TMA4267 Linear statistical models project

candidate numbers 10006 and 10050.

April 2016

1 Introduction

We would like to make a fun exam period procrastination game for the two of us, where we get points for hitting a basket with a ball. Whether one is allowed an easy or a difficult throw will be based on some other factors of the game. As an aid in designing the game, we gather data on initial skill, and perform a linear regression[1] to get information on what factors affect the probability to hit the basket, and how great the element of chance is. Game design already determines type of ball and basket, so what might be varied in addition to person is distance, left or right hand and type of throw.

2 Experiment design and implementation

We are interested in the probability to hit the basket, so the response variable Y is proportion of times we hit the basket out of 20 throws. 20 is selected to not be so big that we would expect to get either better or tired during the course of the the experiments, and to be able to make all the throws in one go. Hit proportion is thus the sum of 20 random variables, and should be normally distributed if 20 is sufficiently large.

To perform a linear regression on the data we assume

$$Y = X\beta + \epsilon \ , \ \epsilon \sim N(0, \sigma^2 I) \ , \tag{1}$$

that Y is a linear function in the parameters, plus some normal distributed error ϵ with variance $\sigma^2 I$, and the zero vector as expected value.

The selected factors were chosen to be: The person throwing, the distance to the basket and the hand used. Type of throw was deselected as it was not very exciting, and we are free to decide that in the game people shall throw upper hand. Distance should represent near and far, and we found these values by trial and error, and markings on the floor that were easy to remember.

Kristin is right handed and Petter is left handed, so a viable option for the hand levels is dominant and second hand. We chose left and right to mirror

[1] This is also a project in TMA4267 on NTNU, and all information and a lot of the code is from the course webpage https://wiki.math.ntnu.no/tma4267/2016v/forelesning

Table 1: Experiment implementation and response.

Experiment	Original order	A	B	C	Hits	Response
1	6.1	1	-1	1	11	0.55
2	3.1	-1	1	-1	6	0.30
3	2.1	1	-1	-1	9	0.45
4	4.1	1	1	-1	7	0.35
5	7.1	-1	1	1	2	0.10
6	5.1	-1	-1	1	12	0.60
7	8.1	1	1	1	7	0.35
8	1.1	-1	-1	-1	14	0.70
9	5.2	-1	-1	1	14	0.70
10	6.2	1	-1	1	13	0.65
11	7.2	-1	1	1	2	0.10
12	3.2	-1	1	-1	6	0.30
13	1.2	-1	-1	-1	13	0.65
14	2.2	1	-1	-1	9	0.45
15	4.2	1	1	-1	5	0.25
16	8.2	1	1	1	6	0.30

the game, as we will not make the rule "this throw should be made with your dominant hand", in case of arguing if someone else were to enter at a later stage. We expect interactions person and both other factors as we assume we have different specialities in throwing. A throw that gives a high hit proportion is easy. The variance and potentially all factors are interesting, so we perform replicates, so that we perform 16 runs in total. This is not too much to do in one go, and we wont do a blocked design.

The factors are easily controlled to be at their appropriate levels. To mimic the game we set up a friendly and competitive mood. First we make some throws from various distances so that all experiments are preceded by throws. If it is not certain whether something is a hit, we decide it in the way we will when playing, so that we estimate probability to get points in the game.

For ease of calculation the experiment is designed to be orthogonal. The order of the runs are randomized as shown in table [1] to make external factors have a random impact. The biggest challenge to the experiments in this implementation representing genuine run replicates is if our skill were to improve with experiment number, or if we were to tire. Then σ would not be constant as we assume. Daily fitness might also impact our skill so that the true variance is greater than the one we measure. We could have blocked the design into good day and bad day, but this is very difficult to plan.

3 Analysis of data

The code used for the analysis is included here for your convenience. We have handed the report in both electronically and on paper so that you might copy

the coe directly into R.

```
library(FrF2)
plan <- FrF2(nruns=8,nfactors=3,replications=2,randomize=
    ↪ FALSE)
plan
y <- c(14,9,6,7,12,11,2,7,13,9,6,5,14,13,2,6)
y=y/20
rekkeflge <- c(8,3,2,4,6,1,5,7,   13,14,12,15,9,10,11,16)
plan <- add.response(plan,y)
plan
lm3r <- lm(y~(.)^3,data=plan)

anova(lm3r)
summary(lm3r)
MEPlot(lm3r)
IAPlot(lm3r)

effects <- 2*lm3r$coeff
effects

n <- length(y)
sest <- summary(lm3r)$sigma # which is the same sqrt(MSE)
sest # \hat{sigma} in text
seffect <- 2*sest/sqrt(n) #Var(2\beta)
seffect
# The cut-off for significance is
signcut <- qt(0.975,df=8)*seffect

# Pareto plot
names(effects) = c("Interaction", "A", "B", "C", "AB","AC
    ↪ ", "BC", "ABC")
barplot(sort(abs(effects[-1]),decreasing=FALSE),las=1,
    ↪ horiz=TRUE)
abline(v=signcut,col=2,lwd=2)

# Residual plots
rres <- rstudent(lm3r)
qwerty= lm3r$fitted

plot(lm3r$fitted,rres,ylab = "residual",
xlab = "Estimated␣proportion␣of␣success")

plot(rekkeflge,rres,xlab = "Experiment␣number",ylab = "
    ↪ Residual")
```

```
# 3 normality of residuals
qqnorm(rres)
qqline(rres)
library(nortest)
#Anderson-Darling normality test
ad.test(rstudent(lm3r))

# Estimates
betas = matrix(c(-0.16875,        0.0625, 0.05,    -0.0375),
nrow=4, ncol=1, byrow = TRUE)
intercept = 0.42500
#Easiest to hardest X
Xeasy = matrix( c(-1,1,1,1, -1,1,-1,-1, -1,0,0,-1,
    ↪ -1,0,0,1,             -1,-1,-1,-1,  -1,-1,-1,1,
    ↪ 1,1,1,1,   1,1,-1,-1,
1,0,0,-1,  1,0,0,1,  1,-1,-1,-1,  1,-1,-1,1),
                    nrow=12, ncol=4, byrow = TRUE)
yhatEasy = Xeasy %*% betas + intercept
plot(c(yhatEasy[1],yhatEasy[2],yhatEasy[11],yhatEasy[12])
    ↪ ,c(0,0,0,0),
xlim= c(0.1,0.7), xlab =  "Hit proportion",ylab="",col="
    ↪ red", axes=FALSE)
Axis(side=1)
points(c(yhatEasy[3],yhatEasy[4],yhatEasy[9],yhatEasy
    ↪ [10]),c(0,0,0,0))
points(yhatEasy[5:8],c(0,0,0,0),col="cyan")
```

3.1 Linear model

The summary of the regression is displayed in table [2]. The effects are twice the value of the βs. The p-values from

$$\mathrm{H}_0: \ \beta_i = 0 \qquad \mathrm{H}_1: \ \beta_i \neq 0\,,$$

are displayed in column 5. With a significance level of 0.05 the significant effects are thus B1, A1:B1, A1:C1, B1:C1. These are the same effects that surpass the red line in the Pareto plot in figure 1, which means that they have size greater than

$$|t_{0.025,8} \cdot s_{eff}| \tag{2}$$

where s_{eff} is the estimated standard deviation of the effects. The estimate for σ is $\hat{\sigma} = 0.04677$, which makes $s_{eff} = 0.02339$. $R^2 = 0.97$ which means that 97% of the data is explained by the regression. As the design is orthogonal the three smallest parameters might be removed, and this will not affect the values of the other parameters. Since

$$R^2 = \sqrt{n\beta_1 + n\beta_2 + \cdots}\,, \tag{3}$$

Table 2: Summary

	βs	Std. Error	t value	Pr($>|t|$)	Effects
(Intercept)	0.42500	0.01169	36.348	3.6e−10	
A	-0.00625	0.01169	-0.535	0.60751	-0.01250
B	-0.16875	0.01169	-14.432	5.2e−7	-0.33750
C	-0.00625	0.01169	-0.535	0.60751	-0.01250
AB	0.06250	0.01169	5.345	0.00069	0.12500
AC	0.05000	0.01169	4.276	0.00270	0.10000
BC	-0.03750	0.01169	-3.207	0.01248	-0.07500
ABC	0.00625	0.01169	0.535	0.60751	0.01250

the amount of the data explained by the regression will change very little as the parameters are removed, and so will $\hat{\sigma}$. The model we use is

$$\hat{Y} = 0.42500 - 0.16875x_A + 0.06250x_{AB} + x_{AC} - 0.0375x_{BC} \tag{4}$$

as this reduce the risk that we are over fitting the data.

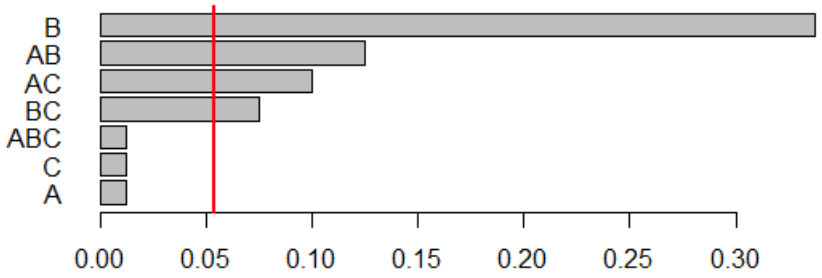

Figure 1: Pareto plot of the main effects

3.2 Check of assumptions

For the obtained linear model to be applicable the regression assumptions from equation (1) must hold. The residuals are the part of the observed Ys that are not explained by the regression. They are what we have estimated the error to be, and should appear to be drawn from the normal distribution $N(0, \sigma^2)$. Figure 2 displays the residuals plotted against estimate $\hat{Y}$. The residuals seem to be centered at zero, and the variance does not seem to change.

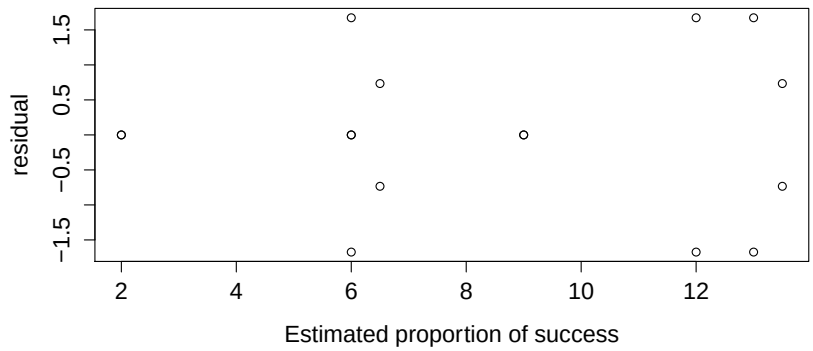

Figure 2: Residuals plotted against estimated proportion of hits.

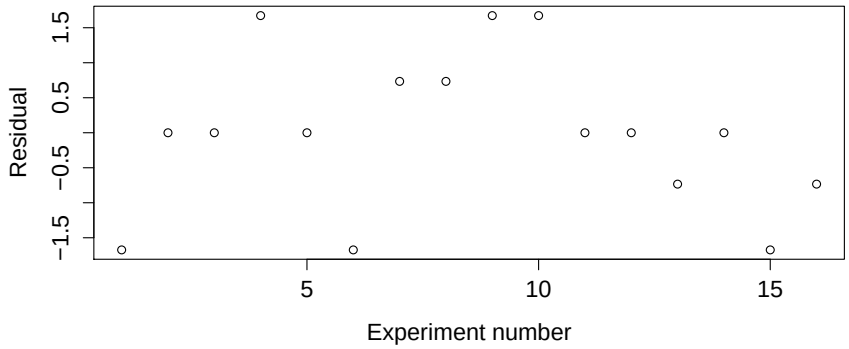

Figure 3: Residuals plotted against experiment number.

To check if our skills improved during the course of the experiments such that the variance is smaller or the expected value is higher in late experiments we plot the order of the experiment against the residuals in figure 3. This also looks like a random plot with no tendency to the residuals, although it is not as pleasing to the eye as the last third of the data have residuals below or at zero. However with so few points, it is not unlikely for the plot to look like this if the residual value is independent of experiment number.

Next we test the normality by looking at the QQ-plot in figure 4 of the quantiles of the measured data against the quantiles of the normal distribution with the estimated σ. The points lie around the theoretical line like they should if the assumptions are correct, although we would have preferred to have more

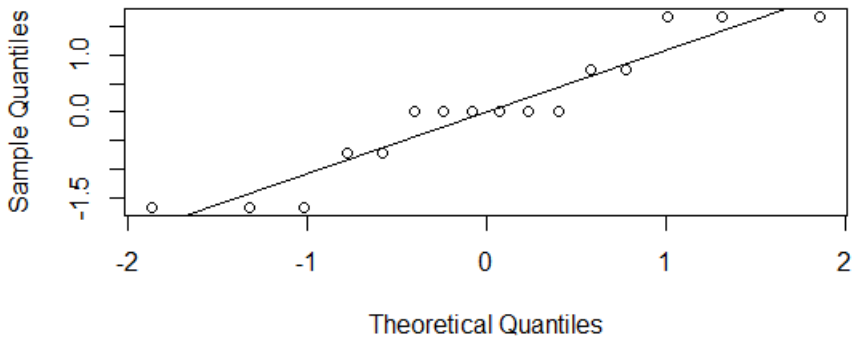

Figure 4: QQ-plot for normal distribution.

of the points on the line.

We also perform the Anderson-Darling test, and get p-value 0.09. As the null hypothesis of the Anderson-Darling test is that it was correct to assume normality, we do not reject our assumptions with confidence level 0.05.

As none of the tests we have performed to assess the assumptions indicate something wrong, we accept the assumptions to hold, and do not search for a transformation for our data with a Box-Cox plot.

3.3 Inference

From the size of the main effects in figure 1 we know that distance (B) has the most to say for how difficult it is to hit. Who is throwing or with which hand, is insignificant when the information is not combined, which means that we are equally talented at throwing. The interaction effects between person and both of the other parameters are significant, which means that we have different strengths.

To learn more we must look at the effect plots. The main effect plots are not so interesting as only one of the main effects are significant, and the negative value of the β in table [2] tells us that it is easier to hit when the bucket is close. None of the lines are parallel in figure 5, as all the interaction effects are significant. In the second row distance is short for the red line, and the red line is always above the black line. This shows that the hit percentage is better when we are close. In the rest of the matrix the lines cross, and this goes well with that the other main effects were not significant.

First row second column displays interaction effects for distance and person. Kristin is relatively better at short distance and Petter is relatively better at long distance; Kristin is more sensitive to distance than Petter.

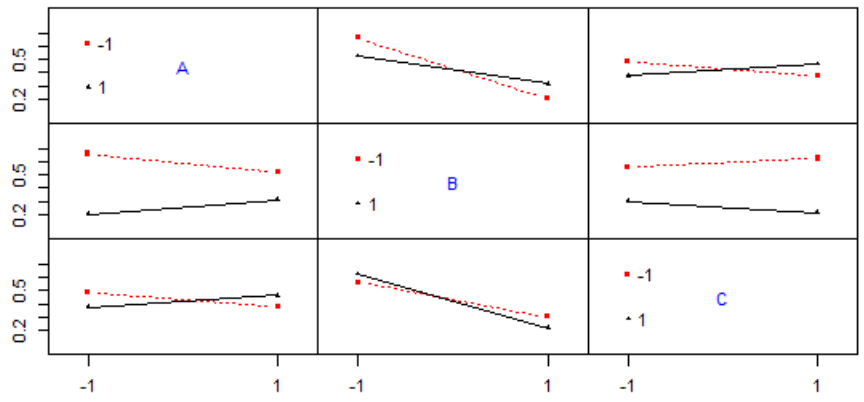

Figure 5: Interaction plot matrix for proportion of hits. High A is Petter, high B is far, high C is left hand. The row letter denotes the variable on the y-axis, with red line for low and black line for high. In first row second column the slope of the red line indicates main effect of B when A is at low level. Interaction effect is seen as difference in slope.

Table 3: Throws by person ordered from easiest to most difficult

Average	Close, Left	Close, Right	Far, Right	Far, Left
Kristin	Close, Right	Close, Left	Far, Right	Far, Left
Petter	Close, Left	Close, Right	Far, Left	Far, Right

First row third column shows that Kristin throws better with right hand, and Petter throws better with left hand. Left hand (high level of C) is better for close throws and right hand (low level) is better for far throws in this experiment; hit proportion with right hand is more stable with respect to distance.

3.4 Throw difficulty level

Table [3] holds the order of difficulty from easiest as found from the main and interaction effects. The effect of distance is far greater than the other effects, so the two easiest throws will be with short distance. On average, interactions with person will not count, and so the second factor to sort by in the average row is interaction between distance and left or right hand. As the interaction effect of distance and hand is smaller than that of hand and person the rows for Kristin and Petter are secondarily sorted by preferred hand.

This gives a broad image of which throw is better for which person. We return to the regression model from equation (4) to construct figure 6 that shows how much better. In addition the regression model predicts an error term

$\epsilon \sim N(0, 0.047^2)$, such that the predicted value of each is within 9.4 percentage points of expected value with 95% probability.

The confidence interval for the error is large, which is good for the game as an element of chance is fun. An infinity of games would however not give the expected values as average, we would expect to improve if we played infinitely many times.

The model could be used to say that 6 close right handed throws gives the same expected value as 15 far right handed throws, as $0.56 \cdot 6 = 3.36$ and $0.22 \cdot 15 = 3.30$, but that it would benefit Kristin to get 6 of the close throws, and Petter to get 15 of the far throws.

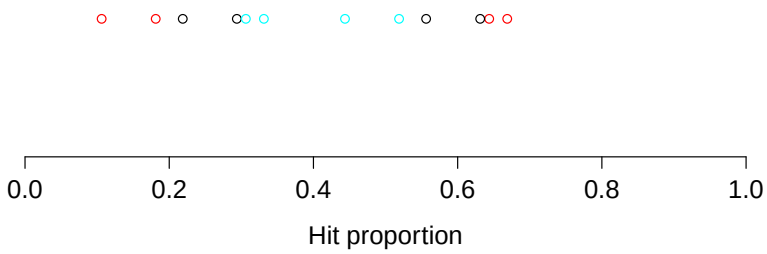

Figure 6: Expected values of proportion of hits by set of parameters. Black is average, red is Kristin, cyan is Petter. See table [3] for the ordering of the other parameters.

3.5 Conclusion

The experiments were conducted in a fashion that made it possible to apply a linear regression model. The assumptions of the regression hold, and the game may now be designed with data from the linear model 4. The difference in skill between Petter and Kristin is not significant, so no handicap will be needed. After the exam period it would be fun to do the exact same experiment. Then we could see if we had made an improvement in skill, and whether that improvement was symmetrical or asymmetrical. Another thing we could investigate is equal type throws made from multiple distances.