**Prova parcial del 11 de novembre de 2013**

**Problema 1**

**(a)** En aquest model les f.p.e són les combinacions lineals *aα*+*bβ*+*cγ*+*dδ* tals que

*a*−*b*+*c*−*d*=0

**(b)**

(i) *α* no és fpe (ii) *α*+*β* sí (iii) *α*+*β*+*γ*+*δ* sí

Els estimadors MQ de (ii) i (iii) són:

library(MASS)

y <- c(2.1, 1.9, 2, 2.2, 1.2, 0.8, 1.1, 0.9, 5.1, 5.2, 4.9, 4.9, 7.9, 8, 8.2,

7.9)

x <- c(rep(c(1, 0, -1, 0), 4), rep(c(0, 1, 0, -1), 4), rep(c(1, 1, 0, 0), 4),

rep(c(2, 2, -1, -1), 4))

x <- matrix(x, ncol = 4, byrow = T)

betas <- ginv(t(x) %\*% x) %\*% t(x) %\*% y

# alpha + beta

sum(betas[1:2])

## [1] 5.006

# alpha + beta + gamma + delta

sum(betas)

## [1] 7

**( c )** Primer calcularem l'estimació de la variància dels errors

residus <- y - x %\*% betas

RSS <- sum(residus^2)

sigma2 <- RSS/(16 - 3)

La variància estimada de l'estimació de *α*+*β* és

a <- c(1, 1, 0, 0)

as.numeric(sigma2 \* t(a) %\*% ginv(t(x) %\*% x) %\*% a)

## [1] 0.004084

La variància estimada de l'estimació de 2*α*+*β*−*γ* és

b <- c(2, 1, -1, 0)

as.numeric(sigma2 \* t(b) %\*% ginv(t(x) %\*% x) %\*% b)

## [1] 0.005445

i la covariància

as.numeric(sigma2 \* t(a) %\*% ginv(t(x) %\*% x) %\*% b)

## [1] 0.002722

**(d)** Per fer el contrast de la hipòtesi *H*(1)0:*α*+*β*=5 considerem el estadístic *t* de Student:

t.est <- as.numeric((sum(betas[1:2]) - 5)/sqrt(sigma2 \* t(a) %\*% ginv(t(x) %\*%

x) %\*% a))

# p-valor

pt(t.est, 16 - 3, lower.tail = F) \* 2

## [1] 0.9236

De manera que acceptem la hipòtesi nul·la.

Per fer el contrast de la hipòtesi *H*(2)0:*α*+*β*=5,2*α*+*β*−*γ*=7 ho farem com un contrast de models. El model complet és l'anterior del que ja hem calculat la suma de quadrats residual. Ara hem de fer el mateix amb el model de la hipòtesi nul·la. En primer lloc observem que *α*+*β*=5,2*α*+*β*−*γ*=7 és equivalent a *β*=5−*α* i *γ*=*α*−2, llavors tenim

*α*−*γ*=2

*β*−*δ*=5−*α*−*δ*

*α*+*β*=5

2*α*+2*β*−*γ*−*δ*=12−*α*−*δ*

Si pasem les constants a la banda de les observacions tenim un nou vector d'observacions

y0 <- y - c(rep(2, 4), rep(5, 4), rep(5, 4), rep(12, 4))

La nova matriu de disseny té, en principi, dos paràmetres *α* i *δ*

x0 <- c(rep(c(0, 0), 4), rep(c(-1, -1), 4), rep(c(0, 0), 4), rep(c(-1, -1),

4))

matrix(x0, ncol = 2, byrow = T)

## [,1] [,2]

## [1,] 0 0

## [2,] 0 0

## [3,] 0 0

## [4,] 0 0

## [5,] -1 -1

## [6,] -1 -1

## [7,] -1 -1

## [8,] -1 -1

## [9,] 0 0

## [10,] 0 0

## [11,] 0 0

## [12,] 0 0

## [13,] -1 -1

## [14,] -1 -1

## [15,] -1 -1

## [16,] -1 -1

però és evident que el rang és 1 i en realitat els dos paràmetres són el mateix. Així doncs la matriu de disseny de la hipòtesi nul·la es pot simplificar

x0 <- c(rep(0, 4), rep(-1, 4), rep(0, 4), rep(-1, 4))

Ara ja podem calcular la suma de quadrats residual de la hipòtesi nul·la i el test *F*

beta0 <- solve(t(x0) %\*% x0) %\*% t(x0) %\*% y0

residus0 <- y0 - x0 \* beta0

RSS0 <- sum(residus0^2)

f.est <- ((RSS0 - RSS)/2)/(RSS/(16 - 3))

pf(f.est, 2, 16 - 3, lower.tail = F)

## [1] 0.8556

Acceptem la hipòtesi nul·la.

**Problema 2 (homes)**

Seleccionem les dades:

library(HistData)

data(GaltonFamilies)

# help(GaltonFamilies) escull el teu cas:

sexe <- "male"

# sexe <- 'female'

GaltonFills <- GaltonFamilies[GaltonFamilies$gender == sexe, ]

# selecció a l'atzar d'un únic fill o filla per familia

rownames(GaltonFills) <- 1:dim(GaltonFills)[1]

set.seed(123)

ff <- function(x) as.numeric(sample(as.character(which(GaltonFills$family ==

x)), 1))

ind <- sapply(unique(GaltonFills$family), ff)

# data.frame per treballar

GaltonFills1 <- GaltonFills[ind, ]

**(a)** Fem un gràfic de dispersió:

with(GaltonFills1, plot(father, childHeight))

Fem també una primera regressió:

g <- lm(childHeight ~ father, data = GaltonFills1)

with(GaltonFills1, plot(father, childHeight))

abline(g)

i considerem una anàlisi gràfica dels residus amb la instrucció

plot(g)

Dels quatre gràfics resultants observem dos residus molt grans: 245 i 216.

plot(g, which = 1)

**(b)** Aquest apartat es pot contestar amb el següent resultat:

(ss <- summary(g))

##

## Call:

## lm(formula = childHeight ~ father, data = GaltonFills1)

##

## Residuals:

## Min 1Q Median 3Q Max

## -9.43 -1.43 0.09 1.63 5.79

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 38.4308 4.6288 8.30 2.5e-14 \*\*\*

## father 0.4460 0.0669 6.66 3.3e-10 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 2.27 on 177 degrees of freedom

## Multiple R-squared: 0.201, Adjusted R-squared: 0.196

## F-statistic: 44.4 on 1 and 177 DF, p-value: 3.28e-10

i de forma individual així:

coef(g)

## (Intercept) father

## 38.431 0.446

ss$sigma^2

## [1] 5.173

ss$r.squared

## [1] 0.2005

La regressió és significativa ja que *F* = 44.3911, 1, 177 i el seu p-valor és inferior a 0.05.

**( c )** La normalitat dels residus es pot comprovar amb un gràtic

plot(g, which = 2)

o amb un test

shapiro.test(residuals(g))

##

## Shapiro-Wilk normality test

##

## data: residuals(g)

## W = 0.9869, p-value = 0.09606

Acceptem la normalitat dels residus.

La normalitat no és cap garantia de mínima variància. El teorema de Gauss-Markov diu que el métode dels mínims quadrats amb les tres hipòtesis de Gauss-Markov (no cal la normalitat) és el que garantitza la mínima variància dels estimadors dels paràmetres o f.p.e.

**(d)** L'interval de confiança al 90% és

confint(g, level = 0.9)

## 5 % 95 %

## (Intercept) 30.7770 46.0846

## father 0.3353 0.5567

Per a la variància del model cal fer-ho a ma

ss$df[2]

## [1] 177

RSS <- sum(ss$residuals^2)

c(RSS/qchisq(0.95, ss$df[2]), RSS/qchisq(0.05, ss$df[2]))

## [1] 4.380 6.219

**(e)** La primera hipòtesi es pot contrastar fent servir la *t* de Student

(0.446 - 0.495)/0.0669

## [1] -0.7324

o com un contrast de models

g0 <- lm(childHeight ~ offset(I(0.495 \* father)), data = GaltonFills1)

anova(g0, g)

## Analysis of Variance Table

##

## Model 1: childHeight ~ offset(I(0.495 \* father))

## Model 2: childHeight ~ father

## Res.Df RSS Df Sum of Sq F Pr(>F)

## 1 178 918

## 2 177 916 1 2.77 0.54 0.47

Les dades de Galton no permeten rebutjar la hipòtesi.

Per a la següent hipòtesi cal que fem un nou model amb les dades de Galton:

gm <- lm(childHeight ~ mother, data = GaltonFills1)

g0 <- lm(childHeight ~ offset(I(0.45 \* mother)), data = GaltonFills1)

anova(g0, gm)

## Analysis of Variance Table

##

## Model 1: childHeight ~ offset(I(0.45 \* mother))

## Model 2: childHeight ~ mother

## Res.Df RSS Df Sum of Sq F Pr(>F)

## 1 178 1119

## 2 177 1081 1 38.4 6.29 0.013 \*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Ara rebutjem la hipòtesi nul·la.

La darrera hipòtesi es contrasta així:

g0 <- lm(childHeight ~ 0 + offset(I(104.023/2.54 + 0.45 \* mother)), data = GaltonFills1)

anova(g0, gm)

## Analysis of Variance Table

##

## Model 1: childHeight ~ 0 + offset(I(104.023/2.54 + 0.45 \* mother))

## Model 2: childHeight ~ mother

## Res.Df RSS Df Sum of Sq F Pr(>F)

## 1 179 1164

## 2 177 1081 2 83.3 6.82 0.0014 \*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

El rebuig de la hipòtesi era d'esperar, ja que hem rebutjat l'anterior.

En el cas de les dades de Galton, el pendent de la recta no té unitats ja que in2 / in2 simplifica. El mateix passa amb el pendent amb les dades espanyoles cm2 / cm2. El coeficient d'intercepció amb les dades de Galton està en polzades (in). En l'altre cas, les unitats són centímetres.

**(f)** La predicció per a la resposta mitjana és

predict(g, newdata = data.frame(father = 70), interval = "confidence", level = 0.99)

## fit lwr upr

## 1 69.65 69.18 70.12

En el model que tenim no podem fer servir l'altura de la mare.

**(g)** És evident que seria millor fer servir les altures de pare i mare en el model. Si tenim en compte que Galton va ser el pioner en el tema de la regressió, la idea de combinar les dues altures en una única variable regressora va ser bona. Ara podem considerar millor un model de regressió múltiple amb les dues variables com regressores o explicatives.

**(h)** Si considerem tots els fills homes d'una mateixa familia podem tenir problemes de correlació (o dependència) entre les dades o els errors, això aniria en contra de la tercera condició de Gauss-Markov i el mètode dels mínims quadrats no tindria garantides les seves bones propietats.

**Problema 2 (dones)**

Seleccionem les dades:

library(HistData)

data(GaltonFamilies)

# help(GaltonFamilies) escull el teu cas: sexe <- 'male'

sexe <- "female"

GaltonFills <- GaltonFamilies[GaltonFamilies$gender == sexe, ]

# selecció a l'atzar d'un únic fill o filla per familia

rownames(GaltonFills) <- 1:dim(GaltonFills)[1]

set.seed(123)

ff <- function(x) as.numeric(sample(as.character(which(GaltonFills$family ==

x)), 1))

ind <- sapply(unique(GaltonFills$family), ff)

# data.frame per treballar

GaltonFills1 <- GaltonFills[ind, ]

**(a)** Fem un gràfic de dispersió:

with(GaltonFills1, plot(mother, childHeight))

Fem també una primera regressió:

g <- lm(childHeight ~ mother, data = GaltonFills1)

with(GaltonFills1, plot(mother, childHeight))

abline(g)

i considerem una anàlisi gràfica dels residus amb la instrucció

plot(g)

Dels quatre gràfics resultants observem dos residus molt grans: 281 i 453.

plot(g, which = 1)

**(b)** Aquest apartat es pot contestar amb el següent resultat:

(ss <- summary(g))

##

## Call:

## lm(formula = childHeight ~ mother, data = GaltonFills1)

##

## Residuals:

## Min 1Q Median 3Q Max

## -6.765 -1.493 0.018 1.462 5.168

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 44.1718 4.4990 9.82 < 2e-16 \*\*\*

## mother 0.3110 0.0701 4.44 1.6e-05 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 2.12 on 174 degrees of freedom

## Multiple R-squared: 0.102, Adjusted R-squared: 0.0964

## F-statistic: 19.7 on 1 and 174 DF, p-value: 1.62e-05

i de forma individual així:

coef(g)

## (Intercept) mother

## 44.172 0.311

ss$sigma^2

## [1] 4.509

ss$r.squared

## [1] 0.1016

La regressió és significativa ja que *F* = 19.6755, 1, 174 i el seu p-valor és inferior a 0.05.

**( c )** La normalitat dels residus es pot comprovar amb un gràtic

plot(g, which = 2)

o amb un test

shapiro.test(residuals(g))

##

## Shapiro-Wilk normality test

##

## data: residuals(g)

## W = 0.9957, p-value = 0.8954

Acceptem la normalitat dels residus.

La normalitat no és cap garantia de mínima variància. El teorema de Gauss-Markov diu que el métode dels mínims quadrats amb les tres hipòtesis de Gauss-Markov (no cal la normalitat) és el que garantitza la mínima variància dels estimadors dels paràmetres o f.p.e.

**(d)** L'interval de confiança al 90% és

confint(g, level = 0.9)

## 5 % 95 %

## (Intercept) 36.7321 51.612

## mother 0.1951 0.427

Per a la variància del model cal fer-ho a ma

ss$df[2]

## [1] 174

RSS <- sum(ss$residuals^2)

c(RSS/qchisq(0.95, ss$df[2]), RSS/qchisq(0.05, ss$df[2]))

## [1] 3.813 5.430

**(e)** Ara necessitem les dades dels fills homes:

sexe <- "male"

GaltonFills <- GaltonFamilies[GaltonFamilies$gender == sexe, ]

# selecció a l'atzar d'un únic fill o filla per familia

rownames(GaltonFills) <- 1:dim(GaltonFills)[1]

set.seed(123)

ind <- sapply(unique(GaltonFills$family), ff)

# data.frame per treballar

GaltonFills.homes <- GaltonFills[ind, ]

La primera hipòtesi es pot contrastar si considerem dos models

gh <- lm(childHeight ~ father, data = GaltonFills.homes)

g0 <- lm(childHeight ~ offset(I(0.495 \* father)), data = GaltonFills.homes)

anova(g0, gh)

## Analysis of Variance Table

##

## Model 1: childHeight ~ offset(I(0.495 \* father))

## Model 2: childHeight ~ father

## Res.Df RSS Df Sum of Sq F Pr(>F)

## 1 178 918

## 2 177 916 1 2.77 0.54 0.47

Les dades de Galton no permeten rebutjar la hipòtesi.

Per a la següent hipòtesi cal que fem un nou model amb les dades de Galton:

gm <- lm(childHeight ~ mother, data = GaltonFills.homes)

g0 <- lm(childHeight ~ offset(I(0.45 \* mother)), data = GaltonFills.homes)

anova(g0, gm)

## Analysis of Variance Table

##

## Model 1: childHeight ~ offset(I(0.45 \* mother))

## Model 2: childHeight ~ mother

## Res.Df RSS Df Sum of Sq F Pr(>F)

## 1 178 1119

## 2 177 1081 1 38.4 6.29 0.013 \*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Ara rebutjem la hipòtesi nul·la.

La darrera hipòtesi es contrasta així:

g0 <- lm(childHeight ~ 0 + offset(I(104.023/2.54 + 0.45 \* mother)), data = GaltonFills.homes)

anova(g0, gm)

## Analysis of Variance Table

##

## Model 1: childHeight ~ 0 + offset(I(104.023/2.54 + 0.45 \* mother))

## Model 2: childHeight ~ mother

## Res.Df RSS Df Sum of Sq F Pr(>F)

## 1 179 1164

## 2 177 1081 2 83.3 6.82 0.0014 \*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

El rebuig de la hipòtesi era d'esperar, ja que hem rebutjat l'anterior.

En el cas de les dades de Galton, el pendent de la recta no té unitats ja que in2 / in2 simplifica. El mateix passa amb el pendent amb les dades espanyoles cm2 / cm2. El coeficient d'intercepció amb les dades de Galton està en polzades (in). En l'altre cas, les unitats són centímetres.

**(f)** La predicció per a la resposta mitjana és

predict(g, newdata = data.frame(mother = 63), interval = "confidence", level = 0.99)

## fit lwr upr

## 1 63.77 63.3 64.23

En el model que tenim no podem fer servir l'altura del pare.

**(g)** És evident que seria millor fer servir les altures de pare i mare en el model. Si tenim en compte que Galton va ser el pioner en el tema de la regressió, la idea de combinar les dues altures en una única variable regressora va ser bona. Ara podem considerar millor un model de regressió múltiple amb les dues variables com regressores o explicatives.

**(h)** Si considerem totes les filles d'una mateixa familia podem tenir problemes de correlació (o dependència) entre les dades o els errors, això aniria en contra de la tercera condició de Gauss-Markov i el mètode dels mínims quadrats no tindria garantides les