MA50259: Statistical Design of Investigations

Lab sheet 4: ANOVA and Factorial designs

In the first part of this practical you will learn how to perform ANOVA in a completely randomised design with one factor.

ANOVA for the completely randomized design with one factor

Consider the treaments effects model in a balanced CRD design with t treatment levels and r replications per treatment level, that is:

$$y_{ij} = \mu + \tau_i + \epsilon_{ij} \qquad i = 1, 2, \ldots, t\,, \quad j = 1, \ldots, r$$

where $\epsilon_{ij} \sim N(0, \sigma^2)$ and are all independent. We consider testing the null hypothesis

$$H_0 : \tau_1 = \cdots = \tau_t = 0 \qquad \text{vs} \qquad H_1 : \tau_i \neq 0 \quad \text{for some } i$$

In the lectures we showed that

$$\frac{SSE}{\sigma^2} = \frac{\boldsymbol{y}^T(I - H_{\boldsymbol{X}})\boldsymbol{y}}{\sigma^2} \sim \chi^2_{(n-t,0)}$$

that is, central chi-squared with $n - t = t(r-1)$ degrees of freedom and

$$\frac{SST}{\sigma^2} = \frac{\boldsymbol{y}^T(H_{\boldsymbol{X}} - H_{\boldsymbol{1}})\boldsymbol{y}}{\sigma^2} \sim \chi^2_{(t-1,\lambda)}$$

where

$$\lambda = \frac{\boldsymbol{\mu}^T(I - H_{\boldsymbol{1}})\boldsymbol{\mu}}{\sigma^2} = \frac{r}{\sigma^2}\sum_{i=1}^{t}(\mu_i - \bar{\mu})^2$$

that is, non-central chi-squared with $t-1$ degrees of freedom and noncentrality parameter λ. Recall that $\mu_i = \mu + \tau_i$ and $\boldsymbol{\mu}^T = (\mu_1, \ldots, \mu_1, \mu_2, \ldots, \mu_2, \ldots, \mu_t, \ldots, \mu_t)$ the mean vector of length $n = tr$. Also recall that $\lambda = 0$ if and only if the null hypothesis $\tau_1 = \cdots = \tau_t$ is true.

The test uses following test statistic

$$F = \frac{SST/(t-1)}{SSE/(n-t)} \sim F_{(t-1,n-t,\lambda)}$$

a non-central F distribution with $t-1$ degrees of freedom in the numerator, $n-t$ degrees of freedom in the denominator and noncentrality parameter

$$\lambda = \frac{r}{\sigma^2}\sum_{i=1}^{t}(\mu_i - \bar{\mu})^2$$

the same as above. The test rejects H_0 when $F > F_c$ where F_c is the critical value. The critical value F_c is the unique value that satisfies the equation

$$P[F > F_c|H_0] = \alpha$$

where α is the pre-fixed significance level. Note how σ^2 cancels out in F giving a test statistic that can be computed (under the null hypothesis) as opposed to simply using SST/σ^2 that cannot be computed since σ^2 is unknown

In R, the probability density, cumulative distribution, quantile and random number generating function of the noncentral chi-square and the non-central F distributions are computed using `dchisq, pchisq, qchisq, rchisq` and `df, pf, qf, rf` respectively. See the corresponding help files for details.

1. Consider the case where $t = 3$ and $r = 4$. Compute the critical value F_c of the test when $\alpha = 0.05$.

```
t<-3
r<-4
alpha<-0.05
df1<-t-1
df2<-t*(r-1)
# we use the quantile function
Fc<-qf(1-alpha,df1,df2) # note that by default, the non-centrality parameter is zero
Fc
```

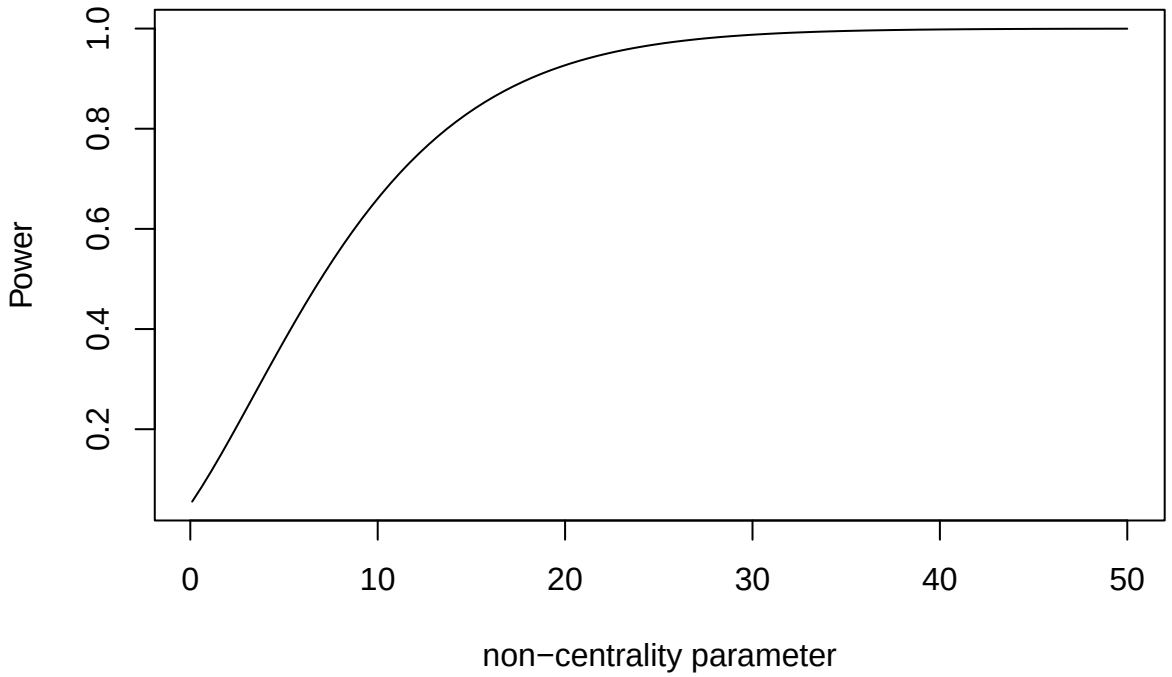
```
## [1] 4.256495
```

2. Using the specifications in question 1, plot the power of the test as a function of the non-centrality parameter in the range $(0, 50)$. The power of the test is given by

   $$P[F > F_c|H_a]$$

   that is, the probability of rejecting H_0 when the alternative hypothesis is true and therefore is a function of the non-centrality parameter. What can you conclude form this plot in terms of the behaviour of the power as function of the number of replicates r and/or as a function of the quantity

```
ncp<-seq(0.1,50,length.out = 100)
pow<-rep(0,100)
for (i in 1:100){
  pow[i]<-pf(Fc,df1,df2,ncp[i],lower.tail = FALSE)
}
plot(ncp,pow,type="l",xlab="non-centrality parameter",ylab="Power")
```



Clearly, as the non-centrality parameter λ increases, the power also increases. Since the non-centrality parameter is of the form

$$\lambda = \frac{r}{\sigma^2}\sum_{i=1}^{t}(\mu_i - \bar{\mu})^2$$

the above plot implies that the power increases if the number of replicates per treatment level (r) increases. Similarly, we obtain a larger power for a smaller variance than with a larger variance. Finally, since $\sum_{i=1}^{t}(\mu_i - \bar{\mu})^2$ is a measure of dispersion on the treatment level means $\{\mu_i\}$ then we can say that the more dispersed the treament level means are, then the power is larger. This last statement makes sense because if the treament level means are very much dispersed, it means they are far from being equal and therefore the test will be able to pick this up with high probability and reject the null hypothesis (high power).

Now you will start learning the basics of how to analyse an experimental factorial design with 2 factors. Take the time to run each of the following commands and analyse the displayed results to understand what the code is doing.

CO Data

Consider a two-factor experiment that consisted of burning an amount of fuel and determining the CO emissions released. The experimental unit is the portion of a standard fuel required for one run, and the response is the carbon monoxide (CO) emissions concentration in grams/cubic meters determined from that run. There were two replicate runs for each combination of factor levels. Factor A is the amount of ethanol added to an experimental unit and factor B is the fuel-to-air ratio used during the burn of that fuel. The data can be downloaded as follows:

```
library(tidyverse)
url_data<-"http://people.bath.ac.uk/kai21/MA50259/Data/COdata.txt"
COdata<-url_data %>% read_delim(delim=" ",col_types=list(col_character(),col_character(),col_integer()))
```

Estimation in the treatment effects model

Consider a factorial design with two factors. Both factors have three levels and there are two replicates per cell, giving a total of 18 observations. The coresponding effects model is given by:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$$

where

- α_i are the treatment effects for the first factor
- β_j are the treatment effects for the second factor
- γ_{ij} are the interaction effects

The model can be writen in matrix form as follows:

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

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

$$\begin{pmatrix} y_{111} \\ y_{121} \\ y_{131} \\ y_{211} \\ y_{221} \\ y_{231} \\ y_{311} \\ y_{321} \\ y_{331} \\ y_{112} \\ y_{122} \\ y_{132} \\ y_{212} \\ y_{222} \\ y_{232} \\ y_{312} \\ y_{322} \\ y_{332} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \gamma_{11} \\ \gamma_{21} \\ \gamma_{31} \\ \gamma_{12} \\ \gamma_{22} \\ \gamma_{33} \\ \gamma_{13} \\ \gamma_{23} \\ \gamma_{33} \end{pmatrix} + \begin{pmatrix} \epsilon_{111} \\ \epsilon_{121} \\ \epsilon_{131} \\ \epsilon_{211} \\ \epsilon_{221} \\ \epsilon_{231} \\ \epsilon_{311} \\ \epsilon_{321} \\ \epsilon_{331} \\ \epsilon_{112} \\ \epsilon_{122} \\ \epsilon_{132} \\ \epsilon_{212} \\ \epsilon_{222} \\ \epsilon_{232} \\ \epsilon_{312} \\ \epsilon_{322} \\ \epsilon_{332} \end{pmatrix}$$

1. Load the corresponding packages to use in this tutorial

```
library(MASS)
```

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

```
# Create factors
COdata<-COdata %>% mutate(Eth=as.factor(Eth))
COdata<-COdata %>% mutate(Ratio=as.factor(Ratio))
# Create interactions
inter<-interaction(COdata$Eth,COdata$Ratio,sep=":")

X1 <- model.matrix(~ Eth-1,data=COdata)
X2 <- model.matrix(~ Ratio-1,data=COdata)
```

```r
X3 <- model.matrix(~ inter-1)

X<-cbind(1,X1,X2,X3)
colnames(X)[1]<-"(Intercept)"
X
```

```
##    (Intercept) Eth0.1 Eth0.2 Eth0.3 Ratio14 Ratio15 Ratio16 inter0.1:14
## 1            1      1      0      0       1       0       0           1
## 2            1      1      0      0       0       1       0           0
## 3            1      1      0      0       0       0       1           0
## 4            1      0      1      0       1       0       0           0
## 5            1      0      1      0       0       1       0           0
## 6            1      0      1      0       0       0       1           0
## 7            1      0      0      1       1       0       0           0
## 8            1      0      0      1       0       1       0           0
## 9            1      0      0      1       0       0       1           0
## 10           1      1      0      0       1       0       0           1
## 11           1      1      0      0       0       1       0           0
## 12           1      1      0      0       0       0       1           0
## 13           1      0      1      0       1       0       0           0
## 14           1      0      1      0       0       1       0           0
## 15           1      0      1      0       0       0       1           0
## 16           1      0      0      1       1       0       0           0
## 17           1      0      0      1       0       1       0           0
## 18           1      0      0      1       0       0       1           0
##    inter0.2:14 inter0.3:14 inter0.1:15 inter0.2:15 inter0.3:15 inter0.1:16
## 1            0           0           0           0           0           0
## 2            0           0           1           0           0           0
## 3            0           0           0           0           0           1
## 4            1           0           0           0           0           0
## 5            0           0           0           1           0           0
## 6            0           0           0           0           0           0
## 7            0           1           0           0           0           0
## 8            0           0           0           0           1           0
## 9            0           0           0           0           0           0
## 10           0           0           0           0           0           0
## 11           0           0           1           0           0           0
## 12           0           0           0           0           0           1
## 13           1           0           0           0           0           0
## 14           0           0           0           1           0           0
## 15           0           0           0           0           0           0
## 16           0           1           0           0           0           0
## 17           0           0           0           0           1           0
## 18           0           0           0           0           0           0
##    inter0.2:16 inter0.3:16
## 1            0           0
## 2            0           0
## 3            0           0
## 4            0           0
## 5            0           0
## 6            1           0
## 7            0           0
## 8            0           0
```

```
## 9            0           1
## 10           0           0
## 11           0           0
## 12           0           0
## 13           0           0
## 14           0           0
## 15           1           0
## 16           0           0
## 17           0           0
## 18           0           1
```

```
Z<-model.matrix(~ Eth*Ratio,data=COdata)
y<-COdata$CO
y
```

```
##  [1] 66 72 68 78 80 66 90 75 60 62 67 66 81 81 69 94 78 58
```

3. Find the rank of $\boldsymbol{X}$

```
qr(X)$rank
```

```
## [1] 9
```

4. Compute the Moore-Penrose generalised inverse of $\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}$ using `ginv`.

```
A<-unname(t(X)%*%X) # remove row/column names
G<-ginv(t(X)%*%X)
```

5. Find a solution to the normal equations using the generalised inverse found above. Verify numerically that it actually solves the normal equations.

```
beta.est<-G%*%t(X)%*%y
beta.est
```

```
##           [,1]
##  [1,] 40.96875
##  [2,]  9.15625
##  [3,] 15.90625
##  [4,] 15.90625
##  [5,] 17.90625
##  [6,] 15.65625
##  [7,]  7.40625
##  [8,] -4.03125
##  [9,]  4.71875
## [10,] 17.21875
## [11,]  3.71875
## [12,]  7.96875
## [13,]  3.96875
## [14,]  9.46875
## [15,]  3.21875
## [16,] -5.28125
```

```r
# now we verify the solve the normal equations
unname(t(X)%*%X%*%beta.est-t(X)%*%y)
```

```
##                [,1]
##  [1,] 9.094947e-13
##  [2,] 2.842171e-13
##  [3,] 3.410605e-13
##  [4,] 2.273737e-13
##  [5,] 3.410605e-13
##  [6,] 3.410605e-13
##  [7,] 2.273737e-13
##  [8,] 1.136868e-13
##  [9,] 1.136868e-13
## [10,] 8.526513e-14
## [11,] 1.136868e-13
## [12,] 1.421085e-13
## [13,] 1.136868e-13
## [14,] 8.526513e-14
## [15,] 8.526513e-14
## [16,] 5.684342e-14
```

6. In the lectures we showed that the means of the model are estimable. Find unbiased estimates for each of the nine means in the above model.

```r
# matrix with rows equal to estimable coefficients
Lambda<-X[1:9,]
# estimates are simple lambda'*beta.est
Lambda%*%beta.est
```

```
##    [,1]
## 1 64.0
## 2 69.5
## 3 67.0
## 4 79.5
## 5 80.5
## 6 67.5
## 7 92.0
## 8 76.5
## 9 59.0
```

```r
# we can confirm this calculkation by calculating the actual means by hand
means<-rep(0,9)
for (i in 1:9){
  means[i]<-mean(y[c(i,9+i)])
}
means
```

```
## [1] 64.0 69.5 67.0 79.5 80.5 67.5 92.0 76.5 59.0
```

7. Verify that the 9 estimable linear combinations above are lineraly independent

```
# rank should be equal to rank of X
qr(Lambda)$rank
```

```
## [1] 9
```

8. Compute the solutions given in the output of the `lm` command. This reparametrised full rank model corresponds to removing which columns from $\boldsymbol{X}$? Using this reparametrised full rank model, estimate the means of each of the 9 combinations. Compare with the answers above.

```
mod.factorial<-lm(CO~Eth*Ratio,data=COdata)
summary(mod.factorial)
```

```
##
## Call:
## lm(formula = CO ~ Eth * Ratio, data = COdata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##   -2.5   -1.5    0.0    1.5    2.5
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      64.000      1.607  39.819 1.98e-11 ***
## Eth0.2           15.500      2.273   6.819 7.74e-05 ***
## Eth0.3           28.000      2.273  12.318 6.16e-07 ***
## Ratio15           5.500      2.273   2.420 0.038631 *
## Ratio16           3.000      2.273   1.320 0.219477
## Eth0.2:Ratio15   -4.500      3.215  -1.400 0.195062
## Eth0.3:Ratio15  -21.000      3.215  -6.533 0.000107 ***
## Eth0.2:Ratio16  -15.000      3.215  -4.666 0.001175 **
## Eth0.3:Ratio16  -36.000      3.215 -11.199 1.38e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.273 on 9 degrees of freedom
## Multiple R-squared:  0.9727, Adjusted R-squared:  0.9483
## F-statistic: 40.02 on 8 and 9 DF,  p-value: 3.861e-06
```

```
# 7 columns removed: 2,5,8,9,10,11,14
```

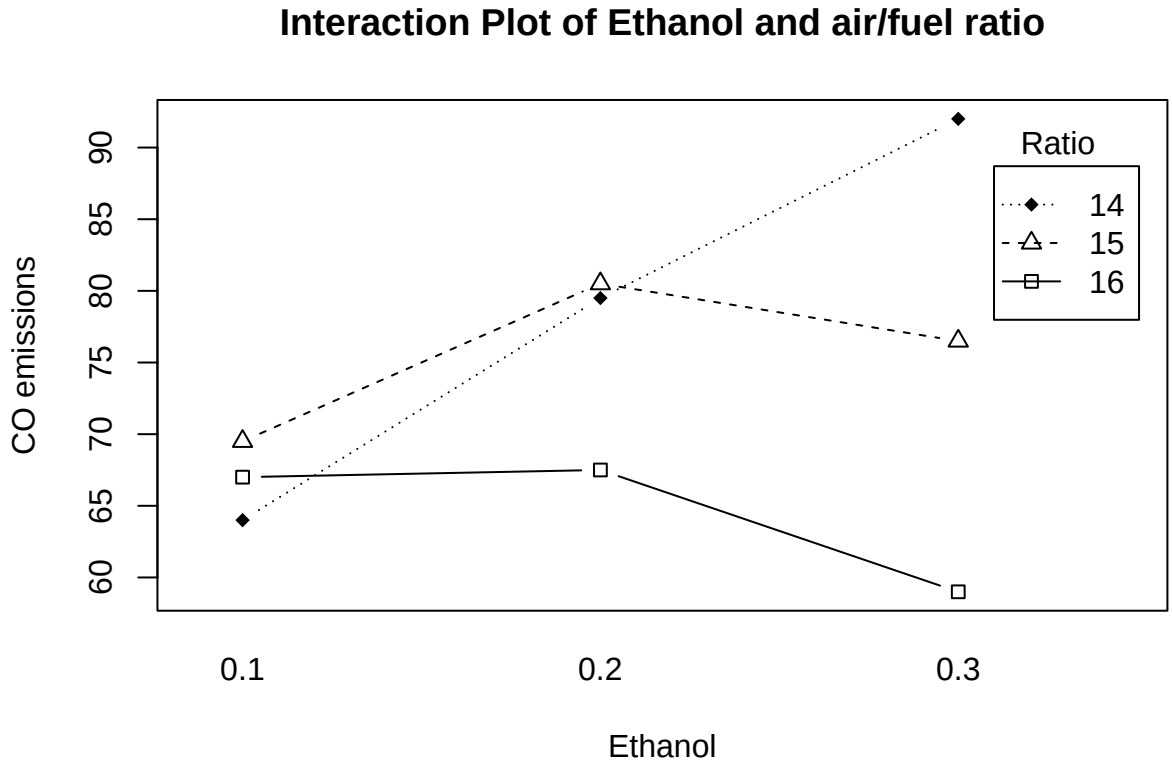
```
Z<-model.matrix(CO~Eth*Ratio,data=COdata)
AA<-t(Z)%*%Z
# new design matrix should be invertible
beta.est.2<-solve(AA)%*%t(Z)%*%y
means.est<-Z%*%beta.est.2
means.est
```

```
##    [,1]
## 1  64.0
## 2  69.5
## 3  67.0
```

```
## 4  79.5
## 5  80.5
## 6  67.5
## 7  92.0
## 8  76.5
## 9  59.0
## 10 64.0
## 11 69.5
## 12 67.0
## 13 79.5
## 14 80.5
## 15 67.5
## 16 92.0
## 17 76.5
## 18 59.0
```

9. The command below, produces an interaction plot for this dataset. What can you conclude from this plot?

```
with(COdata, (interaction.plot(Eth, Ratio, CO, type = "b",
 pch = c(18,24,22), leg.bty = "o",
 main = "Interaction Plot of Ethanol and air/fuel ratio",
 xlab = "Ethanol",ylab = "CO emissions")))
```




```
## NULL
```

10. The error sum of squares is defined as

$$ssE := \|\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}\|^2$$

where $\tilde{\boldsymbol{\beta}}$ is a solution to the normal equations. This can be written as

$$ssE := \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{G}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}\|^2$$

and therefore is invariant to the choice of generalised inverse $\boldsymbol{G}$ since $\boldsymbol{X}\boldsymbol{G}\boldsymbol{X}^{\boldsymbol{T}}$ is also invariant. Compute the sum of squares for the CO data model.

```
d<-y-X%*%G%*%t(X)%*%y
SSE<-t(d)%*%d
SSE
```

```
##          [,1]
## [1,] 46.5
```

11. To test the hypotheses

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$$

or

$$H_0' : \beta_1 = \beta_2 = \beta_3 = 0$$

or

$$H_0'' : \gamma_{11} = \gamma_{21} = \cdots = \gamma_{23} = \gamma_{33} = 0$$

compute the error sum of squares in the reduced models where each null hypothesis is true. Call these sums of squares SS_1, SS_2 and SS_3 respectively. Compute $SS_1 - SS_3$, $SS_2 - SS_3$ and $SS_3 - SSE$ in the CO data and compare to the output given by the R command `aov( CO  Eth * Ratio, data = COdata)`.

```
d1<-y-X1%*%solve(t(X1)%*%X1)%*%t(X1)%*%y
d2<-y-X2%*%solve(t(X2)%*%X2)%*%t(X2)%*%y
XX3<-model.matrix(~Eth+Ratio,data=COdata)
d3<-y-XX3%*%solve(t(XX3)%*%XX3)%*%t(XX3)%*%y

SS1<-t(d1)%*%d1
SS2<-t(d2)%*%d2
SS3<-t(d3)%*%d3

SS2-SS3
```

```
##          [,1]
## [1,]  324
```

```
SS1-SS3
```

```
##          [,1]
## [1,]  652
```

```
SS3-SSE
```

```
##          [,1]
## [1,]  678
```

```
SSE
```

```
##      [,1]
## [1,] 46.5
```

```
mod1 <- aov( CO ~ Eth * Ratio, data = COdata )
summary(mod1)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Eth          2  324.0   162.0   31.36 8.79e-05 ***
## Ratio        2  652.0   326.0   63.10 5.07e-06 ***
## Eth:Ratio    4  678.0   169.5   32.81 2.24e-05 ***
## Residuals    9   46.5     5.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
X11<-cbind(1,X1)
X22<-cbind(1,X2)
X33<-cbind(1,X1,X2)

d1<-y-X11%*%ginv(t(X11)%*%X11)%*%t(X11)%*%y
d2<-y-X22%*%ginv(t(X22)%*%X22)%*%t(X22)%*%y
d3<-y-X33%*%ginv(t(X33)%*%X33)%*%t(X33)%*%y

SS1<-t(d1)%*%d1
SS2<-t(d2)%*%d2
SS3<-t(d3)%*%d3

SS2-SS3
```

```
##      [,1]
## [1,]  324
```

```
SS1-SS3
```

```
##      [,1]
## [1,]  652
```

```
SS3-SSE
```

```
##      [,1]
## [1,]  678
```

```
SSE
```

```
##      [,1]
## [1,] 46.5
```

ANOVA for the factorial design with two factors

Consider the treaments effects model in a balanced factorial design with two factors and r replications, that is:

$$y_{ijk} = \mu + \tau_i + \alpha_j + \gamma_{ij} + \epsilon_{ijk}$$

where

- τ_i are the treatment effects for the first factor $i = 1, \ldots, t$
- α_j are the treatment effects for the second factor $j = 1, \ldots, s$
- γ_{ij} are the interactions
- $k = 1, 2, \ldots, r$

where $\epsilon_{ij} \sim N(0, \sigma^2)$ and are all independent. We consider testing the null hypothesis

$$H_0 : \gamma_{11} = \gamma_{12} = \cdots = \gamma_{ts} = 0 \qquad \text{vs} \qquad H_1 : \gamma_{ij} \neq 0 \quad \text{for some } i \text{ and } j$$

1. Consider the CO emissions data from Lab 4

```
library(MASS)
COdata<-read.table("http://people.bath.ac.uk/kai21/MA50259/Data/COdata.txt",header = TRUE)
COdata$Eth<-as.factor(COdata$Eth)
COdata$Ratio<-as.factor(COdata$Ratio)
```

Compute the sum of squares of the error in general (and call it `SSE`), the sum of squares under the null hypothesis H_0 (and call it `SSE0`) and the sum of squares due to the interactions which is `SSE0-SSE` and call it `SSInt`

```
inter<-interaction(COdata$Eth,COdata$Ratio,sep=":")
X1 <- model.matrix(~ Eth-1,data=COdata)
X2 <- model.matrix(~ Ratio-1,data=COdata)
X3 <- model.matrix(~ inter-1)

X<-cbind(1,X1,X2,X3)
X0<-cbind(1,X1,X2)
G<-ginv(t(X)%*%X)
G0<-ginv(t(X0)%*%X0)

HX<-X%*%G%*%t(X)
HX0<-X0%*%G0%*%t(X0)

y<-COdata$CO
n<-length(y)

r0<-(diag(n)-HX0)%*%y
r<-(diag(n)-HX)%*%y

SSE0<-t(r0)%*%r0
SSE<-t(r)%*%r

SSE0
```

```
##       [,1]
## [1,] 724.5
```

```
SSE
```

```
##      [,1]
## [1,] 46.5
```

```
SSInt<-SSE0-SSE
SSInt
```

```
##      [,1]
## [1,]  678
```

2. Compute the degrees of freedom associated with `SSInt` and `SSE` and then compute the F test statistic as follows

$$F = \frac{SSInt/df_1}{SSE/df_2}$$

where df_1 and df_2 are the degrees of freedom associated with `SSint` and `SSE` respectively

```
df2<-qr(diag(n)-HX)$rank # df of SSE
df1<-qr(HX-HX0)$rank # df of SSInt
df1
```

```
## [1] 4
```

```
df2
```

```
## [1] 9
```

```
Fval<-(SSInt/df1)/(SSE/df2)
Fval
```

```
##          [,1]
## [1,] 32.80645
```

3. Compute the p-value associated using the fact that, under the null hypothesis

$$F = \frac{SSInt/df_1}{SSE/df_2} \sim F_{(df1,df2,0)}$$

a central F distribution with df_1 degrees of freedom in the numerator , df_2 degrees of freedom in the denominator.

```
pf(Fval,df1,df2,lower.tail = FALSE)
```

```
##              [,1]
## [1,] 2.240276e-05
```

4. Compare the results in the previous question with the ones given by running the following commands

```r
mod<-lm(CO~Eth*Ratio,COdata)
summary(aov(mod))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Eth          2  324.0   162.0   31.36 8.79e-05 ***
## Ratio        2  652.0   326.0   63.10 5.07e-06 ***
## Eth:Ratio    4  678.0   169.5   32.81 2.24e-05 ***
## Residuals    9   46.5     5.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```