Introduction à la régression
cours n°3

ENSM.SE – axe MSA

Rappel modèle linéaire

- Y vecteur des réponses
- X matrice du plan d'expériences
- β vecteurs des paramètres
- ε vecteurs des écarts au modèle :

$$Y = X\beta + \varepsilon$$

$$\varepsilon \text{ de loi } N(0, \sigma^2 Id)$$

Intervalles de confiance et tests

(i) Le vecteur $\hat{\beta}$ est de loi normale $N(\beta, \sigma^2(X' X)^{-1})$.

(ii) $\dfrac{(n-p-1)\,\hat{\sigma}^2}{\sigma^2} = \dfrac{\|Y - X\hat{\beta}\|^2}{\sigma^2}$ est de loi χ^2_{n-p-1} et $\hat{\sigma}$ est indépendant de $\hat{\beta}$

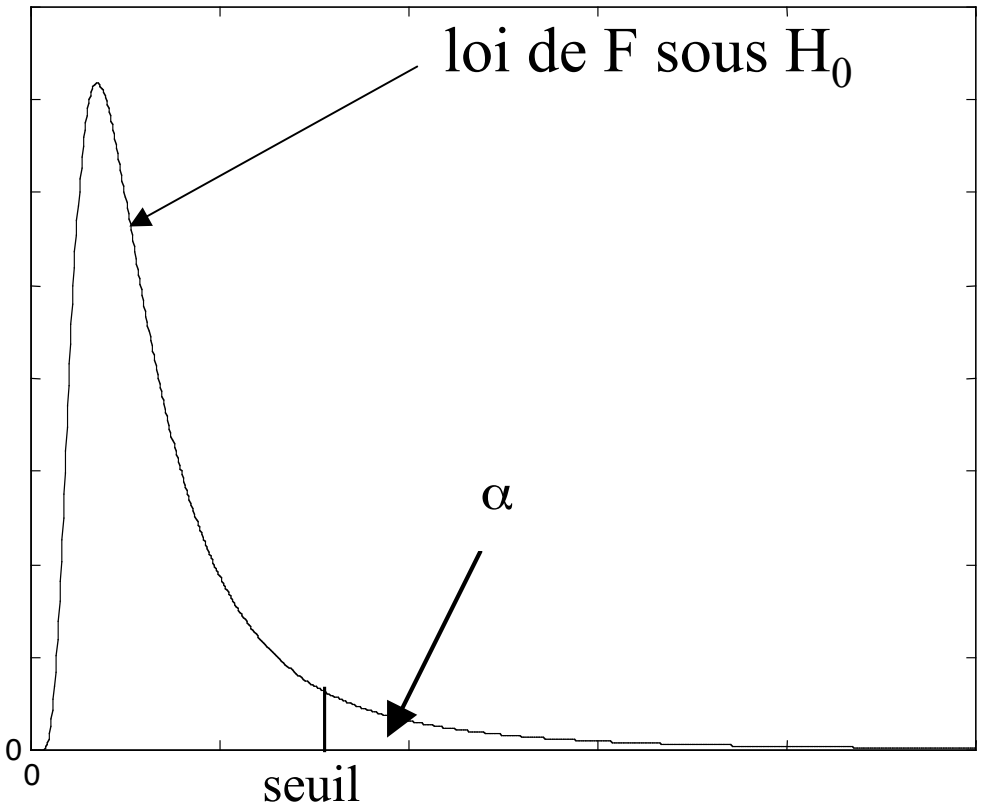
.

(iii) Pour tout $j \in \{0, \ldots, p\}$, en notant c_j le j-ème terme diagonal de la matrice $(X' X)^{-1}$, la variable $\dfrac{\hat{\beta}_j - \beta_j}{\sqrt{c_j}\,\hat{\sigma}}$ est de loi de Student t_{n-p-1}.

(iv) **Sous l'hypothèse** H_0 : $\beta_1 = \ldots = \beta_p = 0$, l'écart relatif

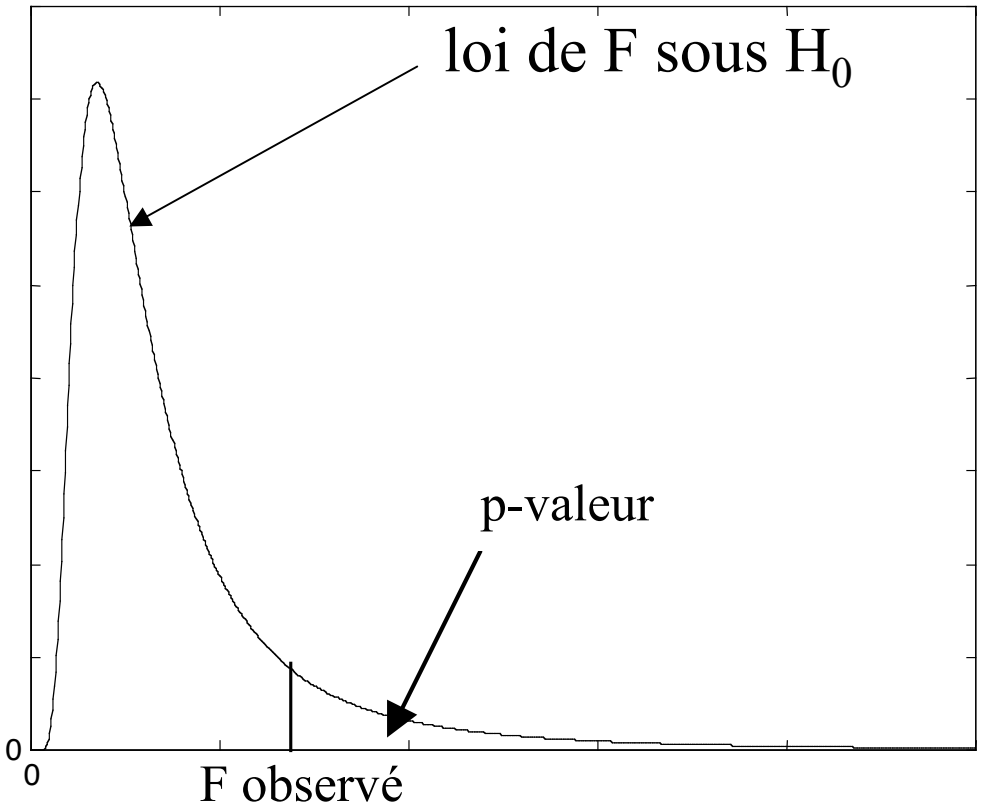
$\dfrac{\|X^0 \bar{y} - X\hat{\beta}\|^2}{p\,\hat{\sigma}^2} = \dfrac{MSR}{MSE}$ est de loi de Fisher-Snedecor F^p_{n-p-1}.

test de niveau α et p-valeur



test basé sur la statistique F, de niveau α

test de niveau α et p-valeur



p-valeur associée à la statistique F

Table d'ANOVA complète

Source	DF	SS	MS	F	p
Regression	p	SSR	SSR/p	MSR/MSE	p value
Error	n-p-1	SSE	SSE/(n-p-1)		
Total	n-1	SST			
R^2	SSR/SST				

Tables d'ANOVA pour les prédicteurs

Variable	Coeff	Std	t value	p value
Intercept	$\hat{\beta}_0$	$\sqrt{c_0}\,\hat{\sigma}$	$Coeff_0/Std_0$	
x_1	$\hat{\beta}_1$	$\sqrt{c_1}\,\hat{\sigma}$	$Coeff_1/Std_1$	
.	.	.	.	
.	.	.	.	
.	.	.	.	
x_p	$\hat{\beta}_p$	$\sqrt{c_p}\,\hat{\sigma}$	$Coeff_p/Std_p$	

Exemple 2

Modèle rendement $\approx \beta_0 + \beta_{pluie}$ pluie

Variable	Coeff	Std	t value	p value
Intercept	0.0662	0.02405	2.752	0.00813
pluie	1.637	0.07526	21.75	0

Modèle rendement $\approx \beta_0 + \beta_{pluie}$ pluie $+ \beta_{pluie\text{^}2}$ pluie?

Variable	Coeff	Std	t value	p value
Intercept	-004246	0.04512	-0.9411	0.3511
pluie	2.502	0.3187	7.85	2.494e-10
pluie?	-1.523	0.547	-2.784	0.007518

Exemple 2 - suite

Modèle rendement $\approx \beta_0 + \beta_{pluie}$ **pluie** $+ \beta_{pluie\hat{}2}$ **pluie?** $+ \beta_{pluie\hat{}3}$ **pluie**^{3}

Variable	Coeff	Std	t value	p value
Intercept	-0.01225	0.07243	-0.1691	0.8664
pluie	2.031	0.9363	2.169	0.03488
pluie?	0.4484	3.721	0.1205	0.9046
pluie^3	-2.419	4.516	-0.5357	0.5945

corrélation des prédicteurs…

Validation - résidus

- Validation basée sur les résidus estimés $\hat{\varepsilon}$
- Tracés des résidus :
  - séquentiels
  - contre la réponse estimée
  - contre les prédicteurs
- Mais variance non constante !!

Standardisation des résidus

- Cas du modèle $y=a+bx+\varepsilon$

$$h_{ii} = \frac{\mathrm{Var}(x) + (x_i - \bar{x})^?}{n\,\mathrm{Var}(x)}$$

- $T_i = \dfrac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$ est le i-ème résidu standardisé

Retour sur l'exemple 1

- Une réponse : tonnage cartons
- Quatre prédicteurs :
  - CA en keuros
  - surface en m^2
  - Prestataire : codé –1 (entreprise « O ») et 1 (entreprise « C »)
  - Catégorie : codée –1 (super) et 1 (hyper)
- Rappel
  - prendre les log (en base 10)
  - centrer les variables

pairs plot (données en log10 et centrées)

Matrice de corrélation des prédicteurs

	log10(CA)	log10(surface)	prestataire	catégorie
log10(CA)	1.0000000	0.9235001	0.3016459	0.8611131
log10(surface)	0.9235001	1.0000000	0.4186272	0.9030056
prestataire	0.3016459	0.4186272	1.0000000	0.3968538
catégorie	0.8611131	0.9030056	0.3968538	1.0000000

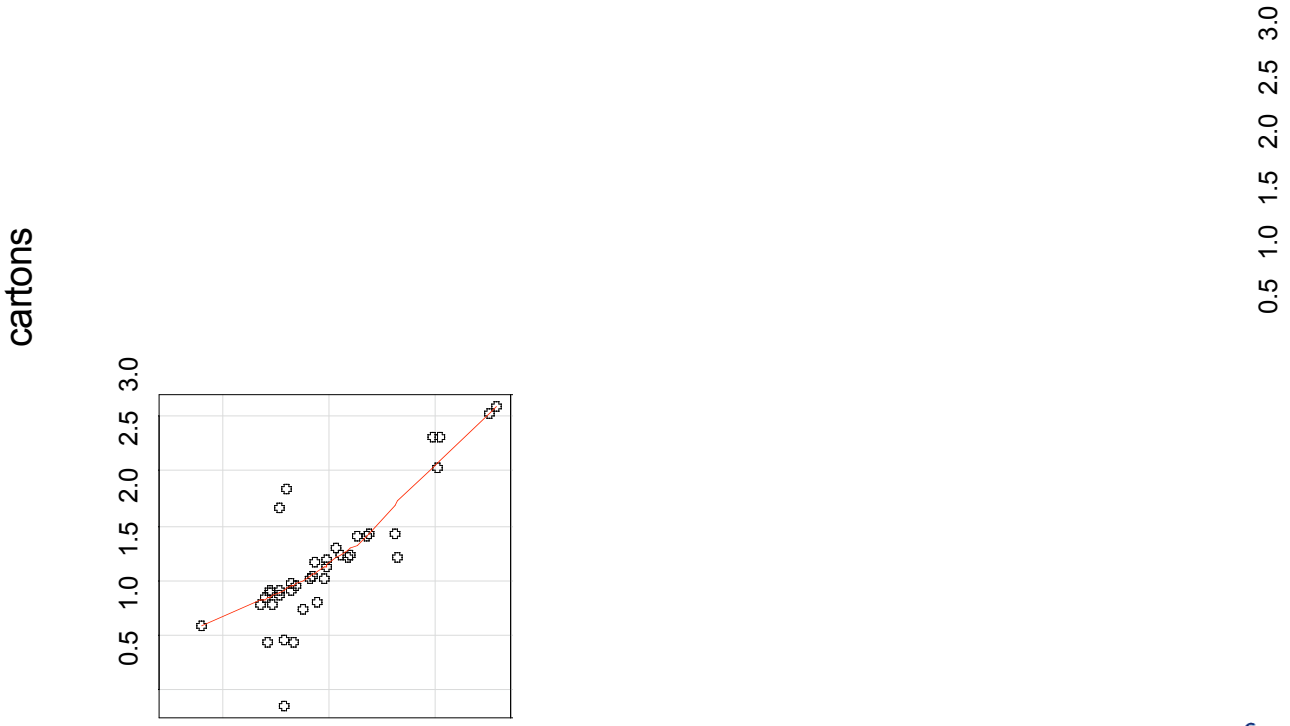
Changement de prédicteurs

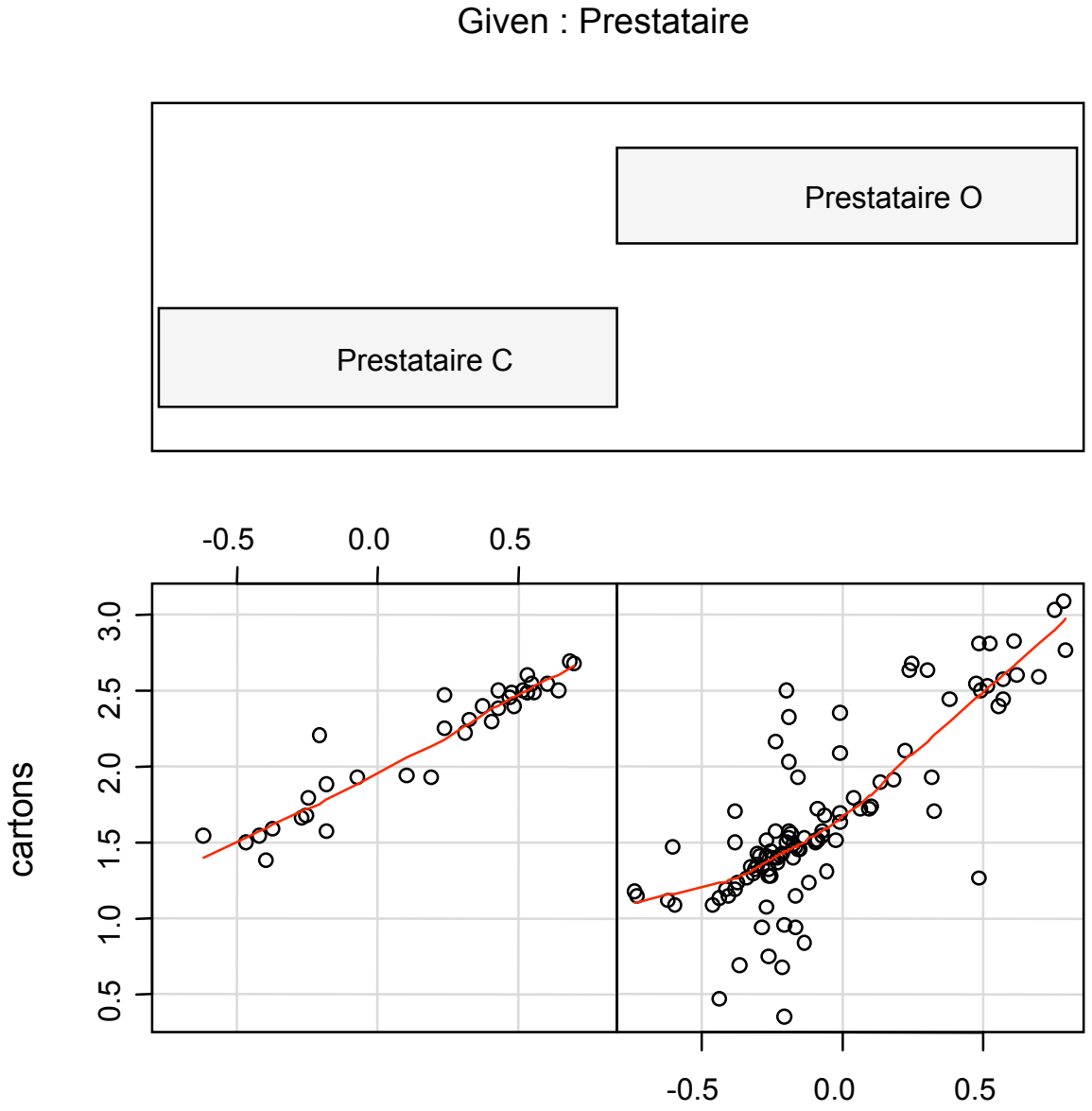
surface devient surface/CA

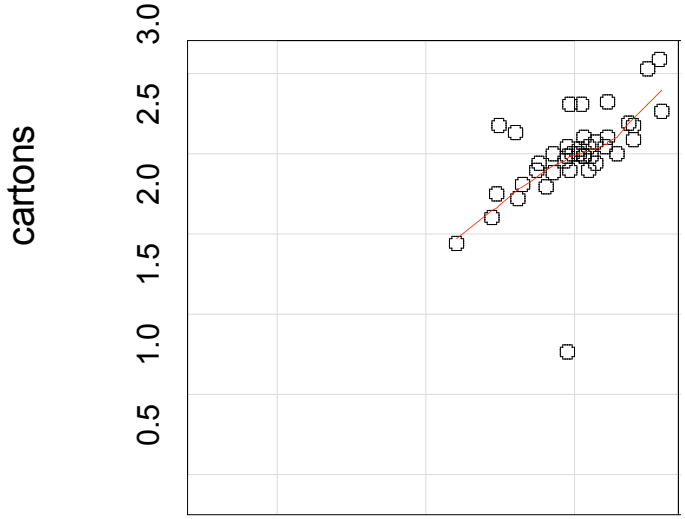
	log10(CA)	log10(surface/CA)	prestataire	catégorie
log10(CA)	1.0000000	-0.20359992	0.3016459	0.86111312
log10(surface/CA)	-0.2035999	1.00000000	0.2960516	0.09973161
prestataire	0.3016459	0.29605164	1.0000000	0.39685378
catégorie	0.8611131	0.09973161	0.3968538	1.00000000


cartons
0.5
0.0
-0.5
1.0
0.0
-1.0
2.5
1.5
0.5
0.4
0.0
-0.4
1.0
0.0
-1.0
Catégorie

Interactions




Given : Prestataire
Prestataire O
Prestataire C
-0.5
0.0
0.5
3.0
2.5
2.0
1.5
1.0
0.5
cartons
-0.5
0.0
0.5
CA


cartons
0.5
1.0
1.5
2.0
2.5
3.0

Conclusion temporaire

- Surface est à éliminer
- Pas d'interaction entre CA et prestataire
- Interaction entre CA et catégorie
- Variance constante ??

ANOVA

lm(formula = cartons ~ CA * Catégorie + Prestataire, data = ldechet)

Residual standard error: 0.3171 on 132 degrees of freedom

Multiple R-Squared: 0.7202, Adjusted R-squared: 0.7117

F-statistic: 84.94 on 4 and 132 DF, p-value: < 2.2e-16

Predictor	Estimate	Std. Error	t-value	Pr(>|t|)
Intercept	1.83420	0.08838	20.753	< 2e-16
CA	1.09201	0.18250	5.984	1.93e-08
Catégorie	0.07336	0.09158	0.801	0.4245
Prestataire	0.08236	0.03413	2.413	0.0172
CA:Catégorie	0.06218	0.18188	0.342	0.7330

ANOVA

lm(formula = cartons ~ CA + Prestataire, data = ldechet)

Residual standard error: 0.3178 on 134 degrees of freedom

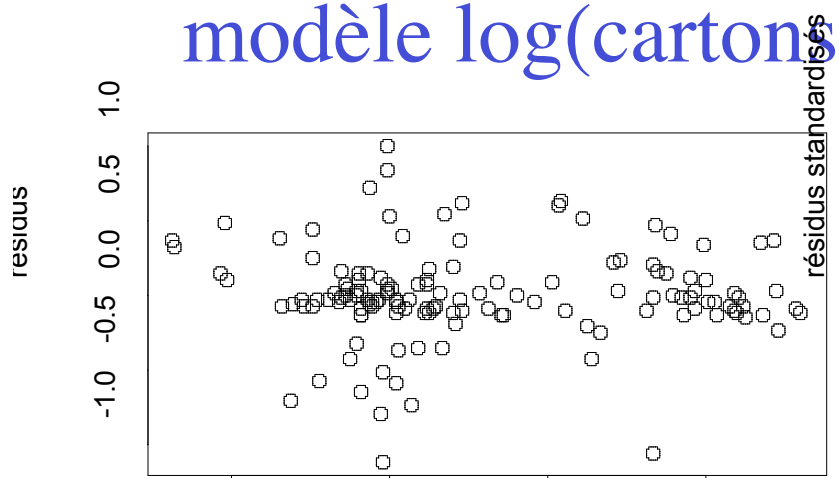
Multiple R-Squared: 0.7146, Adjusted R-squared: 0.7104

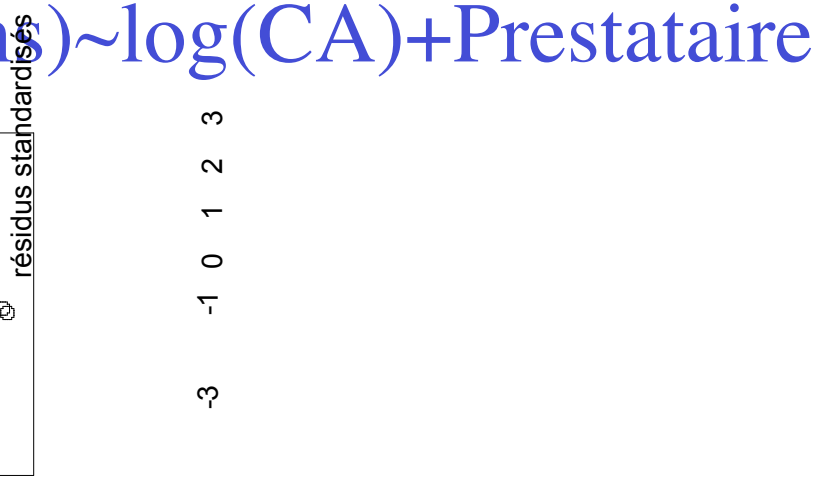
F-statistic: 167.8 on 2 and 134 DF, p-value: < 2.2e-16

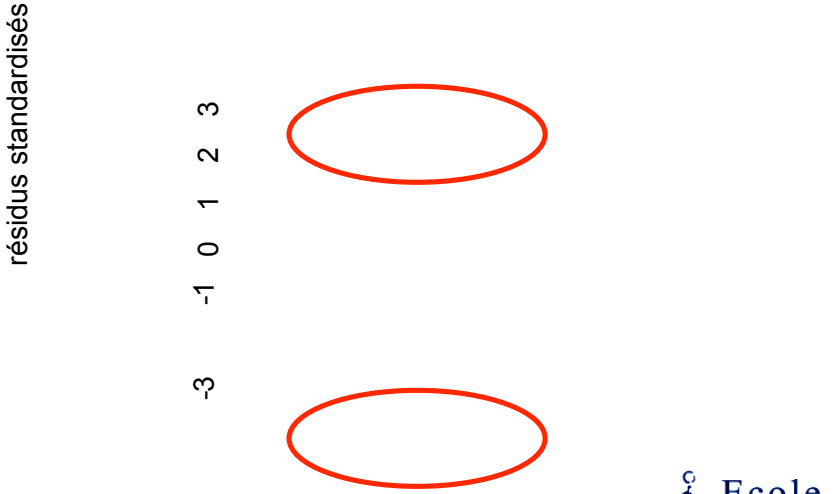
Predictor	Estimate	Std. Error	t-value	Pr(>|t|)
Intercept	1.82875	0.03150	58.058	< 2e-16
CA	1.24504	0.07616	16.347	< 2e-16
Prestataire	0.09640	0.03265	2.953	0.00372

Résidus standardisés

modèle log(cartons)~log(CA)+Prestataire





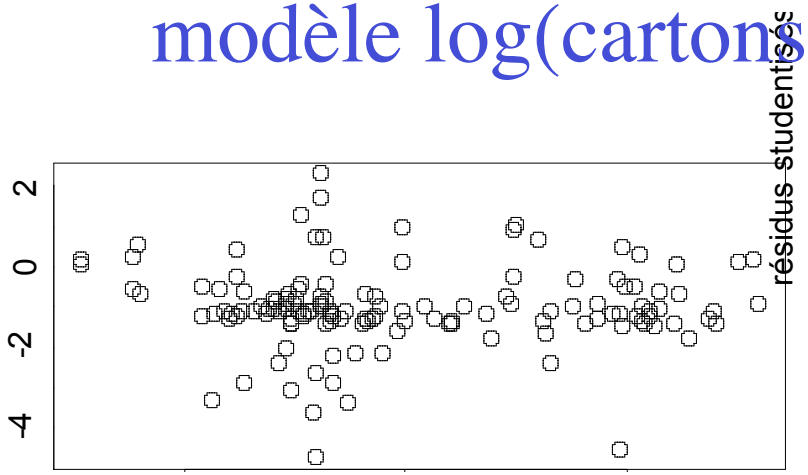


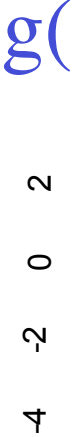
Résidus studentisés T_i^*

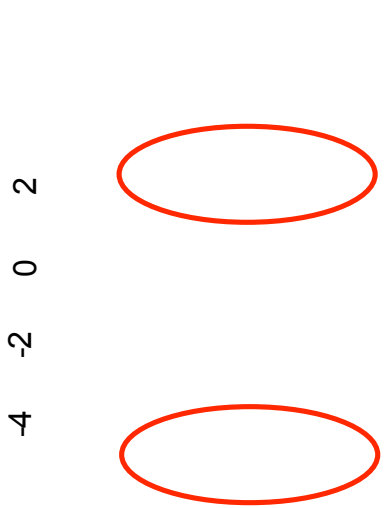
- Loi de T_i^* connue (Student t_{n-p-2}) :
  - Tracé séquentiel pour détecter des résidus trop forts
  - Test d'adéquation à la loi normale N(0,1) (très peu différente de t_{n-p-2})

Résidus studentisés

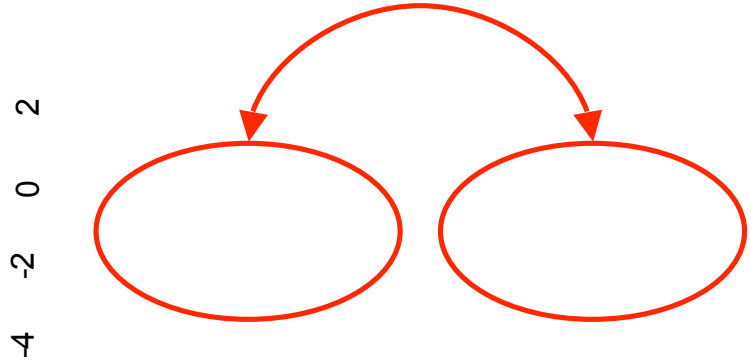
modèle log(cartons)~log(CA)+Prestataire







Effet de variance du prestataire



Catégorie

Prestataire

Exemple 3
distributeurs de boisson

- réponse = temps
- prédicteurs = distance, nombre de caisses
- après analyses graphiques préliminaires (cf. cours 2), modèle candidat :

$$\text{temps} = \beta_0 + \beta_{nb}\, nb + \beta_{dist}\, dist$$

Exemple 3 - ANOVA

Table d'analyse de variance

Source	df	SS	MS	F	p
Regression	2	5544	2772	261.7	4.441e-016
Error	22	233	10.59		
Total	24	5777			

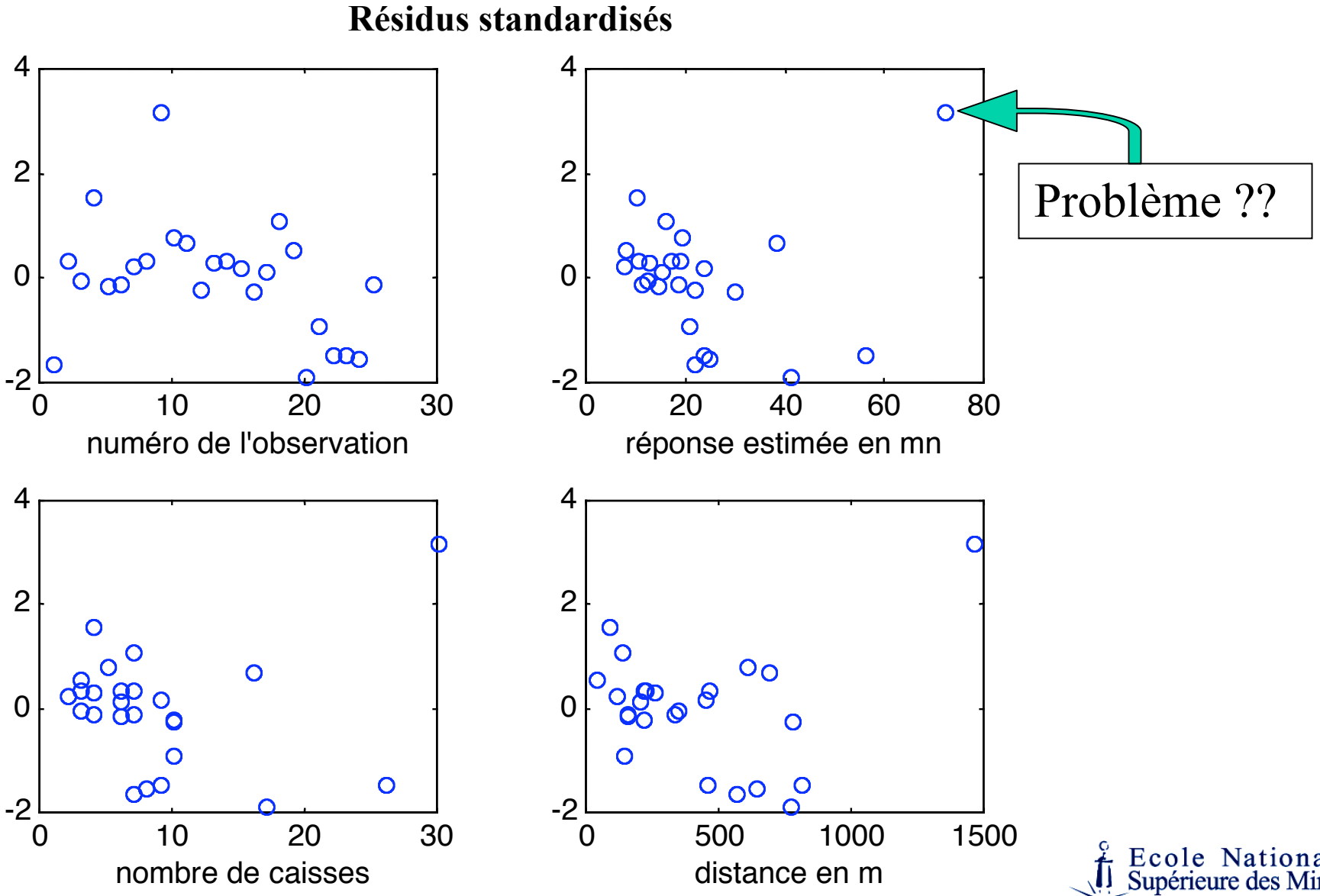
Coefficients

Root MSE	3.255	R-square	0.9597
		R-sq(adj)	0.956

Paramètres estimés

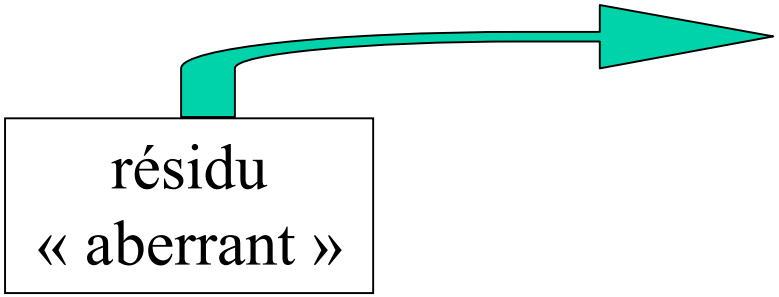
Predictor	Coeff	Stdev	t-ratio	p
intercept	2.353	1.095	2.149	0.04292
nb	1.615	0.1705	9.474	3.199e-009
dist	0.01437	0.003608	3.984	0.0006273

Exemple 3 - résidus



Exemple 3 – résidus studentisés

(avec intervalle à 95%)



Exemple 3 sans obs. n°9 - ANOVA

Table d'analyse de variance

Source	df	SS	MS	F	p
Regression	2	2291	1146	194.6	2.798e-014
Error	21	123.6	5.887		
Total	23	2415			

Coefficients

Root MSE	2.426	R-square	0.9488
		R-sq(adj)	0.9439

Paramètres estimés

Predictor	Coeff	Stdev	t-ratio	p
intercept	4.456	0.951	4.686	0.0001263
X2	1.497	0.13	11.52	1.552e-010
X3	0.01032	0.002849	3.621	0.001601

Exemple 3 – résidus studentisés

(en ôtant l'observation n°9)

Exemple 3 – droite de Henri

(en ôtant l'observation n°9)