Solucions

Problema 1

(a) Tenim 4 paràmetres i el rang de la matriu de disseny és 3, llavors per trobar les funcions paramètriques estimables cal resoldre el següent sistema d'equacions:

$$(a_1,a_2,a_3,a_4) = \lambda_1(1,-1,0,0)+\lambda_2(1,1,1,1)+\lambda_3(1,1,-1,-1)$$

De manera que una combinació lineal $\psi = a_1\alpha + a_2\beta + a_3\gamma + a_4\delta$ serà f.p.e. si $a_3 = a_4$.

(b) Tant $\alpha+\beta$ com α són f.p.e. ja que verifiquen la condició anterior.

Per tal de calcular la seva estimació MQ, procedim a escriure la matriu de disseny i el vector de respostes (3 rèpliques per a cada situació experimental).

```
> x <- c(1,-1,0,0,
+        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,-1,-1,
+        1,1,-1,-1)
> x <- matrix(x, ncol=4, byrow=T)
> y <- c(1.1,0.8,0.9,10.5,9.7,10.1,4.3,3.9,4.2)
```

i ara calculem una estimació dels paràmetres amb l'ajuda de la g-inversa

```
> library(MASS)
> xtx <- t(x)%*%x
> xtxi <- ginv(xtx)
> betas <- xtxi %*% t(x) %*% y
```

Així doncs, les estimacions de las f.p.e són

```
> c(betas[1]+betas[2], betas[1])

[1] 7.116667 4.025000
```

(c) La matriu de variàncies-covariàncies de les estimacions dels paràmetres és $\sigma^2(\mathbf{X}'\mathbf{X})^-$, on σ^2 la podem estimar amb el MSE.

```
> e <- y - x %*% betas
> mse <- sum(e^2)/(9-3)
```

llavors la matriu $\text{var}(\hat{\beta})$ es

```
> (mm <- mse * xtxi)

             [,1]         [,2]        [,3]        [,4]
[1,]  0.009444444 -0.003148148 0.000000000 0.000000000
[2,] -0.003148148  0.009444444 0.000000000 0.000000000
[3,]  0.000000000  0.000000000 0.003148148 0.003148148
[4,]  0.000000000  0.000000000 0.003148148 0.003148148
```

i l'estimació de la variància $\text{var}(\hat{\alpha}) = \text{var}(\hat{\beta}) = 0.00944$ i la covariància $\text{cov}(\hat{\alpha},\hat{\beta}) = -0.00315$.

(d) La primera hipòtesi conté una única equació, de manera que es pot resoldre amb un test t de Student.

```
> t.est <- (betas[1] - 4)/sqrt(mm[1,1])
> (p.value <- pt(t.est, df=6, lower.tail=F) * 2)
```

```
[1] 0.8055822
```

de forma que acceptem la hipòtesi $H_0^{(1)} : \alpha = 4$.

L'altra hipòtesi $H_0^{(2)} : \alpha = 4, \beta = 3$ requereix un test F.

```
> A <- c(1,0,0,0,
+        0,1,0,0)
> A <- matrix(A, ncol=4, byrow=T)
> Abetas_c <- A %*% betas - c(4,3)
> atai <- solve(A %*% xtxi %*% t(A))
> numerador <- t(Abetas_c) %*% atai %*% Abetas_c / 2
> F.est <- numerador / mse
> (p.value <- pf(F.est, 2, 6, lower.tail=F))
```

```
          [,1]
[1,] 0.5650918
```

Està clar que acceptem la hipòtesi nul·la.

Una altra forma de resoldre el mateix test és considerar dos models lineals i fer un ANOVA entre ells.

```
> alpha <- c(rep(1,3),rep(1,3),rep(1,3))
> beta <- c(rep(-1,3),rep(1,3),rep(1,3))
> gamma <- c(rep(0,3),rep(1,3),rep(-1,3))
> delta <- c(rep(0,3),rep(1,3),rep(-1,3))
> m1 <- lm(y ~ 0 + alpha + beta + gamma + delta)
> alphabeta <- c(rep(1,3),rep(7,3),rep(7,3))
> m0 <- lm(y ~ 0 + offset(alphabeta) + gamma + delta)
> anova(m0,m1)
```

```
Analysis of Variance Table

Model 1: y ~ 0 + offset(alphabeta) + gamma + delta
Model 2: y ~ 0 + alpha + beta + gamma + delta
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1      8 0.54833
2      6 0.45333  2     0.095 0.6287 0.5651
```

El resultat és idèntic.

Problema 2

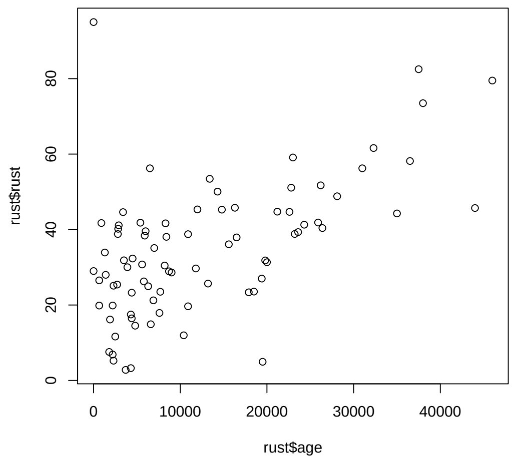
(a) En primer lloc carreguem les dades i els hi fem un cop d'ull.

```
> rust <- read.table("rust.txt", header=T, sep=",")
> rownames(rust) <- rust$name
> rust <- rust[ ,-1]
> rust$type <- factor(rust$type, labels=c("ordinary","non ordinary"))
> str(rust)
```

```
'data.frame':        82 obs. of  3 variables:
 $ type: Factor w/ 2 levels "ordinary","non ordinary": 1 1 1 1 1 1 1 2 2 1 ...
 $ age : int  1 0 650 650 900 1300 1400 1800 1900 2200 ...
 $ rust: num  29 95 19.9 26.5 41.7 ...
```

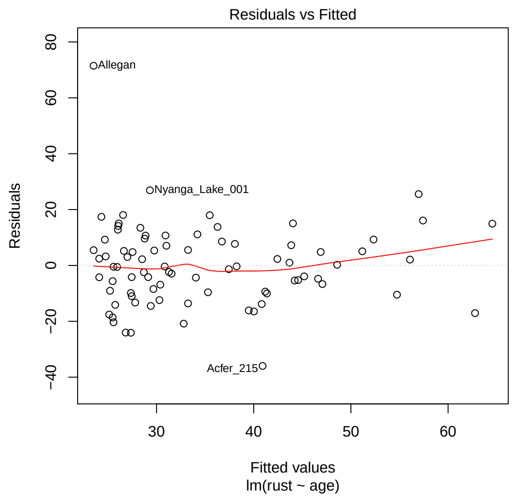
Ara fem el gràfic de dispersió de les variables `age` i `rust`

```
> plot(rust$age,rust$rust)
```



En el gràfic s'observen dos punts estranys que podem identificar amb l'anàlisi dels residus de la regressió lineal.

```
> g <- lm(rust ~ age, data=rust)
> plot(g, which=1)
```



Els punts es corresponen amb les condrites `Allegan` i `Acfer_215`. Procedirem a eliminar aquestes dues de la base de dades.

```
> which(rownames(rust) == "Allegan")

[1] 2

> which(rownames(rust) == "Acfer_215")

[1] 61

> rust0 <- rust[-c(2,61), ]
```

(b) Ara estimem els paràmetres de la regressió.

```
> g <- lm(rust ~ age, data=rust0)
> sg <- summary(g)
> coef(g)
```

```
 (Intercept)         age
21.650786309  0.001004476
```

```
> sg$sigma^2
```

```
[1] 135.9715
```

```
> sg$r.squared
```

```
[1] 0.4933011
```

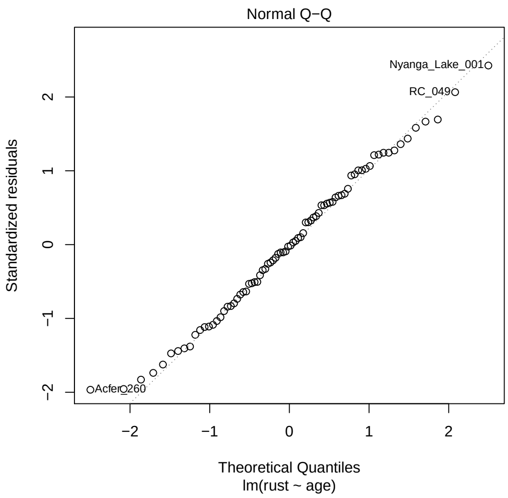
```
> pf(sg$fstatistic[1], 1, 80-2, lower.tail=F)
```

```
       value
3.870635e-13
```

De forma que l'estadístic F és clarament significatiu, el que fa significativa la regressió, encara que l'ajust que mostra el coeficient de determinació no és massa bo.

(c) Mirem la normalitat dels residus amb un gràfic.

```
> plot(g, which=2)
```



A simple vista tot sembla "normal". Fem un test.

```
> shapiro.test(residuals(g))

        Shapiro-Wilk normality test

data:  residuals(g)
W = 0.9884, p-value = 0.6895
```

No hi ha cap raó per dubtar de la normalitat[1].

(d) Els intervals de confiança dels paràmetres es troben amb

```
> confint(g, level=0.99)

                   0.5 %       99.5 %
(Intercept) 1.647573e+01 26.825844643
age         7.001273e-04  0.001308824
```

(e) Per a contrastar aquesta hipòtesi fem un model que la representi.

[1]També podem fer servir qualsevol dels altres test que podeu consultar en el Blog de los erreros.

```
> g0 <- lm(rust ~ 0 + offset(rep(23,80)) + offset(0.0009*age), data=rust0)
> anova(g0,g)
```

```
Analysis of Variance Table

Model 1: rust ~ 0 + offset(rep(23, 80)) + offset(9e-04 * age)
Model 2: rust ~ age
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     80 10718
2     78 10606  2    111.74 0.4109 0.6645
```

Acceptem la hipòtesi nul·la.

(f) Fem la predicció per a un condrita particular

```
> predict(g, newdata=data.frame(age<-rust["Acfer_215","age"]), interval="prediction")
```

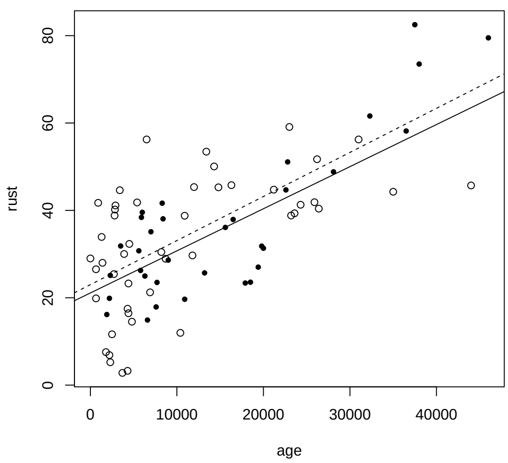
```
       fit      lwr      upr
1 41.23806 17.82667 64.64945
```

(g) Les dues rectes de regressió per separat són

```
> r1 <- lm(rust ~ age, data=rust0, subset=type=="ordinary")
> r2 <- lm(rust ~ age, data=rust0, subset=type=="non ordinary")
```

Amb aquestes ja podem fer un gràfic.

```
> idx <- rust$type == "ordinary"
> plot(rust0$age,rust0$rust, pch=ifelse(idx,1,20), xlab="age", ylab="rust")
> abline(r1)
> abline(r2, lty=2)
```



Però si volem saber si les rectes són paral·leles o coincidents hem de considerar un model conjunt amb quatre paràmetres: α_1, β_1 per a les condrites ordinàries i α_2, β_2 per a les no ordinàries.

```
> n1 <- sum(rust0$type == "ordinary")
> n2 <- sum(rust0$type == "non ordinary")
> n <- n1+n2
> Xc <- matrix(numeric(n*4), ncol=4)
> colnames(Xc) <- c("alpha1", "beta1", "alpha2", "beta2")
> Xc[1:n1,1:2] <- model.matrix(r1)
> Xc[(n1+1):n,3:4] <- model.matrix(r2)
> y <- c(rust0$rust[rust0$type == "ordinary"],rust0$rust[rust0$type == "non ordinary"])
```

En primer lloc hem de comprovar la homocedasticitat del model conjunt amb un contrast sobre les variàncies dels residus de les dues rectes.

```
> var.test(residuals(r1), residuals(r2))

        F test to compare two variances

data:  residuals(r1) and residuals(r2)
F = 1.0132, num df = 46, denom df = 32, p-value = 0.9836
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5200164 1.9005538
sample estimates:
ratio of variances
          1.013213
```

Cap problema.

La hipòtesi de paral·lelisme ens porta a un model on només hi ha un pendent, diguem-li β. Ara podem contrastar els dos models.

```
> beta <- c(Xc[1:n1,"beta1"],Xc[(n1+1):n,"beta2"])
> gc <- lm(y ~ 0 + alpha1 + beta1 + alpha2 + beta2, data=as.data.frame(Xc))
> gp <- lm(y ~ 0 + alpha1 + alpha2 + beta, data=as.data.frame(Xc))
> anova(gp,gc)

Analysis of Variance Table

Model 1: y ~ 0 + alpha1 + alpha2 + beta
Model 2: y ~ 0 + alpha1 + beta1 + alpha2 + beta2
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     77 10489
2     76 10484  1    5.0244 0.0364 0.8492
```

Per a un nivell de significació del 0.05 acceptem la hipòtesi de paral·lelisme.

El test de coincidència també es pot fer per comparació de dos models, el model de dues rectes paral·leles i el model amb una única recta per a tots els punts.

```
> g <- lm(y ~ beta)
> anova(g,gp)

Analysis of Variance Table

Model 1: y ~ beta
Model 2: y ~ 0 + alpha1 + alpha2 + beta
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     78 10606
2     77 10489  1    116.83 0.8576 0.3573
```

De forma que acceptem el model de recta única per a tots el punts.

(h) (opcional) Considerem el model lineal següent:

```
> gm <- lm(rust ~ age * type, data=rust0)
> summary(gm)

Call:
lm(formula = rust ~ age * type, data = rust0)

Residuals:
     Min       1Q   Median       3Q      Max
-21.9758  -8.4534  -0.9796   7.9292  26.6815
```

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          2.110e+01  2.456e+00   8.592 8.09e-13 ***
age                  9.637e-04  1.606e-04   6.000 6.25e-08 ***
typenon ordinary     1.898e+00  4.148e+00   0.458    0.648
age:typenon ordinary 4.523e-05  2.370e-04   0.191    0.849
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.75 on 76 degrees of freedom
Multiple R-squared: 0.4991,       Adjusted R-squared: 0.4794
F-statistic: 25.24 on 3 and 76 DF,  p-value: 1.954e-11
```

En primer lloc, observem que la interacció `age:type` no és significativa. Això és equivalent a acceptar la hipòtesi de paral·lelisme de les dues rectes. Llavors el model és

```
> gm <- lm(rust ~ age + type, data=rust0)
> summary(gm)

Call:
lm(formula = rust ~ age + type, data = rust0)

Residuals:
    Min      1Q  Median      3Q     Max
-21.837  -8.597  -1.044   7.865  26.469

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.087e+01  2.134e+00   9.784 3.74e-15 ***
age              9.845e-04  1.174e-04   8.388 1.82e-12 ***
typenon ordinary 2.497e+00  2.697e+00   0.926    0.357
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.67 on 77 degrees of freedom
Multiple R-squared: 0.4989,       Adjusted R-squared: 0.4859
F-statistic: 38.33 on 2 and 77 DF,  p-value: 2.803e-12
```

Llavors l'efecte del factor `type` també és no significatiu, el que coincideix amb l'acceptació d'una única recta per a tots els punts.

A més, l'estimació dels paràmetres del model `gp` i `gm` coincideix de la següent forma:

```
> coef(gp)

      alpha1       alpha2         beta
2.087408e+01 2.337131e+01 9.845177e-04

> coef(gm)

     (Intercept)              age typenon ordinary
    2.087408e+01     9.845177e-04     2.497230e+00
```

El pendent comú β és el coeficient de la variable `age`. El coeficient α_1 coincideix amb el punt d'intercepció i α_2 és la suma del punt de intercepció i el coeficient del factor `type`. A més, el coeficient del factor `type` és no significatiu de forma que el model final és la recta comuna per a tots els punts, com hem deduït en l'anterior apartat.