent of the level of the other factor. The lines in the plot will then become parallel. This seems to be the case for most of the factors except *obstacle* (D) and *distance* (C). In order to maximize the outcome (*download speed*), D and C should be on their low level (*-1*).

Let us now have a look at the main effects plot.

```
MEPlot(lm4)
```



None of the main factors have a horisontal line with the response, in other words they all affect the response in different ways. The *distance* (C) has the steepest slope and therefore seem to be most significant as assumed. We also notice that for *distance* (C), *heavy program running* (A) and *obstacle* (D) the network speed increases when these factors are at their low level, which in this case is no heavy program, no obstacle and negligeble distance from router. For *bluetooth* (B) devices it seems to be the opposite, where the *downlod speed* actullay increases with more devices using bluetooth nearby. This may seem suprising, but since the change in response is so small, the factor may have no affection in reality.

Checking the assumptions

We will now check the normallity assumptions of the model. First, make a reduced model without any 3rd and 4th order interactions.

```
lm2 <- lm(y~(.)^2,data=plan)
effects <- 2*lm2$coeff
effects
```

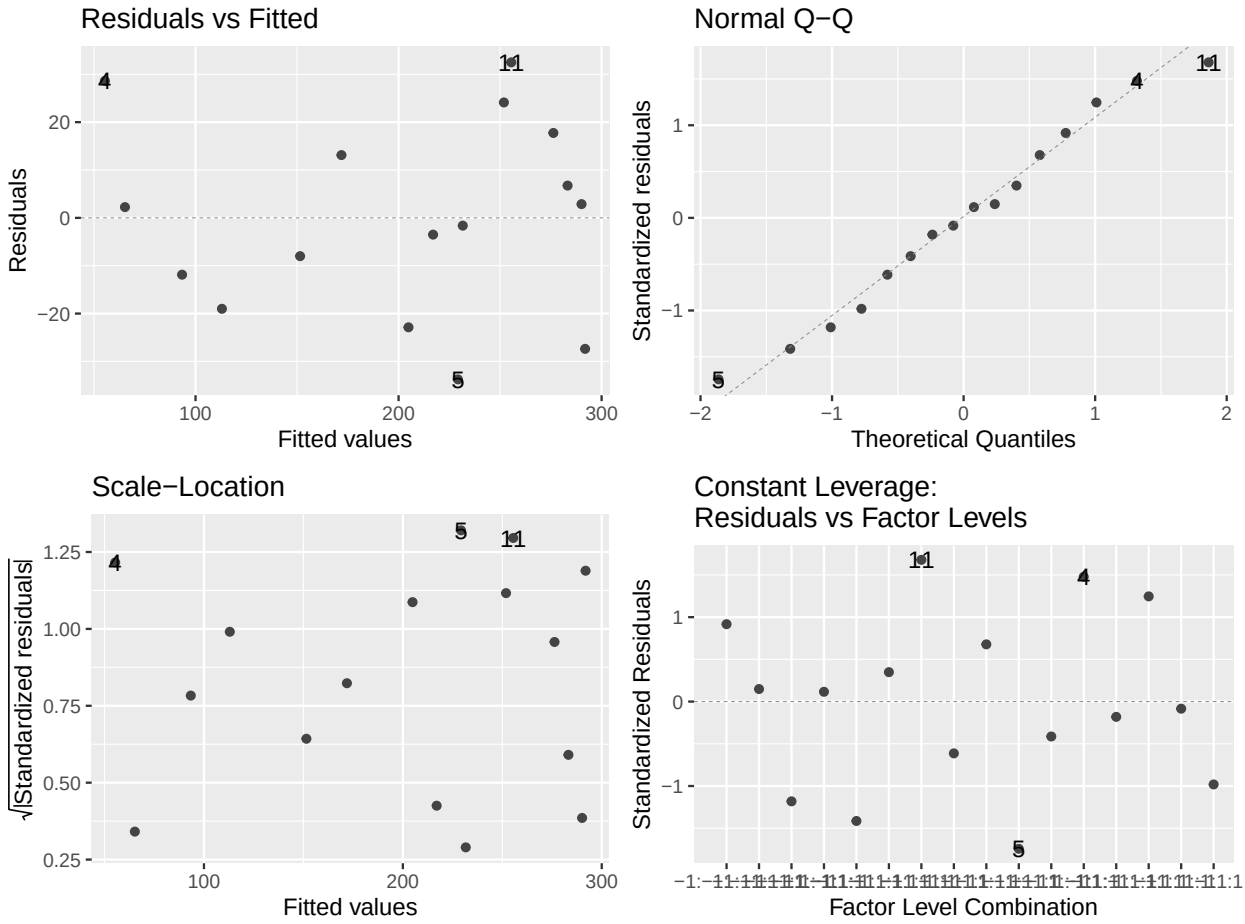
```
## (Intercept)          A1          B1          C1          D1       A1:B1
##     397.750     -42.375      36.625    -105.125     -52.375      14.750
##       A1:C1       A1:D1       B1:C1       B1:D1       C1:D1
##      25.500      21.750      17.500     -11.250     -76.750
```

We will now carry out a residual analysis for the reduced model. We want to verify the assumptions of the linear model, which are:

1. The expected value of ϵ_i is 0: $E(\epsilon_i) = 0$.

2. All ϵ_i have the same variance: $Var(\epsilon_i) = \sigma^2$.
3. The ϵ_i are normally distributed.
4. The ϵ_i are independent of each other.

```
# Residuals plots standard chart
autoplot(lm2,smooth.colour = NA)
```



We start by taking a look at the residual vs. fitted plot. Here we cannot see any clear pattern and the points seem to fall randomly on both sides of zero. We can therefore assume the residuals to be randomly distributed and have constant variance. There does not seem to be any trend in the Scale-Location plot either, reinforcing that the homoscedasticity assumption should be correct. The QQ-plot does not show any evidence against normal distributed residual as the residuals seem to follow a straight line. We have confirmed that there are noe clear evidences against the assumption of normal residuals.

Conclusion and recommendations:

From the experiment it seems as there are few factors that affect the network speed, in other words only the obstacle (D) and distance (C) makes a remarkable change on the response. I expected distance to be the greatest estimates effect, which the experiments seems to agree with.

The plots that I have presented in this report agrees with the conclusion. It seems to be a remarkable interaction between *obstacle* (D) in front of router and *distance* (C), the other interactions are much smaller. Also, the estimated effects show that the network speed will decrease with an increasing distance and obstacle around router as we could expect.

So, to all the people out there who are annoyed with a slow internet speed and want to fix it, this is my conclusion. Several bluetooth-devices (B) near the router will most likely not affect the download speed your device experience, neither will a heavy program running on your computer do (A). But the distance you locate your device from the router, as well as possible furniture blocking the signal, placed in front of the router, can have an affect on the download speed.

References

[1] Anders, David (2019), Internet speed classifications: What's fast, what's slow and what is a good internet speed?

[2] Tyssedal, John Design of Experiments