MAST90104: Introduction to Statistical Learning

Week 7 Lab and Workshop

1. In a manufacturing plant, filters are used to remove pollutants. We are interested in comparing the lifespan of 5 different types of filters. Six filters of each type are tested, and the time to failure in hours is given in the dataset (on the website) `filters` (in `csv` format).

   (a) Use the `read.csv` function to read the data. Then convert the `type` component into a factor.
   Solution:

```
filters <- read.csv("../data/filters.csv")
filters$type <- factor(filters$type)
```

   (b) Using only the `matrix` command, construct a $\mathbf{y}$ vector and two different X for this linear model both of which are full rank, one corresponding to `cont.treatment` and one to `contr.sum`.
   Solution:

```
y <- filters$life
X.treatment <- matrix(0,30,5)
X.treatment [,1] <- 1
for (i in 2:5) { X.treatment [filters$type==i,i] <- 1 }
X.treatment
```

```
##       [,1] [,2] [,3] [,4] [,5]
##  [1,]    1    0    0    0    0
##  [2,]    1    0    0    0    0
##  [3,]    1    0    0    0    0
##  [4,]    1    0    0    0    0
##  [5,]    1    0    0    0    0
##  [6,]    1    0    0    0    0
##  [7,]    1    1    0    0    0
##  [8,]    1    1    0    0    0
##  [9,]    1    1    0    0    0
## [10,]    1    1    0    0    0
## [11,]    1    1    0    0    0
## [12,]    1    1    0    0    0
## [13,]    1    0    1    0    0
## [14,]    1    0    1    0    0
## [15,]    1    0    1    0    0
## [16,]    1    0    1    0    0
## [17,]    1    0    1    0    0
## [18,]    1    0    1    0    0
## [19,]    1    0    0    1    0
## [20,]    1    0    0    1    0
## [21,]    1    0    0    1    0
## [22,]    1    0    0    1    0
## [23,]    1    0    0    1    0
## [24,]    1    0    0    1    0
## [25,]    1    0    0    0    1
## [26,]    1    0    0    0    1
## [27,]    1    0    0    0    1
## [28,]    1    0    0    0    1
## [29,]    1    0    0    0    1
## [30,]    1    0    0    0    1
```

```
X.sum <- matrix(0,30,5)
X.sum[,1] <- 1
for (i in 1:4) { X.sum[filters$type==i,i+1] <- 1 }
for (i in 2:5) {X.sum[filters$type==5,i] <- -1}
X.sum
```

```
##       [,1] [,2] [,3] [,4] [,5]
##  [1,]    1    1    0    0    0
##  [2,]    1    1    0    0    0
##  [3,]    1    1    0    0    0
##  [4,]    1    1    0    0    0
##  [5,]    1    1    0    0    0
##  [6,]    1    1    0    0    0
##  [7,]    1    0    1    0    0
##  [8,]    1    0    1    0    0
##  [9,]    1    0    1    0    0
## [10,]    1    0    1    0    0
## [11,]    1    0    1    0    0
## [12,]    1    0    1    0    0
## [13,]    1    0    0    1    0
## [14,]    1    0    0    1    0
## [15,]    1    0    0    1    0
## [16,]    1    0    0    1    0
## [17,]    1    0    0    1    0
## [18,]    1    0    0    1    0
## [19,]    1    0    0    0    1
## [20,]    1    0    0    0    1
## [21,]    1    0    0    0    1
## [22,]    1    0    0    0    1
## [23,]    1    0    0    0    1
## [24,]    1    0    0    0    1
## [25,]    1   -1   -1   -1   -1
## [26,]    1   -1   -1   -1   -1
## [27,]    1   -1   -1   -1   -1
## [28,]    1   -1   -1   -1   -1
## [29,]    1   -1   -1   -1   -1
## [30,]    1   -1   -1   -1   -1
```

(c) Fit the models and compare with the `lm` output.
Solution:

```
XtXinv <- solve(t(X.treatment )%*%X.treatment )
solve(t(X.treatment )%*%X.treatment )%*%t(X.treatment )%*%y
```

```
##           [,1]
## [1,] 249.16667
## [2,] -61.66667
## [3,] -83.16667
## [4,] 108.16667
## [5,] 112.16667
```

```
solve(t(X.sum)%*%X.sum)%*%t(X.sum)%*%y
```

```
##           [,1]
## [1,] 264.26667
## [2,] -15.10000
## [3,] -76.76667
## [4,] -98.26667
## [5,]  93.06667
```

```
model <- lm(y~type, data=filters)
(b <- model$coefficients)

## (Intercept)       type2       type3       type4       type5
##   249.16667   -61.66667   -83.16667   108.16667   112.16667
```

The default in `lm` is the `.treatment` and the estimates agree. The same is true with `contr.sum` in the second `lm` specification.

(d) Calculate s^2 using the residuals.

Solution:

```
(s2 <- sum(lm(y~type, data=filters)$residuals^2)/25)

## [1] 15304.2
```

(e) Calculate a 95% confidence interval for the difference in lifespan between filter types 3 and 4. This is estimable because it is a treatment contrast. In the `contr.treatment`, it is the difference between the 3rd and 4th parameters,

Solution:

```
tt <- c(0,0,1,-1,0)
hwtreat <- qt(0.975,df=25) * sqrt(s2 * t(tt) %*%solve(t(X.treatment )%*%X.treatment )  %*%
c(tt %*% b - hwtreat, tt %*% b + hwtreat)

## [1] -338.43399  -44.23268
```

(f) Show that the hypothesis that the filters all have the same lifespan is testable.

Solution: This is the hypothesis of no differences between means in the levels of a one factor model which has been shown to be testable in notes because the differences in means are estimable and the hypothesis can be expressed as requiring all the differences to be 0.

(g) Test this hypothesis, using matrix theory.

Solution:

```
C <- matrix(c(0,1,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0,0,1),4,5,byrow=TRUE)
(Fstat <- (t(C%*%b) %*% solve(C %*% XtXinv %*% t(C)) %*% C %*% b/4)/s2)

##          [,1]
## [1,] 3.318776

pf(Fstat, 4, 25, lower.tail=FALSE)

##             [,1]
## [1,] 0.02599945
```

(h) Test the same hypothesis using the `linearHypothesis` function from the `car` package.

Solution:

```
library(car)

## Loading required package:  carData

linearHypothesis(model, C, rep(0,4))

## Linear hypothesis test
##
## Hypothesis:
## type2 = 0
## type3 = 0
## type4 = 0
## type5 = 0
```

```
##
## Model 1: restricted model
## Model 2: y ~ type
##
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     29 585770
## 2     25 382605  4    203165 3.3188  0.026 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2. An industrial psychologist is investigating absenteeism among production-line workers, based on different types of work hours: (1) 4-day week with a 10-hour day, (2) 5-day week with a flexible 8-hour day, and (3) 5-day week with a structured 8-hour day. A study is conducted and the following data obtained of the average number of days missed:

	Work plan		
	1	2	3
Mean	9	6.2	10.1
Number	100	85	90

They also find $s^2 = 110.15$.

(a) Test the hypothesis that the work plan has no effect on the absenteeism.
Solution:

```
b <- c(9,6.2,10.1)
s2 <- 110.15
n <- c(100,85,90)
r <- 3
XtXinv <- diag(c(1/n))
(C <- matrix(c(1,-1,0,0,1,-1),2,3,byrow=T))
```

```
##      [,1] [,2] [,3]
## [1,]    1   -1    0
## [2,]    0    1   -1
```

```
(Fstat <- t(C%*%b)%*%solve(C%*%XtXinv%*%t(C))%*%C%*%b/2/s2)
```

```
##          [,1]
## [1,] 3.200371
```

```
pf(Fstat,2,sum(n)-r,lower=F)
```

```
##            [,1]
## [1,] 0.04228613
```

Therefore we reject the null hypothesis at a 5% level: work plan has an effect on absenteeism.

(b) Test the hypothesis that work plans 1 and 3 have the same rate of absenteeism.
Solution:

```
(C <- matrix(c(1,0,-1),1,3,byrow=T))
```

```
##      [,1] [,2] [,3]
## [1,]    1    0   -1
```

```
(Fstat <- t(C%*%b)%*%solve(C%*%XtXinv%*%t(C))%*%C%*%b/1/s2)
```

```
##           [,1]
## [1,] 0.5203431
```

```
pf(Fstat,1,sum(n)-r,lower=F)
```

```
##           [,1]
## [1,] 0.471315
```

We cannot reject the null hypothesis.

3. We study the effect of various breeds and diets on the milk yield of cows. A study is conducted on 9 cows and the following data obtained:

	Diet		
Breed	1	2	3
1	18.8	16.7	19.8
	21.2		23.9
2	22.3	15.9	21.8
		19.2	

(a) Express this as a two-factor model with no interaction in matrix form.
Solution: $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where

$$\mathbf{y} = \begin{bmatrix} 18.8 \\ 21.2 \\ 16.7 \\ 19.8 \\ 23.9 \\ 22.3 \\ 15.9 \\ 19.2 \\ 21.8 \end{bmatrix}, \quad X = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

and $\boldsymbol{\varepsilon}$ is as expected.

(b) Express this as a two-factor model with interaction in matrix form.
Solution: $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where

$$\mathbf{y} = \begin{bmatrix} 18.8 \\ 21.2 \\ 16.7 \\ 19.8 \\ 23.9 \\ 22.3 \\ 15.9 \\ 19.2 \\ 21.8 \end{bmatrix}, \quad X = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \xi_{11} \\ \xi_{12} \\ \xi_{13} \\ \xi_{21} \\ \xi_{22} \\ \xi_{23} \end{bmatrix}$$

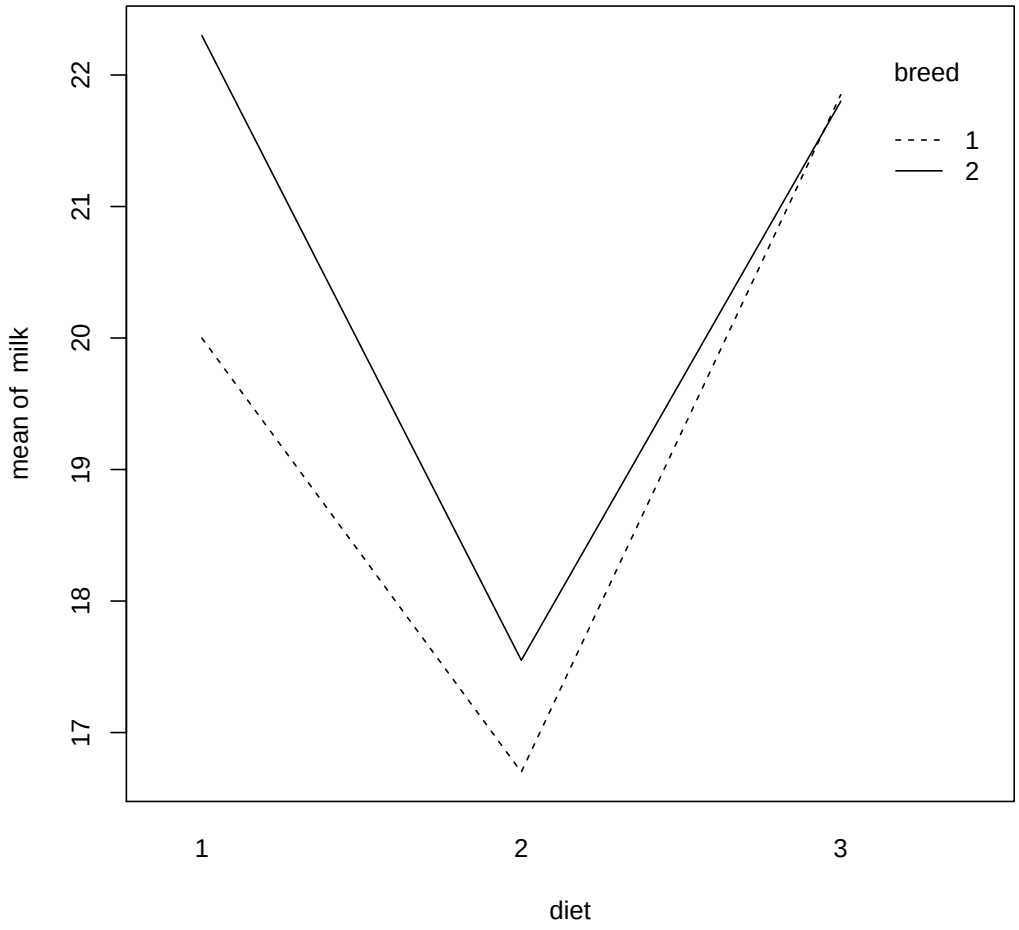
and $\boldsymbol{\varepsilon}$ is as expected.

(c) Express the hypothesis that there is no interaction in terms of your parameters. Eliminate any redundancies.
Solution: We know that we require $(a-1)(b-1) = 2$ hypotheses, so we take the obviously non-redundant hypotheses

$$\begin{aligned} (\xi_{11} - \xi_{12}) - (\xi_{21} - \xi_{22}) &= 0 \\ (\xi_{11} - \xi_{13}) - (\xi_{21} - \xi_{23}) &= 0. \end{aligned}$$

(d) Input this data into R. Plot an interaction plot between breed and diet.
Solution:

```r
milk <- data.frame(milk=c(18.8,21.2,16.7,19.8,23.9,22.3,15.9,19.2,21.8),
diet=factor(c(1,1,2,3,3,1,2,2,3)),
breed=factor(c(1,1,1,1,1,2,2,2,2)))
with(milk, interaction.plot(diet, breed, milk))
```



(e) Test for the presence of interaction.
Solution:

```r
imodel <- lm(milk ~ breed * diet, data=milk)
anova(imodel)

## Analysis of Variance Table
##
## Response: milk
##            Df Sum Sq Mean Sq F value Pr(>F)
## breed       1  0.174  0.1742  0.0312 0.8710
## diet        2 36.204 18.1018  3.2460 0.1777
## breed:diet  2  1.874  0.9372  0.1681 0.8527
## Residuals   3 16.730  5.5767
```

There is clearly no interaction.

(f) What is the degrees of freedom used for the interaction test?
Solution: We use 2 and 3 degrees of freedom.

(g) From the interaction model, what is the estimated amount of milk produced from breed 2 and diet 3?

Solution:

```
imodel$coeff
```

```
##  (Intercept)       breed2        diet2        diet3 breed2:diet2
##        20.00         2.30        -3.30         1.85        -1.45
## breed2:diet3
##        -2.35
```

```
c(1,1,0,1,0,1)%*%imodel$coeff
```

```
##      [,1]
## [1,] 21.8
```

(h) Fit an additive model. What is the estimated amount of milk produced from breed 2 and diet 3 now?

Solution:

```
amodel <- lm(milk ~ breed + diet, data=milk)
amodel$coeff
```

```
## (Intercept)      breed2       diet2       diet3
##   20.422222    1.033333   -3.844444    1.066667
```

```
c(1,1,0,1)%*%amodel$coeff
```

```
##          [,1]
## [1,] 22.52222
```

(i) Test the hypothesis (under the additive model) that the 2nd and 3rd diets are equivalent in terms of milk produced.

Solution:

```
library(car)
linearHypothesis(amodel, c(0,0,1,-1),0)
```

```
## Linear hypothesis test
##
## Hypothesis:
## diet2 - diet3 = 0
##
## Model 1: restricted model
## Model 2: milk ~ breed + diet
##
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1      6 52.000
## 2      5 18.604  1    33.396 8.9752 0.03024 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We reject this hypothesis at a 5% level.

(j) Find a 95% confidence interval, under the additive model, for the amount of milk produced from breed 2 and diet 3. Use both matrix calculations and the `estimable` function from the `gmodels` package.

Solution: Using the design matrix:

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

```
n <- 9
(X <- model.matrix(~breed+diet,data=milk))
```

```
##   (Intercept) breed2 diet2 diet3
## 1           1      0     0     0
## 2           1      0     0     0
## 3           1      0     1     0
## 4           1      0     0     1
## 5           1      0     0     1
## 6           1      1     0     0
## 7           1      1     1     0
## 8           1      1     1     0
## 9           1      1     0     1
## attr(,"assign")
## [1] 0 1 2 2
## attr(,"contrasts")
## attr(,"contrasts")$breed
## [1] "contr.treatment"
##
## attr(,"contrasts")$diet
## [1] "contr.treatment"
```

```
# X <- matrix(0,n,6)
# X[,1] <- 1
# X[cbind(1:n,as.numeric(milk£breed)+1)] <- 1
# X[cbind(1:n,as.numeric(milk£diet)+3)] <- 1
y <- milk$milk
XtXinv <- solve(t(X) %*% X)
#XtXc <- ginv(t(X) %*% X)
b <- XtXinv %*% t(X) %*% y
r <- rankMatrix(X)
s2 <- sum((y - X %*% b)^2)/(n - r)
t <- c(1,1,0,1)
mu23 <- t(t) %*% b
wdth <- qt(.975, n - r)*sqrt(s2 * t(t) %*% XtXinv %*% t)
c(mu23 - wdth, mu23, mu23 + wdth)
```

```
## [1] 18.82634 22.52222 26.21811
```

Alternatively we can use `estimable`. Note that like `linearHypothesis`, `estimable` requires that you express the estimated quantity in terms of the estimates R uses:

```
library(gmodels)
estimable(amodel, c(1,1,0,1), conf.int=0.95)
```

```
##           Estimate Std. Error  t value DF     Pr(>|t|) Lower.CI Upper.CI
## (1 1 0 1) 22.52222   1.437762 15.66477  5 1.927104e-05 18.82634 26.21811
```

(k) Find the same confidence interval under the interaction model.
Solution:

```
estimable(imodel, c(1,1,0,1,0,1), conf.int=0.95)
```

```
##               Estimate Std. Error  t value DF    Pr(>|t|) Lower.CI
## (1 1 0 1 0 1)     21.8   2.361497 9.231434  3 0.002689148 14.28466
##               Upper.CI
## (1 1 0 1 0 1) 29.31534
```

(l) Why is the second interval wider than the first?

Solution: The second interval is wider than the first because we are attributing some degrees of freedom to the interaction term(s). The resulting loss in degrees of freedom for the residuals leads to greater error in our estimations.

4. Suppose each row of a dataset has a response variable and two factors, which have 2 and 3 possible levels respectively. The dataset has 2 rows for each possible combination of factor levels. We model this with a less than full rank model with one parameter for the overall mean, and one parameter for each level of each factor, assuming that the overall mean is adjusted additively by each factor. Write down the linear model in both equation and matrix form.

Solution:

We denote the response variable from the kth sample from the combination of factors with the first factor at level i and the second factor at level j to be y_{ijk}. We also denote the overall mean by μ, and parameters corresponding to each factor by τ_i for the ith level of factor 1, and β_j for the jth level of factor 2. The linear model is

$$y_{ijk} = \mu + \tau_i + \beta_j + \varepsilon_{ijk},$$

for $i = 1, 2$, $j = 1, 2, 3$, and $k = 1, 2$.

Equivalently, $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where

$$\mathbf{y} = \begin{bmatrix} y_{111} \\ y_{112} \\ y_{121} \\ y_{122} \\ y_{131} \\ y_{132} \\ y_{211} \\ y_{212} \\ y_{221} \\ y_{222} \\ y_{231} \\ y_{232} \end{bmatrix}, \quad X = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} e_{111} \\ e_{112} \\ e_{121} \\ e_{122} \\ e_{131} \\ e_{132} \\ e_{211} \\ e_{212} \\ e_{221} \\ e_{222} \\ e_{231} \\ e_{232} \end{bmatrix}.$$

5. Let

$$A = \begin{bmatrix} 1 & 2 & 5 & 2 \\ 3 & 7 & 12 & 4 \\ 0 & 1 & -3 & -2 \end{bmatrix}.$$

(a) Show that $r(A) = 2$.

(b) Construct two different full rank matrices which generate the same column space as A.

Solution:

(a) It is easy to see that the third row of A is the second row minus 3 times the first row, but the first two rows are linearly independent. Therefore $r(A) = 2$.

(b) The first and second columns are linearly independent so they are a basis for the column space. Hence they could form a full rank matrix with the same column space. The same applies to the first and third columns (or the first and fourth).

6. It is known that toxic material was dumped into a river that flows into a large salt-water commercial fishing area. We are interested in the amount of toxic material (in parts per million) found in oysters harvested at three different locations in this area. A study is conducted and the following data obtained:

Site 1	Site 2	Site 3
15	19	22
26	15	26

(a) Write down the linear model in matrix form.
Solution: $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where

$$\mathbf{y} = \begin{bmatrix} 15 \\ 26 \\ 19 \\ 15 \\ 22 \\ 26 \end{bmatrix}, \quad X = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{bmatrix}.$$

(b) Write down the normal equations.
Solution: $X^T X\mathbf{b} = X^T\mathbf{y}$, where

$$X^T X = \begin{bmatrix} 6 & 2 & 2 & 2 \\ 2 & 2 & 0 & 0 \\ 2 & 0 & 2 & 0 \\ 2 & 0 & 0 & 2 \end{bmatrix}, \quad X^T\mathbf{y} = \begin{bmatrix} 123 \\ 41 \\ 34 \\ 48 \end{bmatrix}.$$

(c) Reparameterize the model to a full rank model.
Solution:

$$Y = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \quad \boldsymbol{\gamma} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}.$$

(d) Find a solution for the normal equations.
Solution: A solution for the normal equations for the original model is

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 0 \\ 20.5 \\ 22 \\ 24 \end{bmatrix}$$

and one for the reparameterised model is

$$\hat{\boldsymbol{\gamma}} = \begin{bmatrix} 20.5 \\ 22 \\ 24 \end{bmatrix}$$

.

7. In the one-way classification model, show that any linear combination of $\bar{y}_1 - \bar{y}_\cdot, \ldots, \bar{y}_k - \bar{y}_\cdot$ can be written as a linear combination of $\bar{y}_1, \ldots, \bar{y}_k$. Does the converse hold?

Solution: We have

$$\sum a_i(\bar{y}_i - \bar{y}_\cdot) = \sum a_i\bar{y}_i - \left(\sum a_i\right)\frac{1}{k}\sum \bar{y}_i = \sum (a_i - \bar{a})\bar{y}_i.$$

The converse only holds for contrasts.