TD 2.4 Régression linéaire simple et multiple

# Régression linéaire simple

## Visualisation des données

#### Exclure des valeurs trop extrêmes

> plot(masse~taille, main="Tous")

> tai <- taille[taille>130]

> mas <- masse[taille>130]

> plot(mas~tai, main="La partie lin´eaire")

## Estimation des coefficients

#### Signification des valeurs

Linear form

> lm1 <- lm(mas~tai)

> summary(lm1)

Call:

lm(formula = mas ~ tai)

Residuals:

Min 1Q Median 3Q Max

-23,8116 -7,1624 -0,6126 5,7381 30,5267

Coefficients:

P value

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

(Intercept) -121,17315 16,62910 -7,287 6,14e-11 \*\*\*

Sigma resi

Tai 1,08458 0,09581 11,320 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0,001 ‘\*\*’ 0,01 ‘\*’ 0,05 ‘.’ 0,1 ‘ ’ 1

Residual standard error: 10,03 on 105 degrees of freedom

Multiple R-squared: 0.5496, Adjusted R-squared: 0.5454

F-statistic: 128.1 on 1 and 105 DF, p-value: < 2,2e-16

R2 (part de variation expliquée)

Nombre de Coeff

#### Rajouter droite de régression sur graph (y = bx + a + E)

> plot(mas~tai, pch="+",xlab="Taille en cm", ylab="Masse en kg", main="Masse selon la taille")

> abline(lm1,lty=2)

#### Intervalle de confiance et Validation du modèle ANOVA

Nuage de point

> plot(mas~tai, pch="+",xlab="Taille en cm", ylab="Masse en kg",ylim=c(0,110),col="darkgrey")

> new <- data.frame(tai = seq(130, 200, 0.5))

> ICprediction <- predict(lm(mas ~ tai), new, interval="prediction")

> ICdroite <- predict(lm(mas ~ tai), new, interval="confidence")

Dessine plusieurs lignes simultanément

> matlines(new$tai,cbind(ICdroite, ICprediction[,-1]),

+ lty=c(1,2,2,4,4), type="l", ylab="predicted y",col=c(1,2,2,4,4))

> legend("topleft",lty=c(2,4),c("IC de droite","IC de pr´ediction"), col=c(2,4))

> title("Nuage, droite de r´egression et intervalles de confiance")

# Régression linéaire multiple

## Estimation des coefficients

lm(rep~X1+X2) ## Mod`ele additif, sans interaction

**OU**

lm(rep~X1\*X2) ## Mod`ele multiplicatif, avec interaction

Vérifier rapidement les croisements avant de commencer grâce à « pairs »

> irisQT <- iris[1:4]

> pairs(irisQT, main = "Anderson's Iris Data -- 3 species",

+ pch = 21, bg = c("red", "green3", "blue")[unclass(iris$Species)])

> sl <- irisQT[,1] ; sw <- irisQT[,2] ; pl <- irisQT[,3] ; pw <- irisQT[,4]

> lm2 <- lm(sl~sw\*pl\*pw)

> summary(lm2)

Call:

lm(formula = sl ~ sw \* pl \* pw)

Residuals:

Min 1Q Median 3Q Max

-0,78826 -0,23215 0,01429 0,18533 0,74079

Coefficients:

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

(Intercept) 1,73247 0,87187 1,987 0,0488 \*

sw 0,73953 0,27311 2,708 0,0076 \*\*

pl 1,04893 0,49247 2,130 0,0349 \*

pw -2,18321 1,48898 -1,466 0,1448

sw:pl -0,13072 0,16361 -0,799 0,4256

sw:pw 0,45030 0,48976 0,919 0,3594

pl:pw 0,18729 0,22641 0,827 0,4095

sw:pl:pw -0,04066 0,07567 -0,537 0,5919

---

Signif. codes: 0 ‘\*\*\*’ 0,001 ‘\*\*’ 0,01 ‘\*’ 0,05 ‘.’ 0,1 ‘ ’ 1

Residual standard error: 0,3119 on 142 degrees of freedom

Multiple R-squared: 0.8648, Adjusted R-squared: 0.8581

F-statistic: 129.7 on 7 and 142 DF, p-value: < 2,2e-16

## Sélection de variable

> step(lm(sl~sw+pl+pw) , k=2)

Start: AIC=-343,04

sl ~ sw + pl + pw

Df Sum of Sq RSS AIC

<none> 14,445 -343,04

- pw 1 1,8834 16,329 -326,66

- sw 1 9,4353 23,881 -269,63

- pl 1 15,4657 29,911 -235,86

Call:

lm(formula = sl ~ sw + pl + pw)

Coefficients:

(Intercept) sw pl pw

1,8560 0,6508 0,7091 -0,5565