NTNU

TMA4267 Linear Statistical methods

May 11, 2020

The Dog Poop Project, Compulsory exercise 3

⋆ Olav Milian Scmitt Gran

Abstract

In this experiment I look at how far from the house door the family dog poops depending on the factors time of day, speed, food, and route. After doing the experiments, I reach the conclusion that the family dog poops after 340 meters independent of the factors and levels I have chosen with a significance level $\alpha = 5\%$.

Introduction

How far away from the house door can I get before the family dog needs to poop? This is interesting because it will give a picture of how far away from the house door I can expect to get before the dog needs to poop. Meaning that I for instance know if I need a bag to pick up the poop or if we have i.e. reached the forest, where we don't need to pick up if the dog does not poop on the path. From earlier walks with the family dog I know that the dog usually poops in the start of the exercise, usually also in the same spot depending on the route. With this information in mind I want to get a clear estimate of how far from the house door the family dog will poop, therefore I call it The dog poop project.

Setup

To achieve this I have set up a 2^4 factorial design with the factors time of day, speed, food, and route. For time of day the levels Morning, meaning before 10:00, and evening, after 15:00, because the dog usually wants to walk once in the morning and once in the evening. This also allows for a max and average of two experiments per day, which is needed because of time constraints given by other courses I am taking. For speed, I will be using the levels running and walking, because I would expect that we come further when running than walking, if where the dog poops is dependent on time since we left the house. Food's levels are food, meaning the dog has eaten in the last $1-2$ hours, usually 30 minutes, and no food, meaning that the dog has not eaten in the last $2-3$ hours. I see this as relevant because the dog often wants to go out for pooping after it has eaten food. For route, the levels are urban and forest, since the dog seems to have different areas where it likes to poop, depending on the route. My impression is that it does not like to poop on paved areas. This is at least what I have experienced in earlier walks whit the dog. Note that if the dog does not poop the experiment is seen as invalid and needs to be repeated at a later date. I also expect an interaction between time of day and food, because the family dog gets food for breakfast in the morning (07:00-09:00) and at dinner in the evening (17:00-19:00).

To ensure that the factors are at the desired level, I set up a plan for which experiment to do when, where experiments that may be invalid will be repeated in the end after all 16 experiments are done. For time of day, I need to be aware of the time of day and if the dog wants to go for a walk. For instance I need to wake up when I heard the dog wake up in the morning, and I should let the rest of my family know that I am walking the dog. For speed, the dog knows from my clothing if we are running or walking, training clothes for running and normal clothes for walking. Food is easily controlled if I am the person which gives the dog food or if I clearly ask if the dog has had food. Also telling the rest of my family if the dog needs to walk with me before or after food. Route is also quite easy to control since the urban route is to the left and the forest route is to the right of the house door, so here I just need to be decisive in which direction we are going. Here I need to make it clear that the health and well being of the family dog comes first, so the dog was not harmed in the experiments.

The only response variable, y, will be the distance from the house door measured in kilometres with two decimals. I could also measure from where the dog defines as "home" because the dog does not poop on its own home turf when going on a walk, but this would only offset all measurements by a constant and therefore does not influence the experiment. The response variable is measured via a Garmin Forerunner 235 GPS-watch on my left arm. GPS signal must be enabled and ready before running or walking from the house door. GPS tracking starts when leaving the house door, and ends when and where the dog poops. The accuracy of the measurements are then given by GPS accuracy, which on Garmins support page is given to be within 15 meters 95% of the time (1).

As said before a 2^4 factorial design is chosen. The desired resolution is IV, meaning $1 = ABCD$, and a blocked design is not necessary. Replicates are also not needed, but if the dog does not poop the experiment is invalid and needs to be repeated.

	A	B	C	D	y
1	1	-1	1	1	0.38
2	-1	1	-1	-1	0.24
3	1	-1	-1	1	0.21
4	-1	-1	1	-1	0.21
5	1	-1	-1	-1	0.15
6	-1	-1	1	1	0.08
7	1	-1	1	-1	0.14
8	-1	1	1	1	0.77
9	1	1	1	-1	0.25
10	-1	1	-1	1	0.79
11	1	1	1	1	0.49
12	-1	1	1	-1	0.07
13	1	1	-1	1	0.42
14	1	1	-1	-1	0.32
15	-1	-1	-1	-1	0.15
16	-1	-1	-1	1	0.77

Table 1: A: time of day, 1 morning, -1 evening, B: speed, 1 running, -1 walking, C: food, 1 food, -1 no food, and D: route, 1 urban, -1 forest. The measured y values are also given.

The randomized plan for the experiment is given in table 1 above. Note that the plan is random under the restriction that the variable time of day takes a pattern like “morning, evening, morning, evening, . . . “ most of the time, where breaks in the pattern are given by external factor form ordinary life, i.e. a family event, this will again give some randomization back. Also, this pattern is needed because of personal time restraints from other courses I take, so doing two experiments a day on max and average is desirable. The other factors do not impose any restrictions. Each experiment is a genuine run replicate, that is, it reflects the total variability of the experiment. The pattern above is followed for most of the time and when it is not followed its usually because there is no experiment one day, or the dog was on a long run before dinner, which means that the dog needs rest. The experiments should not harm the dog and the dog goes before the experiment, meaning the dogs health and well being is the most important. Also note that we are an active family that do a lot of sports and the dog usually trains one time a day, so doing to experiments per day effects the dog minimally. By this the experiments are assumed independent.

Results and conclusion

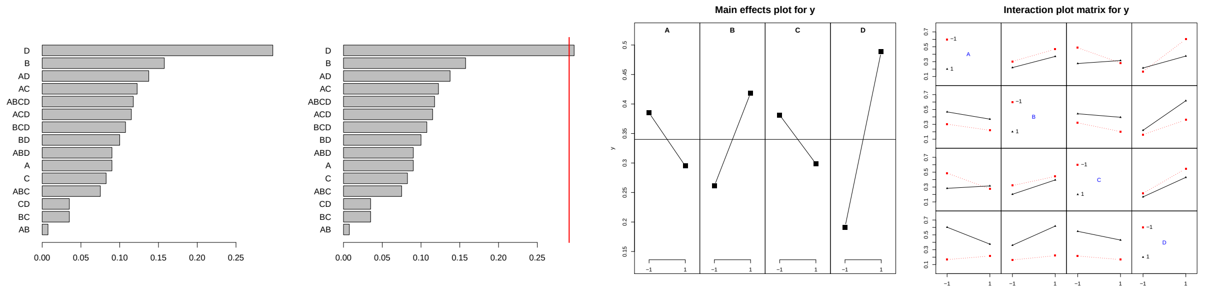

Figure 1: Bar plots for effects with Lengths method with significancelevel $\alpha = 5\%$ and $\alpha = 11\%$, and Main effects plot and interaction plot matrix for y

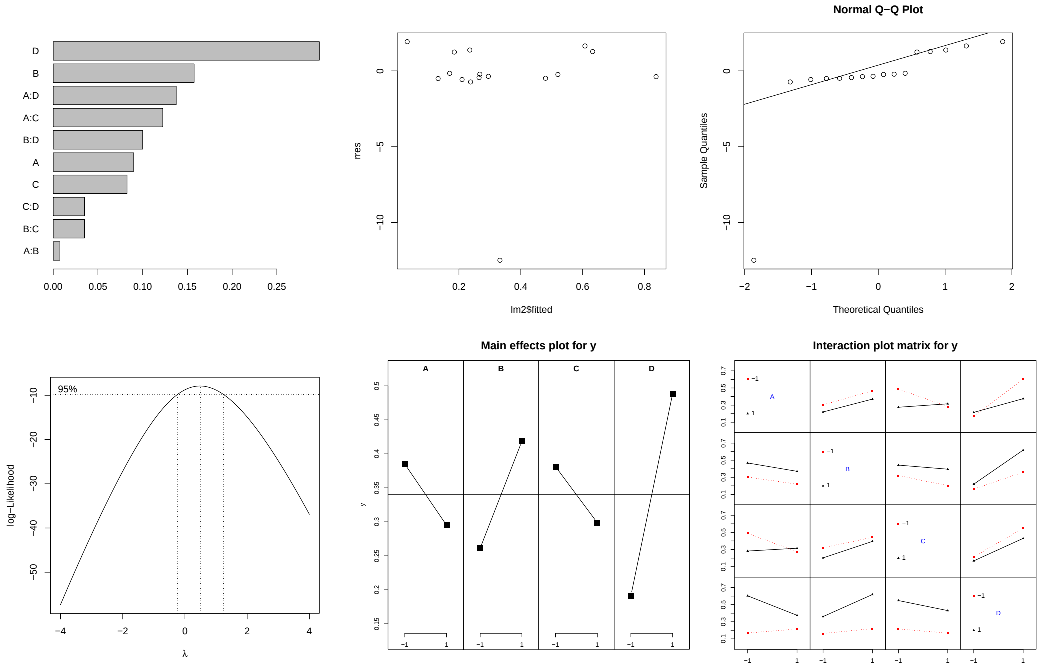

Figure 2: Bar plot, Residual plots, log-Likelihood plot for lm2, and Main effects plot and interaction plot matrix for y

My conclusion from the experiment and analysis done is that the family dog poops after 340 meters independent of the factors and levels I have chosen, with a significance level $\alpha = 5\%$. This is because at significance level $\alpha = 5\%$, all effects can be assumed to be equal to zero, see R printout 4. From significance level $\alpha = 11\%$ and onward, the route plays a role. This also fits with the prior knowledge I have, because the dog's assumed favorite spot for pooping is roughly after 380 and 300 meters for the urban and forest route.

Thinking over this result, i.e. independence on the chosen factors and levels, it may not be an unusual result of an experiment. Finding that one during the design of an experiment has chosen factors which in the end have no influence on the result might happen quite often. For instance in the testing of vaccines or medicines, one would assume that a certain vaccine or medicine could help against a disease, but find that it has no effect. Finding that some factors do not influence the result of the experiment does not mean that all possible factors are insignificant. For instance, in my dog poop project, I believe now, after finishing the experiment with my chosen factors, that the weather or the person the dog walks with might have an influence on the result: The dog does not like rain and it has a different behavior with me than with my sisters or my father. If I had chosen the factors differently, i might have found significant influence on the result.

Apart from this, the interactions in the interaction plot matrix for y, Figure 1 and 2 give us that A and B, and C and D are close to parallel, when A and C, and A and D are crossing each other. The interpretation of the analysis is then that the effect of food is lower when the time of day is late (low level), but higher when the time of day is early (high level), and reverse for A and D.

The residual plot may look off, because of one residual under -12, but $\lambda = 1$ is inside the 95% confidence interval for the log-Likelihood, as seen in Figure 2. The hypothesis test for the factors time of day and food being the same, i.e. $\hat{\beta}_A = \hat{\beta}_C$, using the reduced model, lm2, I get that this is 96% likely (see R printout 4). But again, this could also be because both factors are not significant.

All in all from this design of experiment exercise I get that the family dog poops after 340 meters independent of the factors and levels I have chosen. This does not mean that other factors are insignificant too, but I would need a new experiment to test this.

R printout

Some R printout, (Run code for full R print out):

```
##     A  B  C  D    y
## 1   1 -1  1  1 0.38
## 2  -1  1 -1 -1 0.24
## 3   1 -1 -1  1 0.21
## 4  -1 -1  1 -1 0.21
## 5   1 -1 -1 -1 0.15
## 6  -1 -1  1  1 0.08
## 7   1 -1  1 -1 0.14
## 8  -1  1  1  1 0.77
## 9   1  1  1 -1 0.25
## 10 -1  1 -1  1 0.79
## 11  1  1  1  1 0.49
## 12 -1  1  1 -1 0.07
## 13  1  1 -1  1 0.42
## 14  1  1 -1 -1 0.32
## 15 -1 -1 -1 -1 0.15
## 16 -1 -1 -1  1 0.77
##
## Call:
## lm(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)  0.34000         NA      NA       NA
## A           -0.04500         NA      NA       NA
## B            0.07875         NA      NA       NA
## C           -0.04125         NA      NA       NA
## D            0.14875         NA      NA       NA
## A:B         -0.00375         NA      NA       NA
## A:C          0.06125         NA      NA       NA
## A:D         -0.06875         NA      NA       NA
## B:C          0.01750         NA      NA       NA
## B:D          0.05000         NA      NA       NA
## C:D         -0.01750         NA      NA       NA
## A:B:C       -0.03750         NA      NA       NA
## A:B:D       -0.04500         NA      NA       NA
## A:C:D        0.05750         NA      NA       NA
## B:C:D        0.05375         NA      NA       NA
## A:B:C:D     -0.05875         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
## [1] "--------------------------------------------------"
## (Intercept)           A           B           C           D         A:B
##      0.6800     -0.0900      0.1575     -0.0825      0.2975     -0.0075
##         A:C         A:D         B:C         B:D         C:D       A:B:C
##      0.1225     -0.1375      0.0350      0.1000     -0.0350     -0.0750
##       A:B:D       A:C:D       B:C:D     A:B:C:D
##     -0.0900      0.1150      0.1075     -0.1175
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
## Warning in anova.lm(lm4):  ANOVA F-tests on an essentially perfect fit are unreliable
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value Pr(>F)
## A          1 0.03240 0.03240
## B          1 0.09922 0.09922
## C          1 0.02722 0.02722
## D          1 0.35402 0.35402
## A:B        1 0.00023 0.00023
## A:C        1 0.06003 0.06003
## A:D        1 0.07563 0.07563
## B:C        1 0.00490 0.00490
## B:D        1 0.04000 0.04000
## C:D        1 0.00490 0.00490
## A:B:C      1 0.02250 0.02250
## A:B:D      1 0.03240 0.03240
## A:C:D      1 0.05290 0.05290
## B:C:D      1 0.04622 0.04622
## A:B:C:D    1 0.05523 0.05523
## Residuals  0 0.00000
## [1] "--------------------------------------------------"
## (Intercept)           A           B           C           D         A:B
##        TRUE       FALSE       FALSE       FALSE       FALSE       FALSE
##         A:C         A:D         B:C         B:D         C:D       A:B:C
##       FALSE       FALSE       FALSE       FALSE       FALSE       FALSE
##       A:B:D       A:C:D       B:C:D     A:B:C:D
##       FALSE       FALSE       FALSE       FALSE
## (Intercept)           A           B           C           D         A:B
##        TRUE       FALSE       FALSE       FALSE        TRUE       FALSE
##         A:C         A:D         B:C         B:D         C:D       A:B:C
##       FALSE       FALSE       FALSE       FALSE       FALSE       FALSE
##       A:B:D       A:C:D       B:C:D     A:B:C:D
##       FALSE       FALSE       FALSE       FALSE
##
## Call:
## lm(formula = y ~ (.)^2, data = plan)
##
## Residuals:
##       1       2       3       4       5       6       7       8       9      10
##  0.1450 -0.0275 -0.0550  0.1775 -0.0200 -0.2525 -0.0700  0.1375 -0.0450 -0.0475
##      11      12      13      14      15      16
## -0.0300 -0.0625 -0.0600  0.1350 -0.0875  0.1625
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.34000    0.05114   6.648  0.00116 **
## A           -0.04500    0.05114  -0.880  0.41920
## B            0.07875    0.05114   1.540  0.18423
## C           -0.04125    0.05114  -0.807  0.45655
## D            0.14875    0.05114   2.909  0.03346 *
## A:B         -0.00375    0.05114  -0.073  0.94439
## A:C          0.06125    0.05114   1.198  0.28474
## A:D         -0.06875    0.05114  -1.344  0.23663
## B:C          0.01750    0.05114   0.342  0.74613
## B:D          0.05000    0.05114   0.978  0.37315
## C:D         -0.01750    0.05114  -0.342  0.74613
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2046 on 5 degrees of freedom
## Multiple R-squared:  0.7695,Adjusted R-squared:  0.3085
## F-statistic: 1.669 on 10 and 5 DF,  p-value: 0.2976
## (Intercept)           A           B           C           D         A:B
##      0.6800     -0.0900      0.1575     -0.0825      0.2975     -0.0075
##         A:C         A:D         B:C         B:D         C:D
##      0.1225     -0.1375      0.0350      0.1000     -0.0350
## [1] "--------------------------------------------------"
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value  Pr(>F)
## A          1 0.03240 0.03240  0.7742 0.41920
## B          1 0.09922 0.09922  2.3710 0.18423
## C          1 0.02722 0.02722  0.6505 0.45655
## D          1 0.35402 0.35402  8.4594 0.03346 *
## A:B        1 0.00023 0.00023  0.0054 0.94439
## A:C        1 0.06003 0.06003  1.4343 0.28474
## A:D        1 0.07563 0.07563  1.8070 0.23663
## B:C        1 0.00490 0.00490  0.1171 0.74613
## B:D        1 0.04000 0.04000  0.9558 0.37315
## C:D        1 0.00490 0.00490  0.1171 0.74613
## Residuals  5 0.20925 0.04185
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] "--------------------------------------------------"
## Linear hypothesis test
##
## Hypothesis:
## A - C = 0
##
## Model 1: restricted model
## Model 2: y ~ (A + B + C + D)^2
```

```
##
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1      6 0.20936
## 2      5 0.20925  1 0.0001125 0.0027 0.9607
## [1] "--------------------------------------------------"
## (Intercept)           A           B           C           D         A:B
##        TRUE       FALSE       FALSE       FALSE       FALSE       FALSE
##         A:C         A:D         B:C         B:D         C:D
##       FALSE       FALSE       FALSE       FALSE       FALSE
## [1] "--------------------------------------------------"
##           1           2           3           4           5           6
##   1.3768122  -0.2163371  -0.4404738   1.9285398  -0.1569042 -12.4915268
##           7           8           9          10          11          12
##  -0.5692250   1.2754963  -0.3575326  -0.3780861  -0.2362669  -0.5041127
##          13          14          15          16
##  -0.4827474   1.2432285  -0.7283176   1.6460312
##
##  Anderson-Darling normality test
##
## data:  rstudent(lm2)
## A = 2.9399, p-value = 8.808e-08
```

References

[1] Garmin support center, GPS accuracy, https://support.garmin.com/en-US/?faq=aZc8RezeAb9LjCDpJplTY7