MA50259: Statistical Design of Investigations

Lab sheet 2: Estimability under non full rank

In this practical you will start learning the basics of how to analyse an experimental design that does not have full rank. Take the time to run each of the following commands and analyse the displayed results to understand what the code is doing.

Estimation in the means model

Consider the means model of a completely randomized design (CRD) with $t = 3$ treatment levels and $r_i = 4$ replicates per level:

$$\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\mu} + \boldsymbol{\epsilon}$$

where $\boldsymbol{\epsilon} \sim MVN(\mathbf{0}, \sigma^2\, \boldsymbol{I})$ and

$$\boldsymbol{y} = \begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{24} \\ y_{31} \\ y_{32} \\ y_{33} \\ y_{34} \end{pmatrix}, \; \boldsymbol{Z} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \; \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}, \; \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{14} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \\ \epsilon_{24} \\ \epsilon_{31} \\ \epsilon_{32} \\ \epsilon_{33} \\ \epsilon_{34} \end{pmatrix}$$

1. Load the corresponding packages to use in this tutorial

```
library(tidyverse)
library(Matrix)
library(MASS)
```

2. Construct the matrix $\boldsymbol{Z}$ in R

```
t<-3; # the number of treatmenet levels
r<-4; # the number of replicates
n<-t*r # total number of experimental units
levels<-c("level 1","level 2","level 3");
fact <- rep(levels,each = r) %>% factor()
#fact<-gl(t,r,labels=levels) # alternative code

crd <- tibble(  treatment=fact )

Z <- model.matrix(~ fact-1)
Z
```

```
##    factlevel 1 factlevel 2 factlevel 3
## 1            1           0           0
## 2            1           0           0
## 3            1           0           0
## 4            1           0           0
## 5            0           1           0
## 6            0           1           0
## 7            0           1           0
## 8            0           1           0
## 9            0           0           1
## 10           0           0           1
## 11           0           0           1
## 12           0           0           1
## attr(,"assign")
## [1] 1 1 1
## attr(,"contrasts")
## attr(,"contrasts")$fact
## [1] "contr.treatment"
```

3. Verify that $\boldsymbol{Z}$ has rank equal to $t = 3$. In this case we say the model is **full rank** since the matrix $\boldsymbol{Z}$ has rank equal to the its number of columns

```
rankMatrix(Z)[1]
```

```
## [1] 3
```

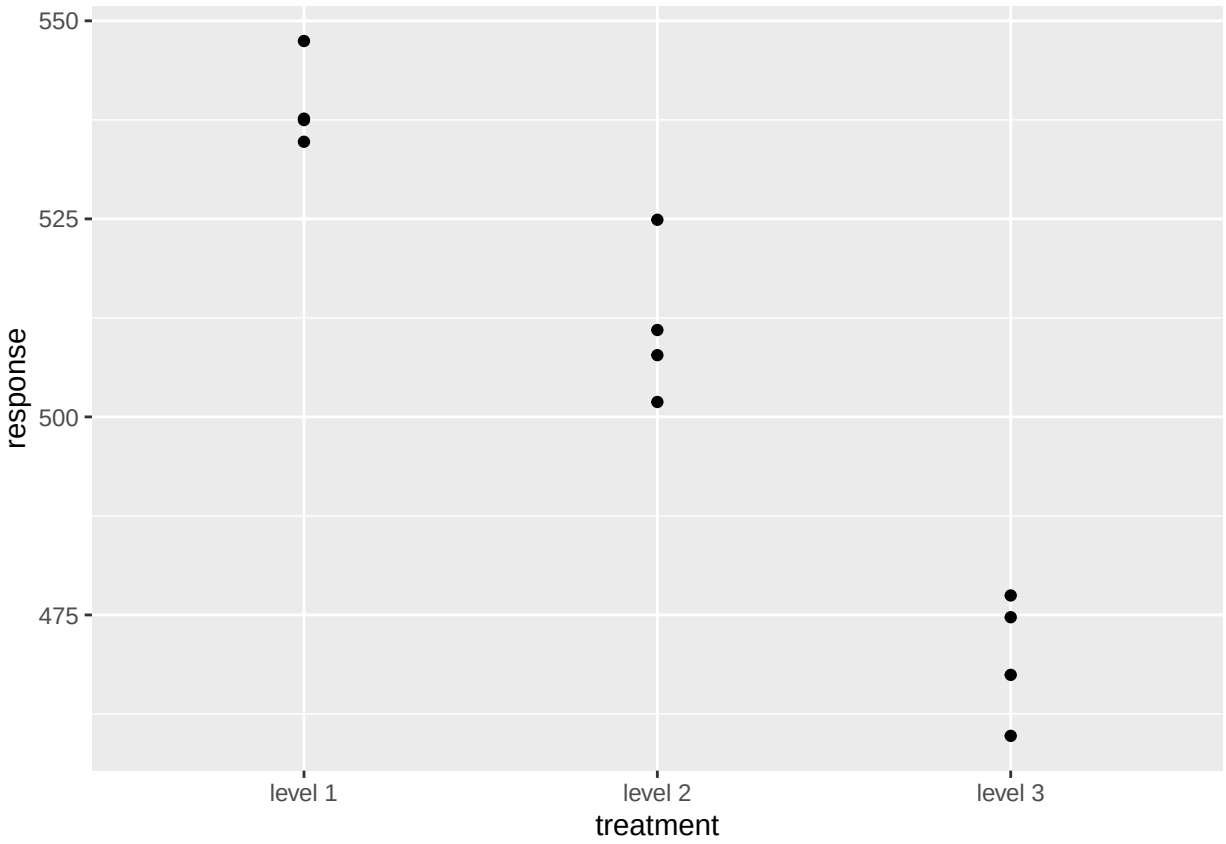
```
#qr(Z)$rank # alternative code
```

4. Simulate the response values and plot them agains the treatment levels

```
set.seed(13579)
mu<-500 # reference value
tau<-c(50,0,-30) # treatment effects= differences wrt to mu
sd<-10 # overall standard deviation
means<-mu+tau %>% rep(each=r) # vector of means
y<-rnorm(n,mean=means,sd=sd)
crd$response<-y
y
```

```
##  [1] 537.6528 537.4717 547.4522 534.7335 510.9711 524.8874 507.7948 501.8838
##  [9] 459.7355 467.4329 477.4605 474.7122
```

```
ggplot(crd,aes(treatment,response))+geom_point()
```



5. The least squares estimate $\widehat{\boldsymbol{\mu}}$ of $\boldsymbol{\mu}$, which under the assumptions stated in the lecture is equivalent to maximum likelihood, is given by the unique solution of the normal equations

$$\boldsymbol{Z^T Z\,\widehat{\mu}} = \boldsymbol{Z^T y}$$

Find the least squares estimate $\widehat{\boldsymbol{\mu}}$ of μ as follows

$$\widehat{\mu} = (\boldsymbol{Z^T Z})^{-1}\boldsymbol{Z^T y}$$

```
S<-t(Z)%*%Z
mu.hat<-solve(S)%*%t(Z)%*%y
mu.hat
```

```
##                 [,1]
## factlevel 1 539.3276
## factlevel 2 511.3843
## factlevel 3 469.8353
```

6. Verify that the entries of $\widehat{\boldsymbol{\mu}}$ are simply the arithmetic means of the response values at each treatment level

```
by_group <- group_by(crd, treatment)
means.crd<-summarize(by_group, means = mean(response))
glimpse(means.crd)
```

```
## Rows: 3
## Columns: 2
## $ treatment <fct> level 1, level 2, level 3
## $ means     <dbl> 539.3276, 511.3843, 469.8353
```

7. Verify that the entries of $\widehat{\boldsymbol{\mu}}$ are also given in the output of the `lm` command

```
mod.crd.means<-lm(response~treatment-1,data=crd)
summary(mod.crd.means)
```

```
##
## Call:
## lm(formula = response ~ treatment - 1, data = crd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.100  -3.841  -1.765   5.564  13.503
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## treatmentlevel 1  539.328      3.975   135.7 3.26e-16 ***
## treatmentlevel 2  511.384      3.975   128.7 5.26e-16 ***
## treatmentlevel 3  469.835      3.975   118.2 1.13e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.949 on 9 degrees of freedom
## Multiple R-squared:  0.9998, Adjusted R-squared:  0.9998
## F-statistic: 1.631e+04 on 3 and 9 DF,  p-value: < 2.2e-16
```

8. Using the invariance property of maximum likelihood estimates, find estimates of the following quantities

- $\mu_2 - \mu_1$
- $\mu_3 - \mu_1$
- $\mu_3 - \mu_2$

```
mu.hat[2]-mu.hat[1]
```

```
## [1] -27.94328
```

```
mu.hat[3]-mu.hat[1]
```

```
## [1] -69.49227
```

```
mu.hat[3]-mu.hat[2]
```

```
## [1] -41.54899
```

Estimation in the treatment effects model

Consider the treatment effects model of a completely randomized design (CRD) with $t = 3$ levels and $r_i = 4$ replicates

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{\epsilon} \sim MVN(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$ and

$$\boldsymbol{y} = \begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{24} \\ y_{31} \\ y_{32} \\ y_{33} \\ y_{34} \end{pmatrix}, \ \boldsymbol{X} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{pmatrix}, \ \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{14} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \\ \epsilon_{24} \\ \epsilon_{31} \\ \epsilon_{32} \\ \epsilon_{33} \\ \epsilon_{34} \end{pmatrix}$$

9. Construct the design matrix $\boldsymbol{X}$

```
X<-cbind(1,as.matrix(Z))
colnames(X)<-c("reference","effect 1","effect 2","effect 3")
X
```

```
##    reference effect 1 effect 2 effect 3
## 1          1        1        0        0
## 2          1        1        0        0
## 3          1        1        0        0
## 4          1        1        0        0
## 5          1        0        1        0
## 6          1        0        1        0
## 7          1        0        1        0
## 8          1        0        1        0
## 9          1        0        0        1
## 10         1        0        0        1
## 11         1        0        0        1
## 12         1        0        0        1
```

10. Verify that $\boldsymbol{X}$ has rank $t = 3$. We say that this **model does not have full rank** since the rand of $\boldsymbol{X}$ is less its number of columns.

```
rankMatrix(X)[1]
```

```
## [1] 3
```

11. We can attempt to find the least squares estimate of $\boldsymbol{\beta}$ and we arrive again at the normal equations

$$\boldsymbol{X^T X\beta} = \boldsymbol{X^T y}$$

Try to find the inverse of $\boldsymbol{X^T X}$. Why this does not work?

```
solve(t(X)%*%X)
```

12. **Definition:** A generalized inverse of a square matrix $\boldsymbol{A}$ is another square matrix $\boldsymbol{A}^{-}$ satisfying

$$\boldsymbol{A}\boldsymbol{A}^{-}\boldsymbol{A} = \boldsymbol{A}$$

The command `ginv` in the package `MASS` computes one type of generalised inverse called the Moore-Penrose generalised inverse. Compute the Moore-Penrose generalised inverse of $\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}$ using `ginv`.

```
ginv1<-ginv(t(X)%*%X)
ginv1
```

```
##           [,1]      [,2]      [,3]      [,4]
## [1,] 0.046875  0.015625  0.015625  0.015625
## [2,] 0.015625  0.171875 -0.078125 -0.078125
## [3,] 0.015625 -0.078125  0.171875 -0.078125
## [4,] 0.015625 -0.078125 -0.078125  0.171875
```

13. Another type of generalized inverse can be calculated for the specific case of $\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}$ as follows

$$(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-} = \begin{pmatrix} 0 & 0 \\ 0 & (\boldsymbol{Z}^{\boldsymbol{T}}\boldsymbol{Z})^{-1} \end{pmatrix}$$

Compute this type of generalised inverse in our example

```
ginv2<-rbind(0,cbind(0,solve(t(Z)%*%Z)))
ginv2
```

```
##               factlevel 1 factlevel 2 factlevel 3
##             0        0.00        0.00        0.00
## factlevel 1 0        0.25        0.00        0.00
## factlevel 2 0        0.00        0.25        0.00
## factlevel 3 0        0.00        0.00        0.25
```

an alternative

```
Z2<-X[,c(1,3,4)]
ginv3<-matrix(0,4,4)
ginv3[c(1,3,4),c(1,3,4)]<-solve(t(Z2)%*%Z2)
```

14. The system of normal equations has many solutions! Verify (analytically, not in R) that if $(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-}$ is a generalized inverse of $\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}$ then

$$\tilde{\boldsymbol{\beta}} = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}$$

is a solution to the normal equations

$$\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}\,.$$

15. Find two solutions to the normal equations based on the two types of generalised inverses described above. Verify numerically that they actually solve the normal equations. Comment on the differences between the two solutions found

```r
beta1<-ginv1%*%t(X)%*%y
beta1
```

```
##           [,1]
## [1,] 380.13679
## [2,] 159.19078
## [3,] 131.24750
## [4,]  89.69851
```

```r
beta2<-ginv2%*%t(X)%*%y
beta2
```

```
##                 [,1]
##               0.0000
## factlevel 1 539.3276
## factlevel 2 511.3843
## factlevel 3 469.8353
```

```r
beta3<-ginv3%*%t(X)%*%y
beta3
```

```
##           [,1]
## [1,] 539.32756
## [2,]   0.00000
## [3,] -27.94328
## [4,] -69.49227
```

```r
# now we verify the solve the normal equations
t(X)%*%X%*%beta1-t(X)%*%y
```

```
##                   [,1]
## reference 3.637979e-12
## effect 1  4.547474e-13
## effect 2  9.094947e-13
## effect 3  6.821210e-13
```

```r
t(X)%*%X%*%beta2-t(X)%*%y
```

```
##                   [,1]
## reference 9.094947e-13
## effect 1  0.000000e+00
## effect 2  0.000000e+00
## effect 3  0.000000e+00
```

Note that `beta3` gives the same solution as the default choice in R via `lm` as shown below

```r
model.fit<-lm(response~treatment,crd)
coefficients(model.fit)
```

```
##      (Intercept) treatmentlevel 2 treatmentlevel 3
##        539.32756        -27.94328        -69.49227
```

16. As done in question 7, find estimates for

- $\mu_2 - \mu_1$
- $\mu_3 - \mu_1$
- $\mu_3 - \mu_2$

using the each of the two estimates found in the previous question. Compare your answers with those from question 8.

```
beta1[3]-beta1[2]
```

```
## [1] -27.94328
```

```
beta1[4]-beta1[2]
```

```
## [1] -69.49227
```

```
beta1[4]-beta1[3]
```

```
## [1] -41.54899
```

```
beta2[3]-beta2[2]
```

```
## [1] -27.94328
```

```
beta2[4]-beta2[2]
```

```
## [1] -69.49227
```

```
beta2[4]-beta2[3]
```

```
## [1] -41.54899
```

17. Now estimate $\mu_i = \mu + \tau_i$ for $i = 1, 2, 3$ using both estimates of $\boldsymbol{\beta}$ and compare your answers with those in question 5!

```
beta1[1]+beta1[2]
```

```
## [1] 539.3276
```

```
beta1[1]+beta1[3]
```

```
## [1] 511.3843
```

```r
beta1[1]+beta1[4]
```

```
## [1] 469.8353
```

```r
beta2[1]+beta2[2]
```

```
## [1] 539.3276
```

```r
beta2[1]+beta2[3]
```

```
## [1] 511.3843
```

```r
beta2[1]+beta2[4]
```

```
## [1] 469.8353
```

18. Compute the solutions given in the output of the `lm` command and compare with the answers in the previous two questions

```r
mod.crd.treat<-lm(response~treatment,data=crd)
summary(mod.crd.treat)
```

```
##
## Call:
## lm(formula = response ~ treatment, data = crd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.100  -3.841  -1.765   5.564  13.503
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        539.328      3.975 135.692 3.26e-16 ***
## treatmentlevel 2  -27.943      5.621  -4.971 0.000769 ***
## treatmentlevel 3  -69.492      5.621 -12.363 5.97e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.949 on 9 degrees of freedom
## Multiple R-squared:  0.9451, Adjusted R-squared:  0.9328
## F-statistic:  77.4 on 2 and 9 DF,  p-value: 2.137e-06
```