

Introduction à la régression

cours n°2

ENSM.SE – axe MSA
L. Carraro

Plan

- ANOVA
- Prédicteurs qualitatifs
- Prédicteurs corrélés
- Interaction de prédicteurs
- Choix de modèles
- Validation, résidus.

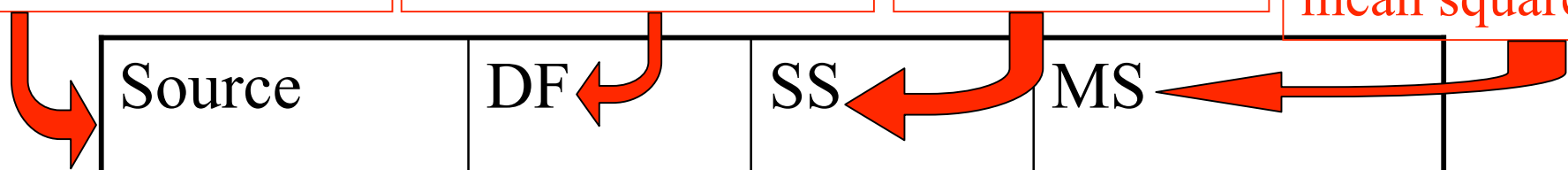
Analyse de variance (ANOVA)

source de variation

degrees of freedom

sum of squares

mean squares



Source	DF	SS	MS
Regression	p	SSR	SSR/p
Error	n-p-1	SSE	SSE/(n-p-1)
Total	n-1	SST	

Sum of squares

- $SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$
- $SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$
- $SST = \sum_{i=1}^n (y_i - \bar{y})^2$

Exercice

- Représenter graphiquement les vecteurs $\mathbf{Y}$, $\hat{\mathbf{Y}}$ et $\bar{\mathbf{Y}}$ de sorte que les quantités SSR, SSE et SST soient apparentes. Indiquer tous les angles droits et en déduire des relations.
- Donner la dimension des e.v. auxquels appartiennent les vecteurs :

$$\mathbf{Y} - \bar{\mathbf{Y}}, \hat{\mathbf{Y}} - \bar{\mathbf{Y}} \text{ et } \mathbf{Y} - \hat{\mathbf{Y}}$$

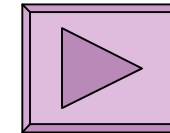
Formule d'ANOVA

- $SST = SSR + SSE$
- Coefficient de détermination :
 $R^2 = SSR/SST = 1 - SSE/SST$
- Utilisations de R^2
- Coefficient de détermination ajusté :

$$R^2_{\text{-adj}} = 1 - MSE/MST = 1 - \frac{\hat{\sigma}^2}{MST}$$

Questions de prédicteurs

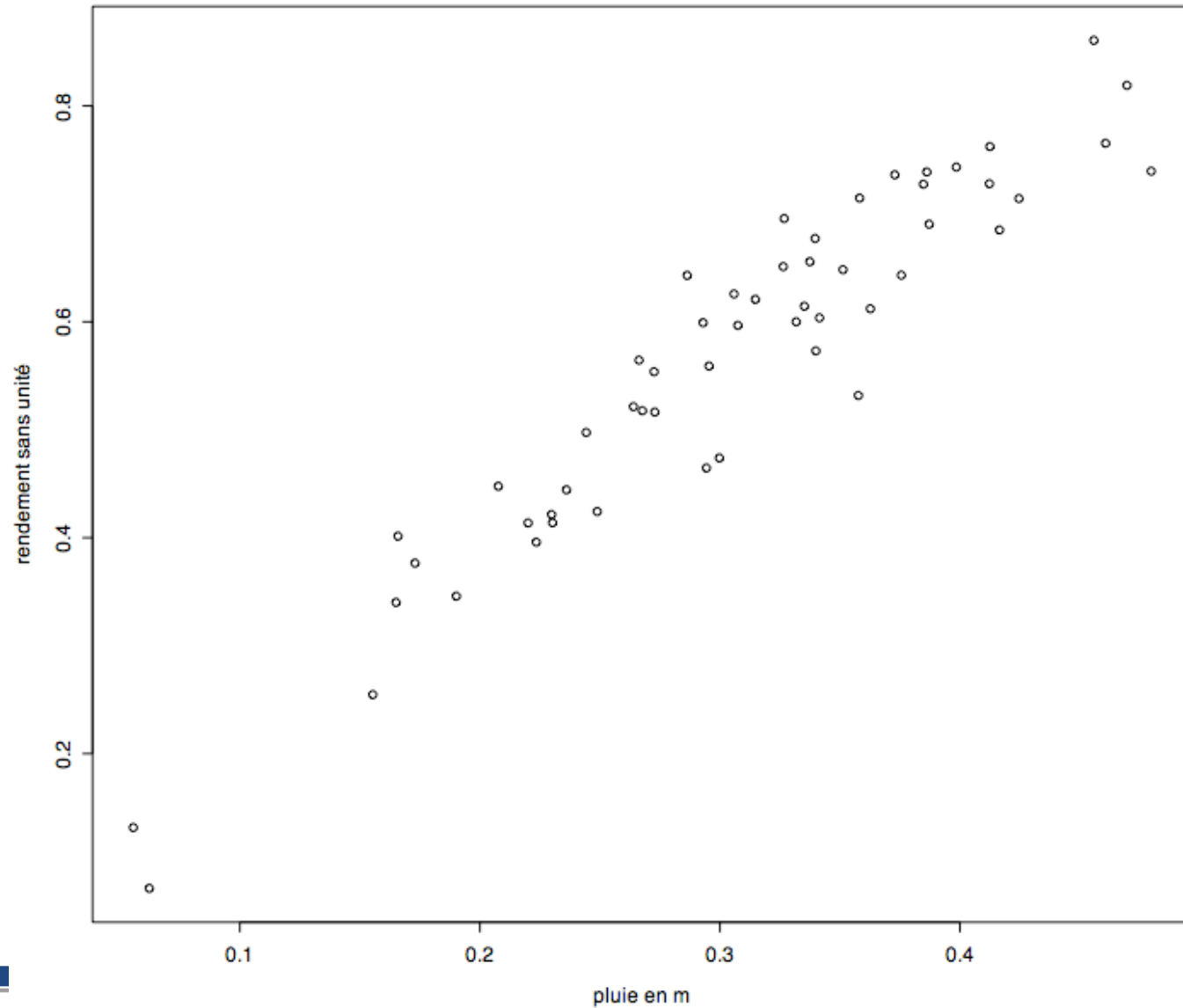
- prédicteurs qualitatifs
 - codage des niveaux
 - retour sur l'exemple des cartons
 - remarque sur R
- prédicteurs corrélés
 - problème d'identifiabilité
 - utilisation des tables d'ANOVA ou de sélection de paramètres.
- interaction de prédicteurs
 - notion **liée à la réponse**
 - sans rapport avec la corrélation



Corrélation de prédicteurs

- Exemple 2 :
 - Réponse = pourcentage d'un rendement maximal de blé
 - Prédicteur = quantité de pluies printanières (en m).
 - 54 observations

Exemple 2 : les données



Exemple 2

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

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

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 ²	-1.523	0.547	-2.784	0.007518

Exemple 2 - suite

$$\text{Modèle rendement} \approx \beta_0 + \beta_{\text{pluie}} \text{ pluie} + \beta_{\text{pluie}^2} \text{ pluie}^2 + \beta_{\text{pluie}^3} \text{ pluie}^3$$

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 ²	0.4484	3.721	0.1205	0.9046
pluie ³	-2.419	4.516	-0.5357	0.5945



corrélation des prédicteurs...

Matrice de corrélation des prédicteurs

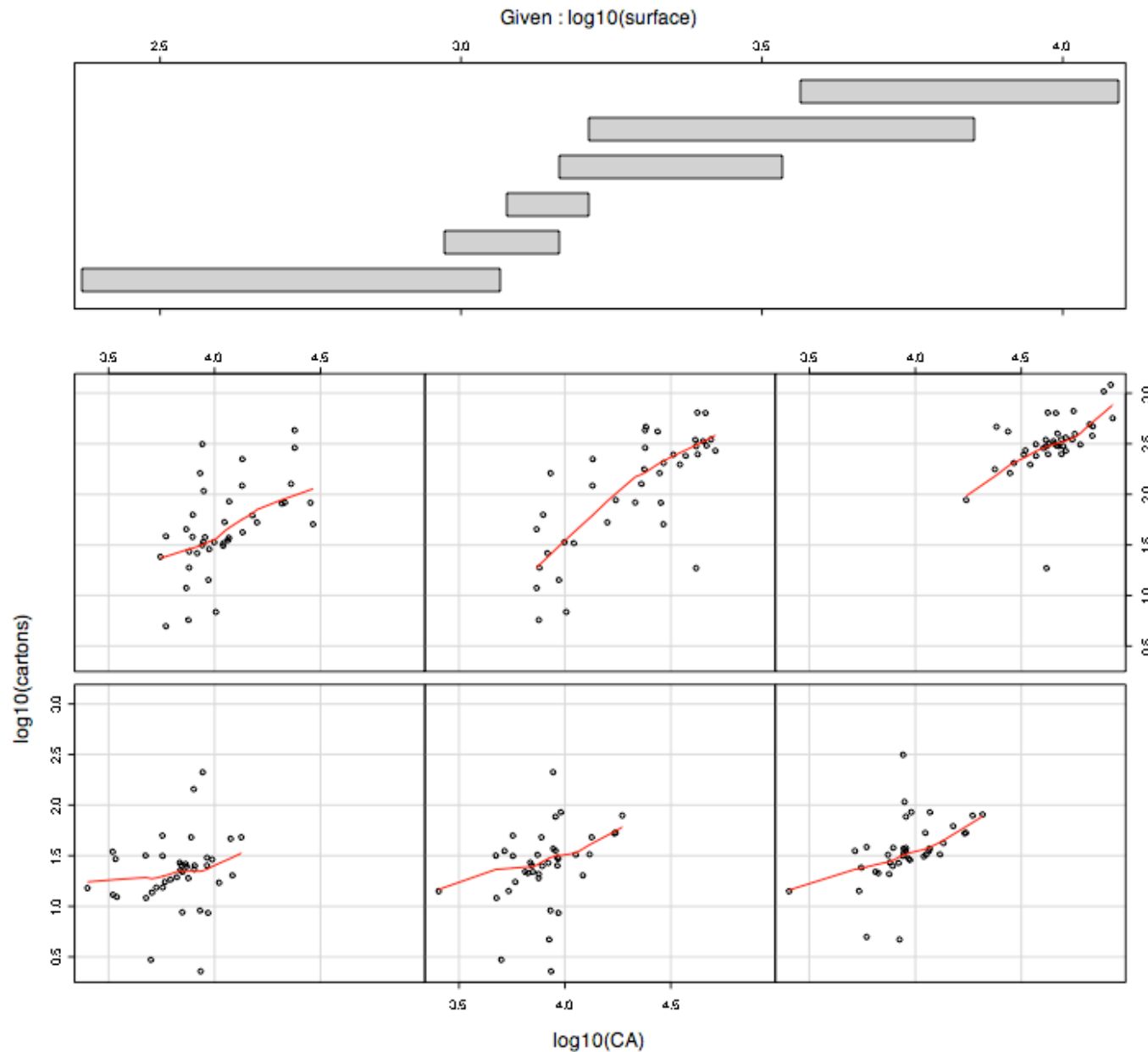
	log10(CA)	log10(surface)	prestataire	catégorie
log10(CA)	1.0000000	0.9235001	0.3016459	0.8611131
log10(surface)	0.9235001	1.0000000	0.4186272	0.9030056
prestataire	0.3016459	0.4186272	1.0000000	0.3968538
catégorie	0.8611131	0.9030056	0.3968538	1.0000000

Changement de prédicteurs

surface devient surface/CA

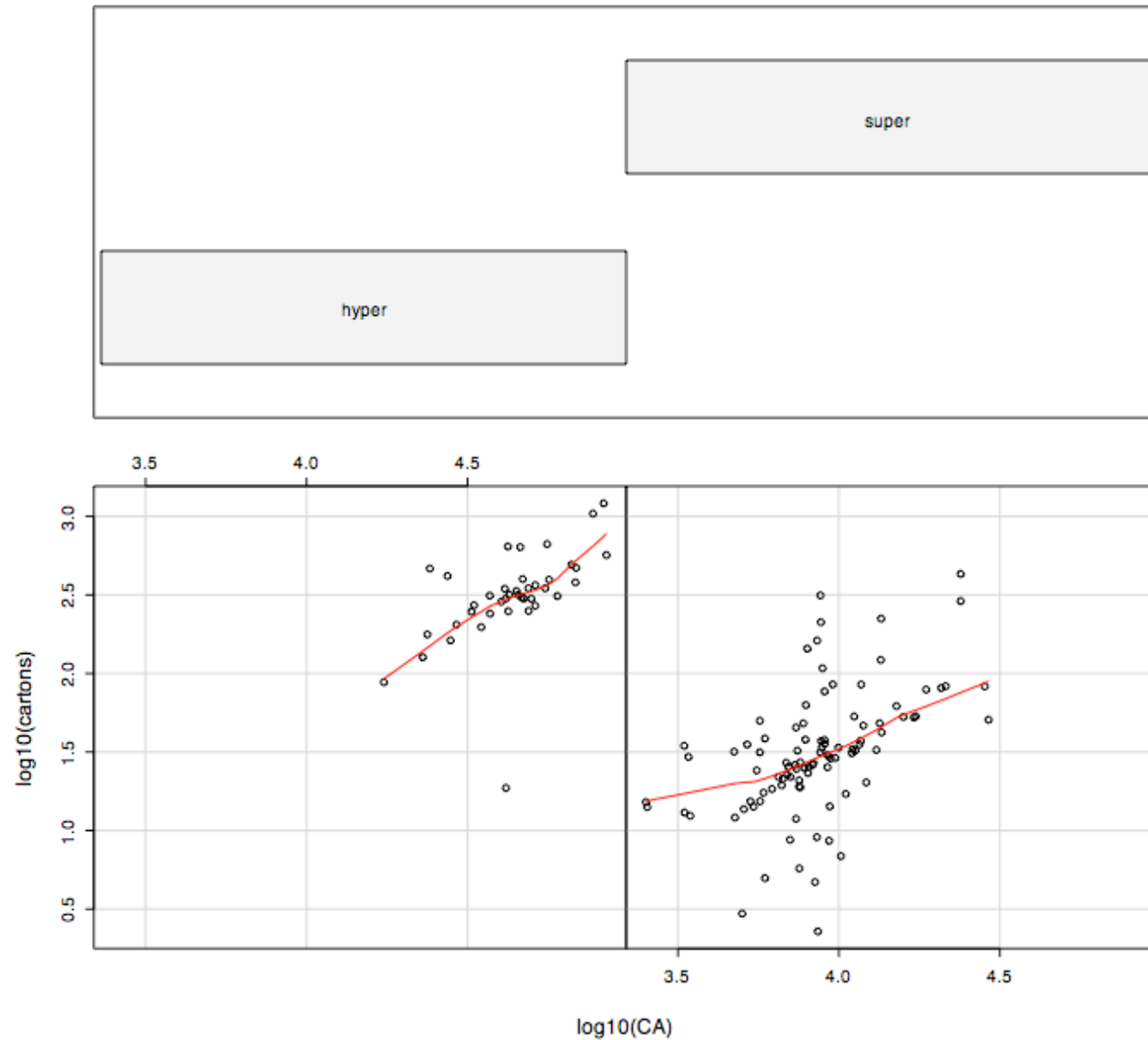
	log10(CA)	log10(surface/CA)	prestataire	catégorie
log10(CA)	1.0000000	-0.20359992	0.3016459	0.86111312
log10(surface/CA)	-0.2035999	1.00000000	0.2960516	0.09973161
prestataire	0.3016459	0.29605164	1.0000000	0.39685378
catégorie	0.8611131	0.09973161	0.3968538	1.00000000

Cartons - interaction CA-surface



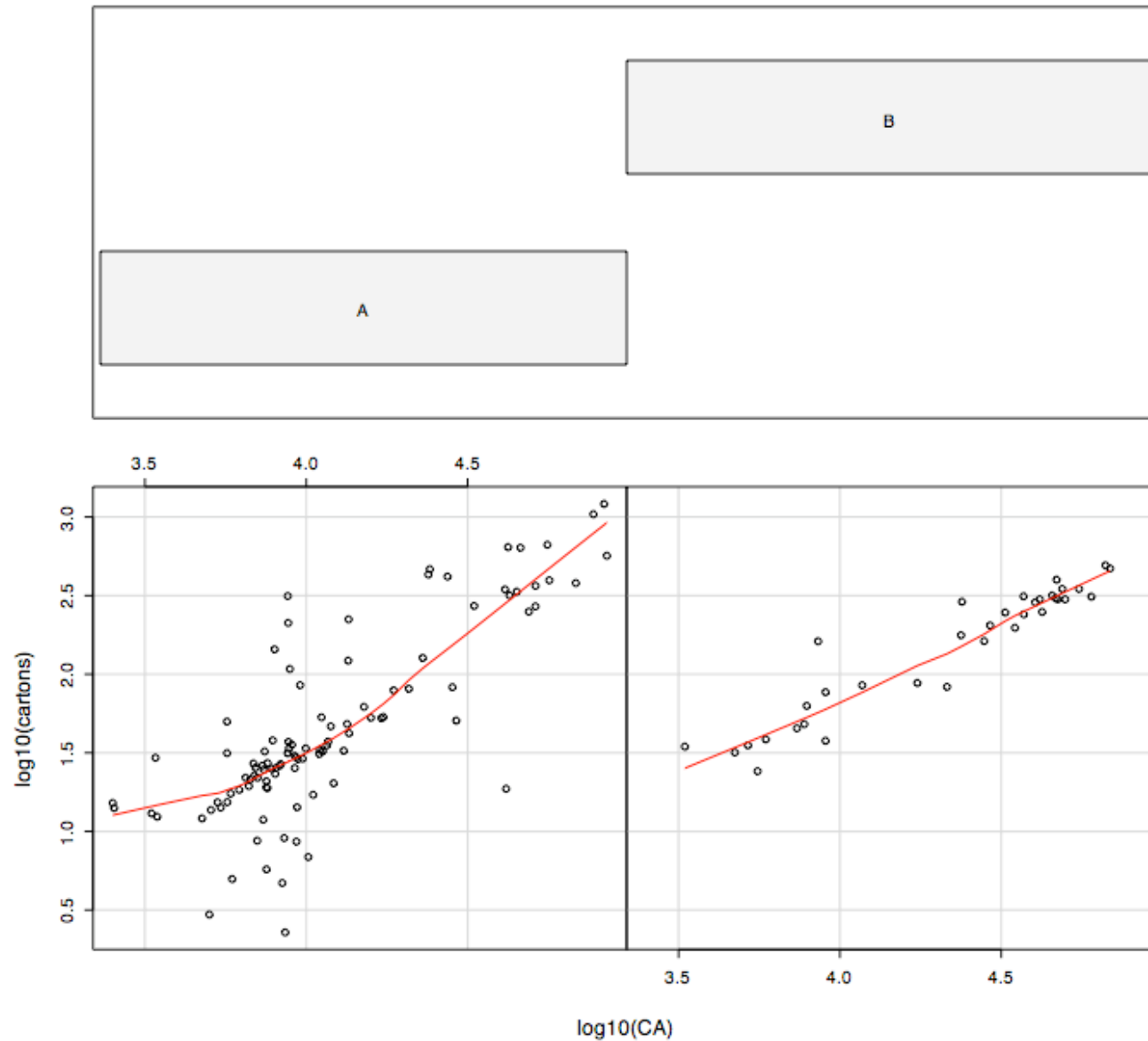
Cartons - interaction CA-categorie

Given : categorie

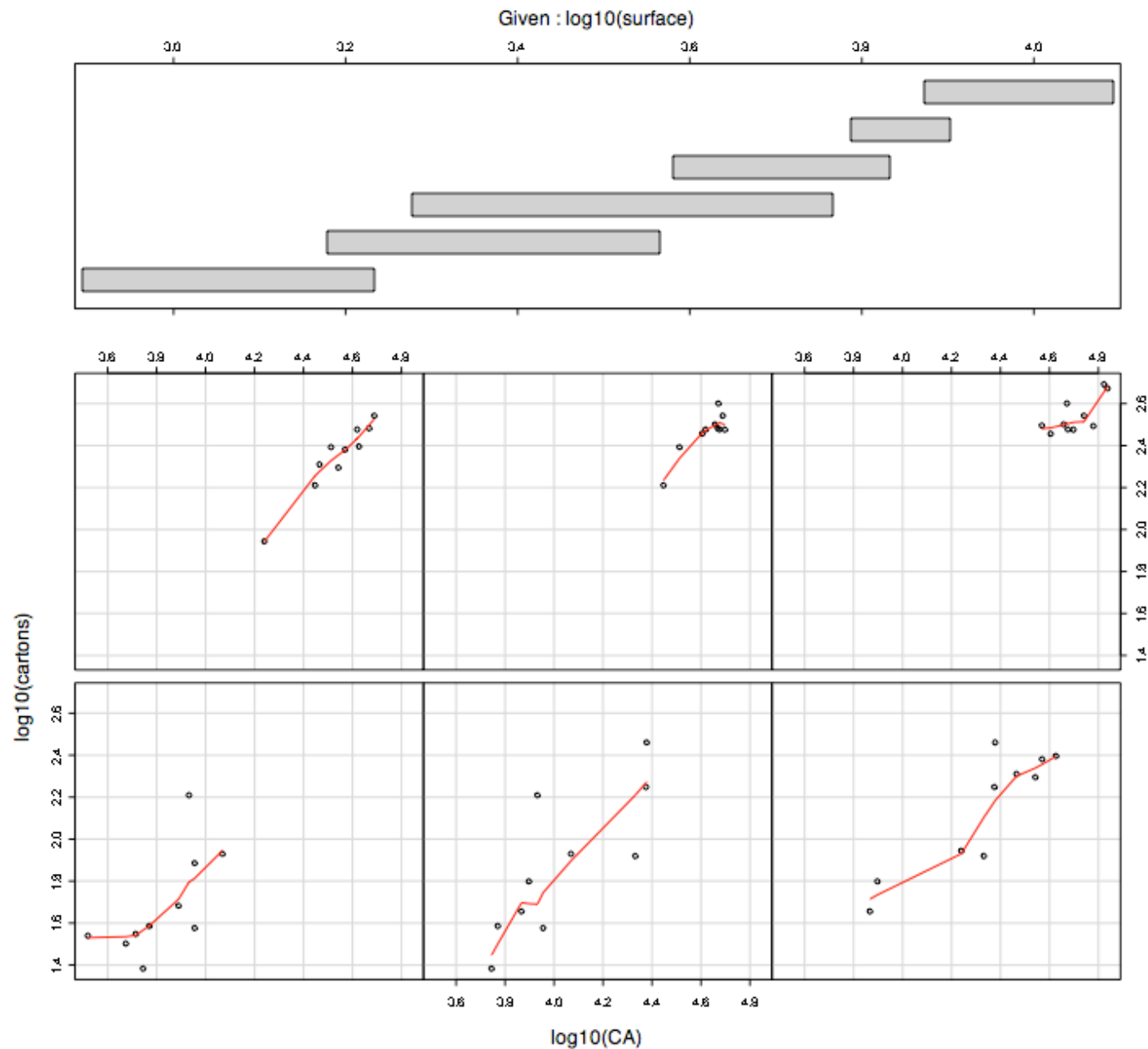


Cartons - interaction CA-prestataire

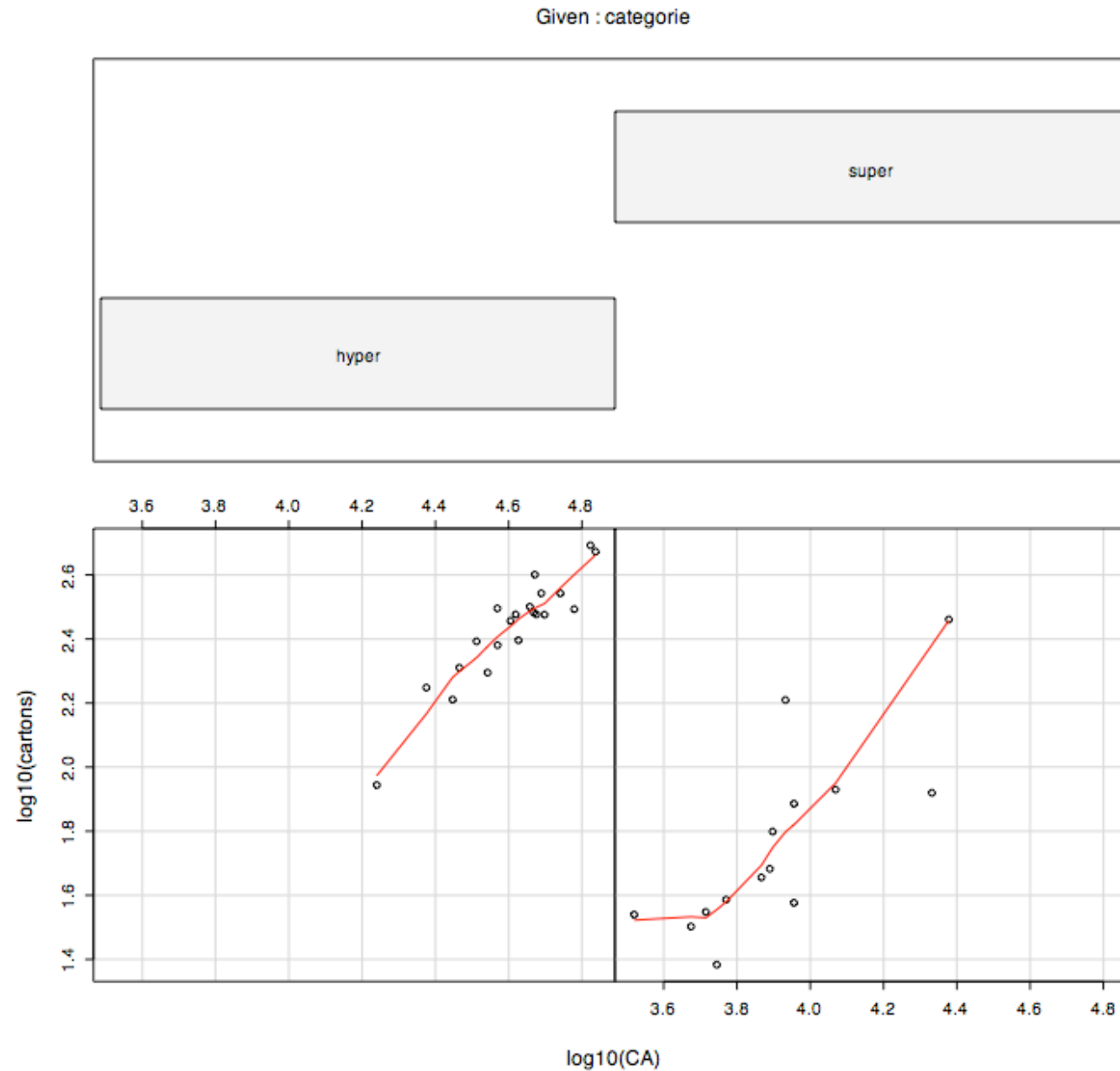
Given : prestataire

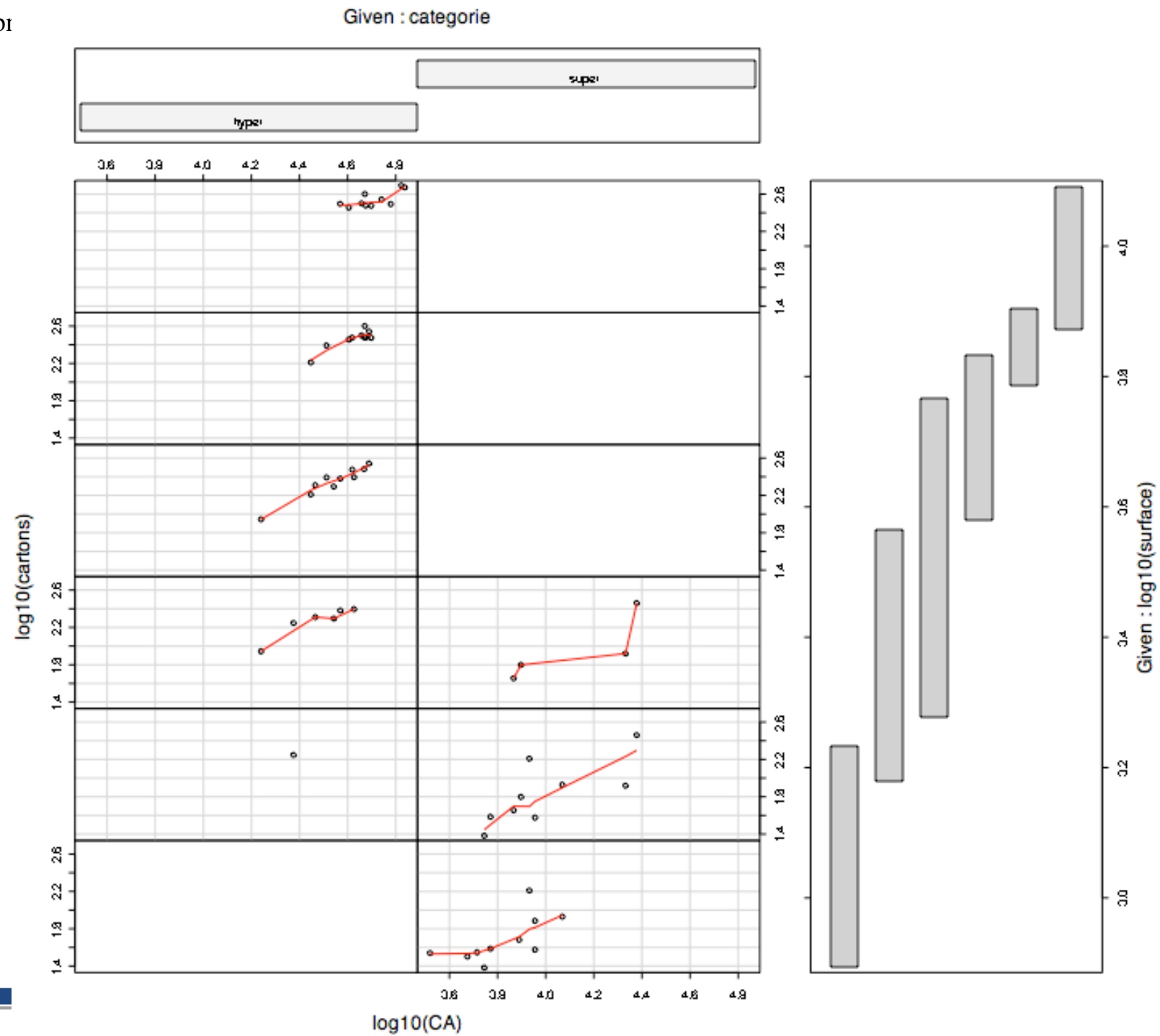


Prestataire B - interaction CA-surface



Prestataire B - interaction CA-categorie





Prestataire B - modèle initial

```
lm(formula = log10(cartons) ~ log10(CA) + log10(surface) +
    categorie,
    data = dechetsB)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.254735	-0.049060	-0.001419	0.041061	0.417644

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-2.3793	0.6175	-3.853	0.000549	***
log10(CA)	0.9259	0.2139	4.330	0.000145	***
log10(surface)	0.1416	0.2814	0.503	0.618407	
categoriesuper	0.0725	0.1121	0.647	0.522447	

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

Residual standard error: 0.1294 on 31 degrees of freedom

Multiple R-Squared: 0.9052, Adjusted R-squared: 0.896

F-statistic: 98.68 on 3 and 31 DF, p-value: 5.992e-16

Prestataire B - modèle réduit

```
lm(formula = log10(cartons) ~ log10(CA), data = dechetsB)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.24801	-0.05692	-0.01017	0.04435	0.42809

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-2.01914	0.23857	-8.464	8.85e-10 ***
log10(CA)	0.96648	0.05489	17.609	< 2e-16 ***

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

Residual standard error: 0.1264 on 33 degrees of freedom

Multiple R-Squared: 0.9038, Adjusted R-squared: 0.9009

F-statistic: 310.1 on 1 and 33 DF, p-value: < 2.2e-16

ANOVA avec R - modèle initial

- La table d'ANOVA est contenue dans le tableau donné par la commande summary.
- Commande anova :

Analysis of Variance Table

Response: log10(cartons)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
log10(CA)	1	4.9519	4.9519	295.5775	<2e-16	***
log10(surface)	1	0.0006	0.0006	0.0387	0.8454	
categorie	1	0.0070	0.0070	0.4185	0.5224	
Residuals	31	0.5194	0.0168			

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

ANOVA pour modèles emboîtés

- **Modèle complet**

$$y \approx \beta_0 + \beta_1 x_1 + \dots + \beta_q x_q + \dots + \beta_p x_p$$

- **Modèle réduit**

$$y \approx \beta_0 + \beta_1 x_1 + \dots + \beta_q x_q \text{ (avec } q < p \text{)}$$

$$H_0 : \beta_{q+1} = \dots = \beta_p = 0$$

Sous l'hypothèse $H_0 : \beta_{q+1} = \dots = \beta_p = 0$,

$$F = \frac{(SSE_R - SSE_C)/(p - q)}{SSE_C/(n - p - 1)}$$

suit une loi de Fisher $F_{p-q, n-p-1}$.

Modèles emboîtés - cas cartons

```
> mod_lB1<-
  lm(log10(cartons)~log10(CA)+log10(surface)+categorie,
    data=dechetsB)
> mod_lB2<-lm(log10(cartons)~log10(CA),data=dechetsB)

> anova(mod_lB2,mod_lB1)
```

Analysis of Variance Table

Model 1: $\log_{10}(\text{cartons}) \sim \log_{10}(\text{CA})$

Model 2: $\log_{10}(\text{cartons}) \sim \log_{10}(\text{CA}) + \log_{10}(\text{surface}) + \text{categorie}$

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	33	0.52701				
2	31	0.51935	2	0.00766	0.2286	0.797

Choix automatique de modèles

- Approche forward
 - On part d'un modèle simple et on ajoute des termes tant qu'ils sont significatifs
- Approche backward
 - On part d'un modèle complet, et on enlève des termes tant que leur suppression ne donne pas une détérioration significative
 - Approche stepwise : les deux en même temps.

Critères de choix

- Test : statistique F (cf. modèles emboîtés)
- Numérique :
 - $\text{MSE} = \hat{\sigma}^2$ (ou $R^2\text{-adj}$)
 - Critères de type MSE pénalisé :
$$\text{AIC} = n \log(\hat{\sigma}^2) + k p$$
$$k = 2 : \text{AIC d'Akaike originel}$$
$$k = \log(n) : \text{BIC}$$

Exemple pour le cas carton

- Ajouts de prédicteurs par le test F
 - on ajoute successivement $\log_{10}(\text{CA})$,
prestataire, $\text{prestataire} * \log_{10}(\text{CA})$.
- Suppression de prédicteurs par le test F
 - Modèle complet linéaire sans interaction : on enlève categorie puis $\log_{10}(\text{surface})$.
 - Modèle complet avec interaction : on enlève $\log_{10}(\text{surface}) * \text{prestataire}$,
 $\log_{10}(\text{CA}) * \text{prestataire}$, $\log_{10}(\text{CA}) * \text{categorie}$

Stepwise avec AIC et BIC

- AIC en partant de rien :
 - + $\log_{10}(\text{CA})$ + prestataire
 - + $\log_{10}(\text{CA}) * \text{prestataire}$ + categorie
 - + $\text{prestataire} * \text{categorie} - \log_{10}(\text{CA}) * \text{prestataire}$
 - + $\log_{10}(\text{surface})$
- BIC en partant de rien :
 - + $\log_{10}(\text{CA})$ + prestataire
 - + $\log_{10}(\text{CA}) * \text{prestataire}$

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 $y=a+bx+\varepsilon$

$$h_{ii} = \frac{\text{Var}(x) + (x_i - \bar{x})^2}{n \text{Var}(x)}$$

- $T_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma} \sqrt{1-h_{ii}}}$ 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 : A ou B
 - Catégorie : super ou hyper
- Rappel
 - prendre les log (en base 10)
 - séparer les données selon le prestataire
 - seul CA est conservé dans le modèle

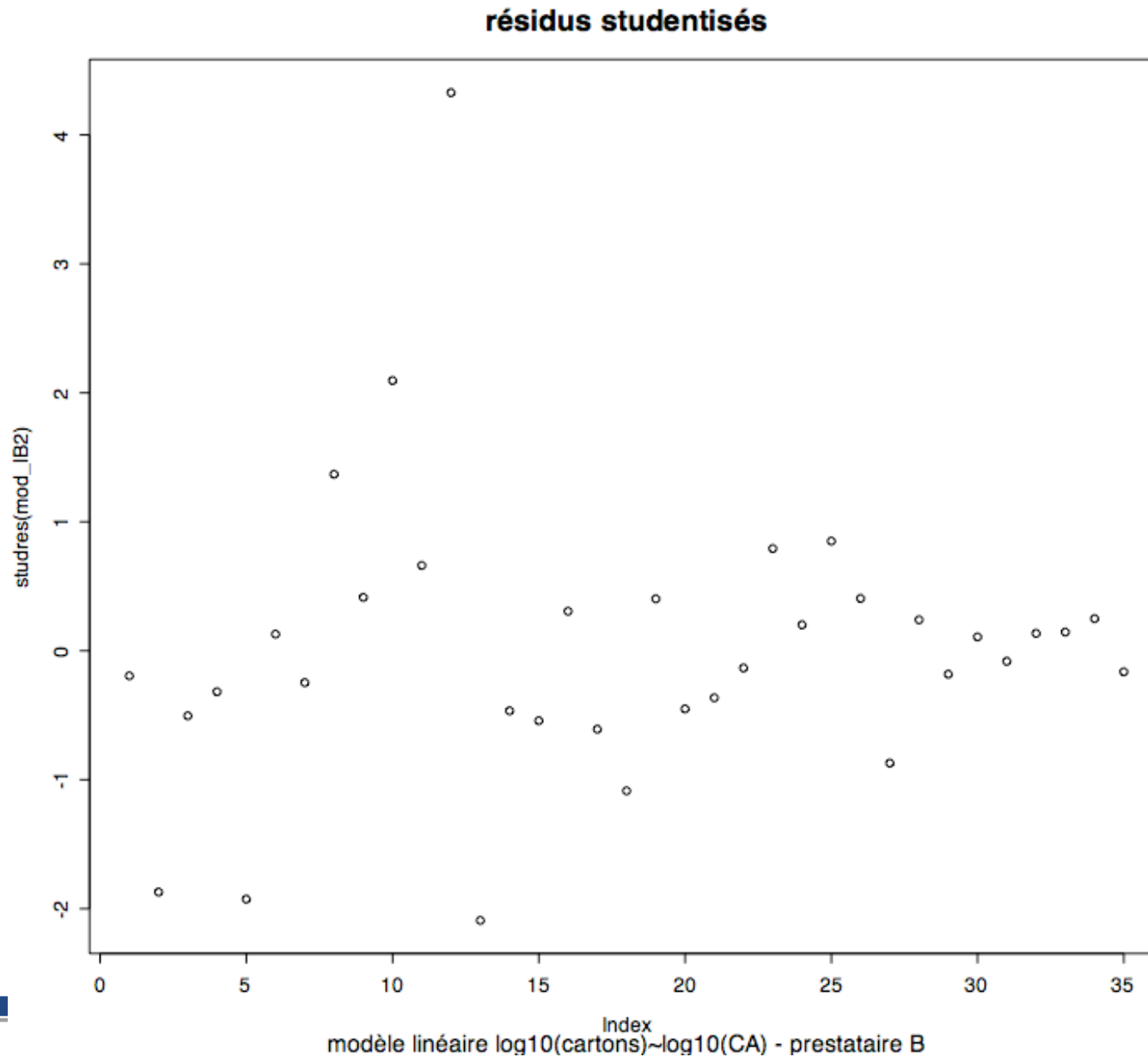
Résidus studentisés T_i^*

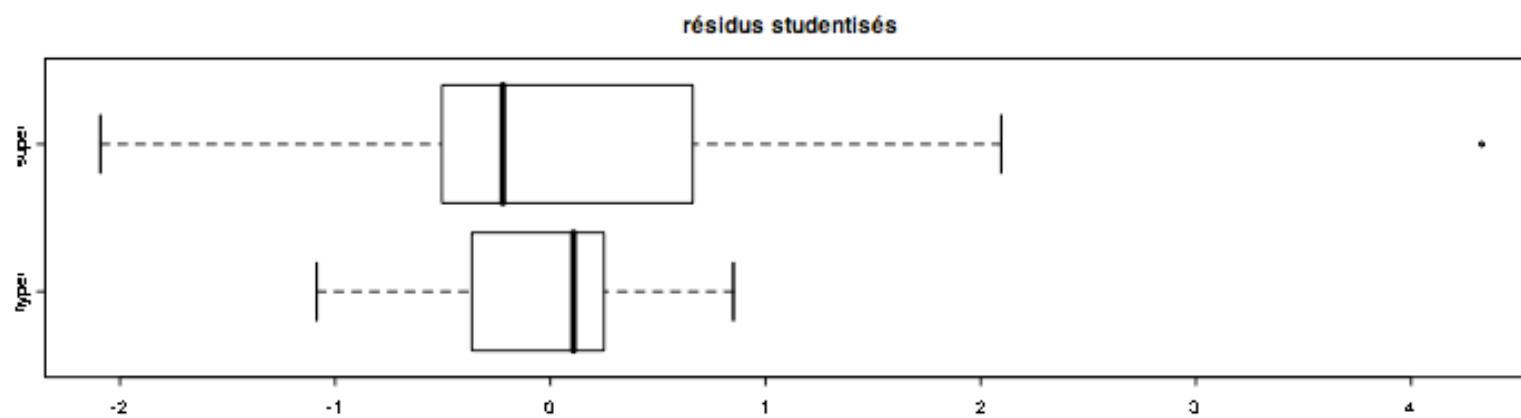
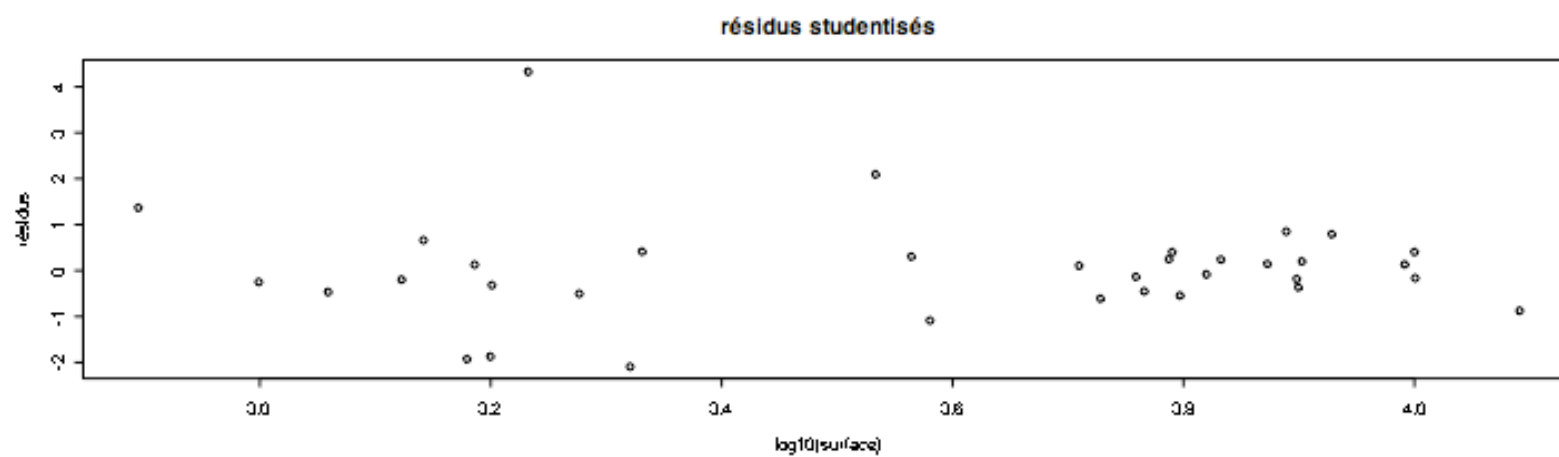
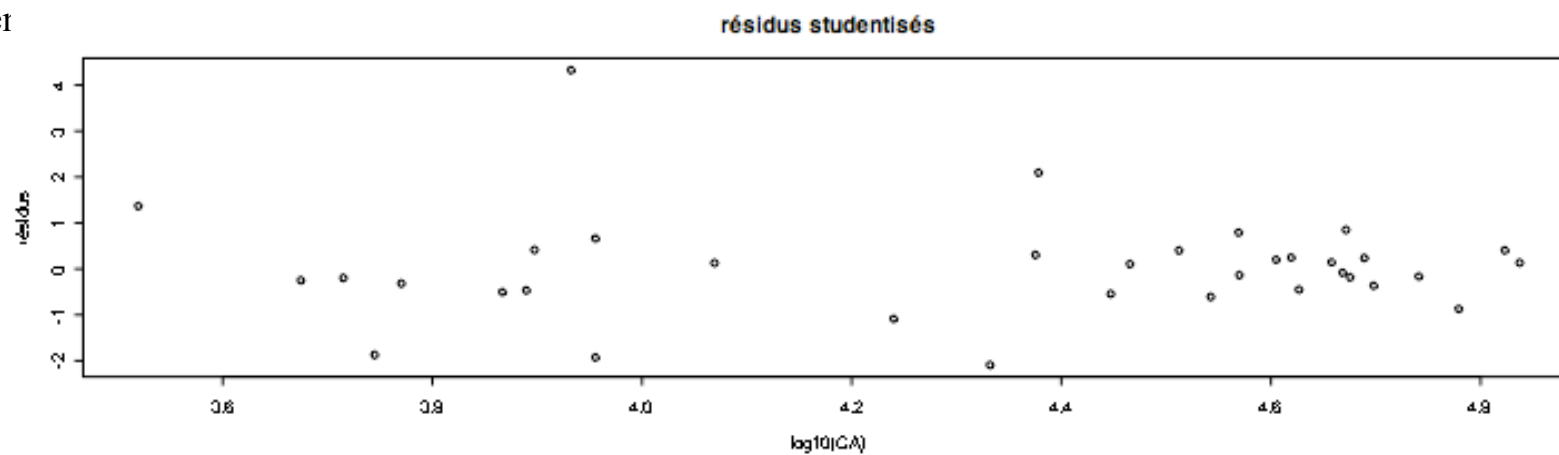
- Définition analogue à celle de T_i :

$$T_i^* = \frac{\hat{\varepsilon}_i}{\sigma_{(i)} \sqrt{1 - h_{ii}}}$$

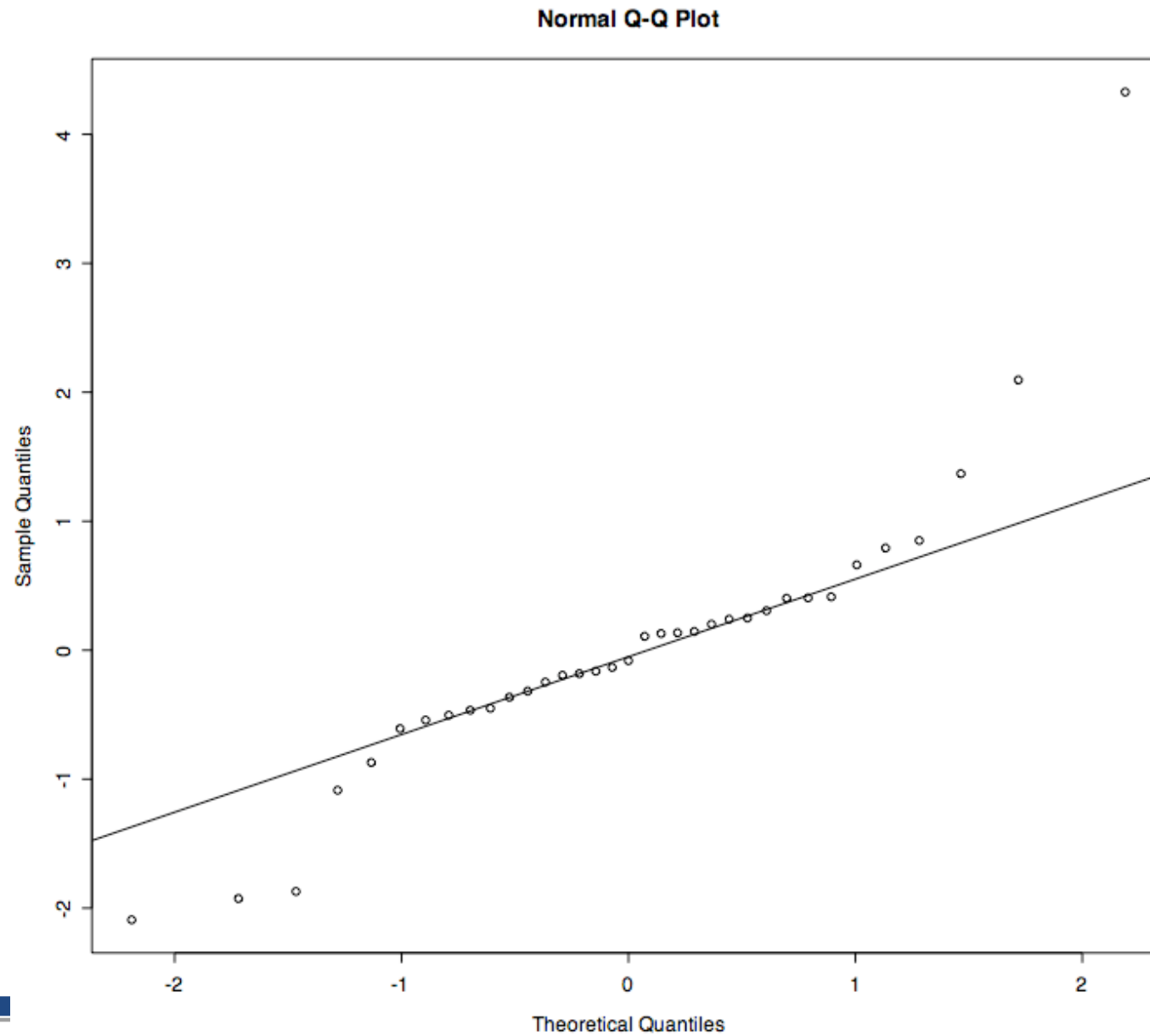
- Loi de T_i^* connue (Student t_{n-p-2}) :
 - Tracé séquentiel pour détecter des résidus trop forts
 - Test d'adéquation à la loi normale $N(0,1)$ (très peu différente de t_{n-p-2})

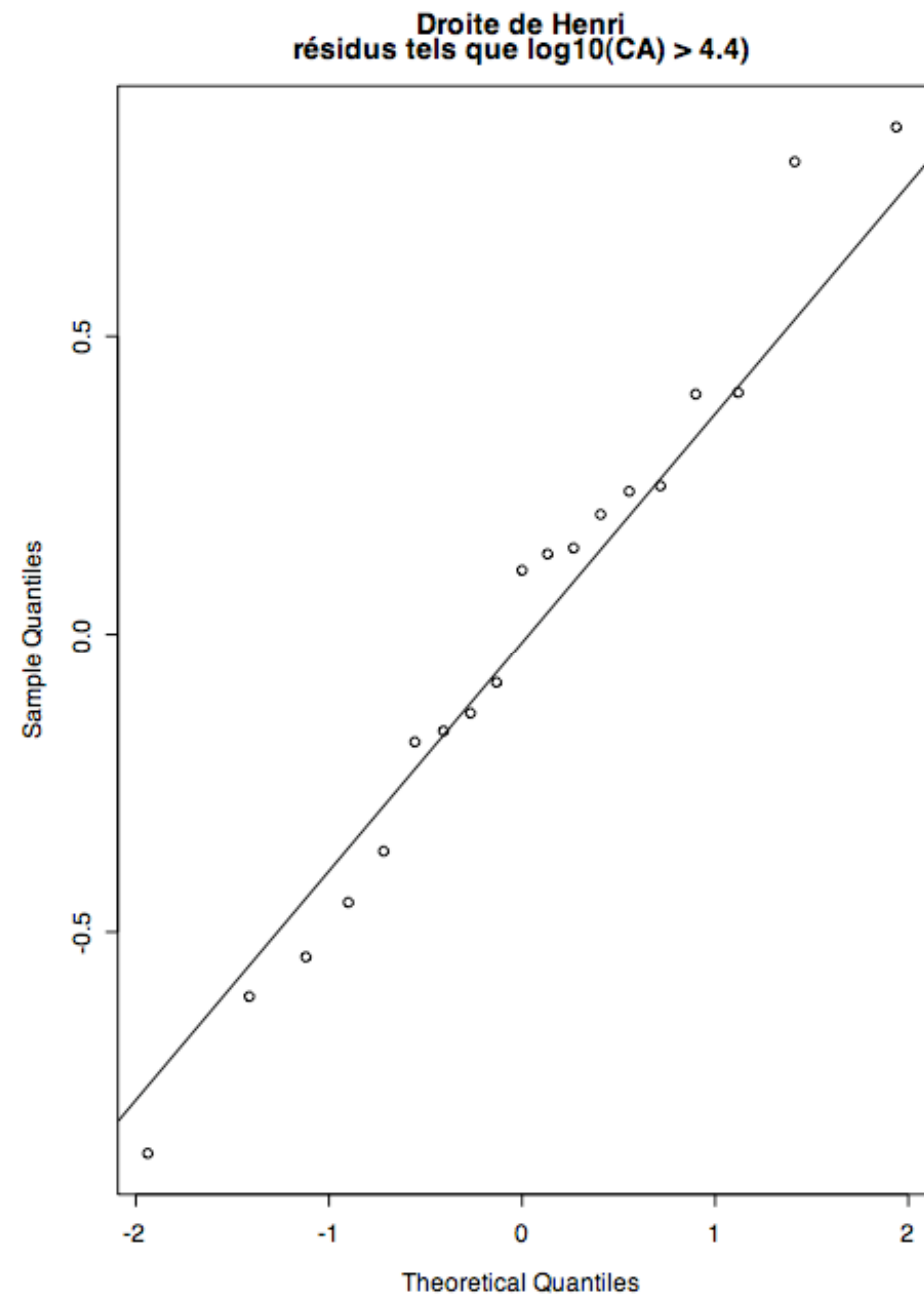
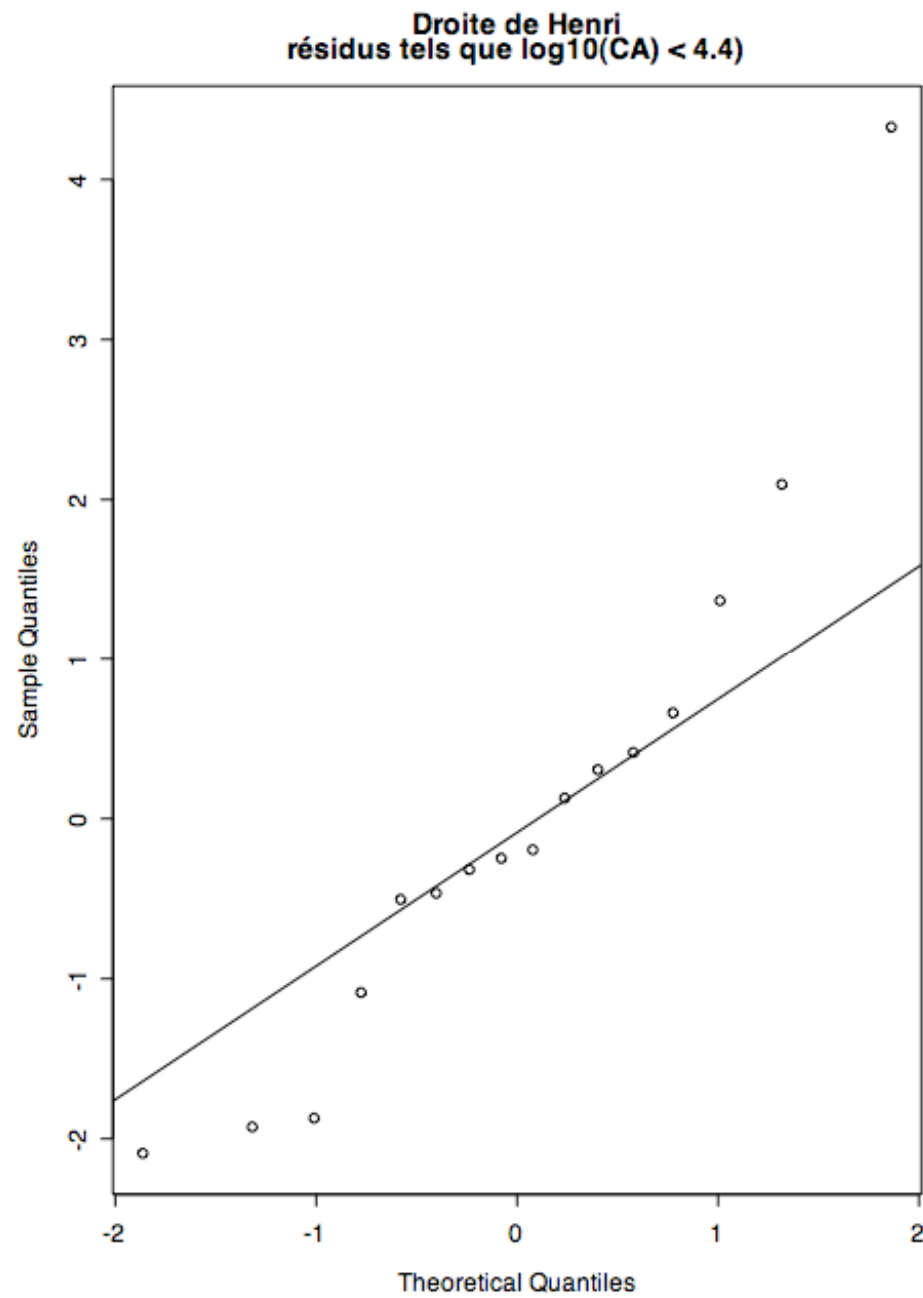
Retour sur l'exemple « cartons »





Droite de Henri





Prestataire B - modèle sans l'obs n°12

```
lm(formula = log10(cartons) ~ log10(CA), data = dechetsB2)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.235126	-0.042072	0.007826	0.044136	0.259799

Coefficients:

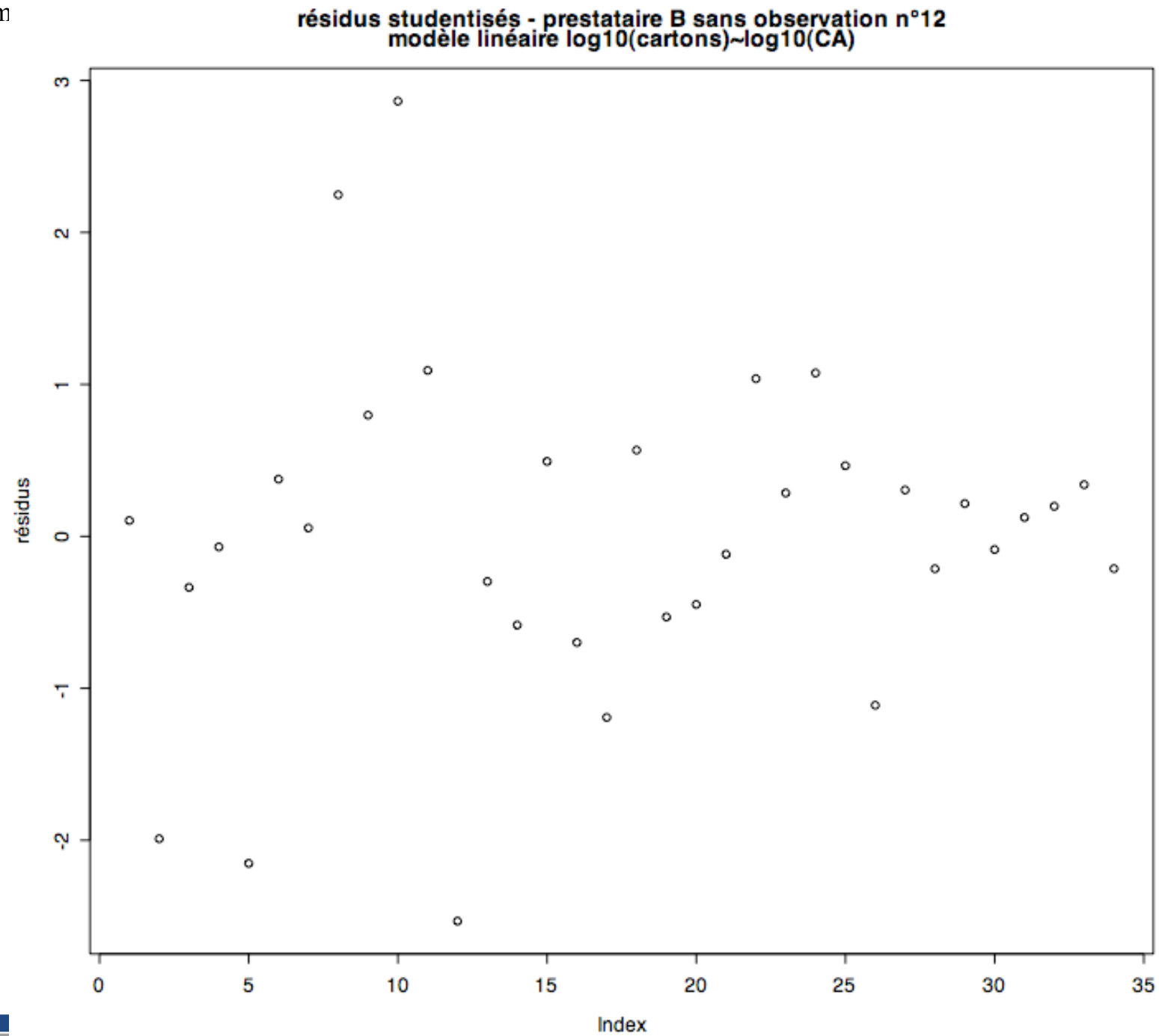
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-2.17938	0.19594	-11.12	1.58e-12 ***
log10(CA)	1.00050	0.04496	22.25	< 2e-16 ***

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

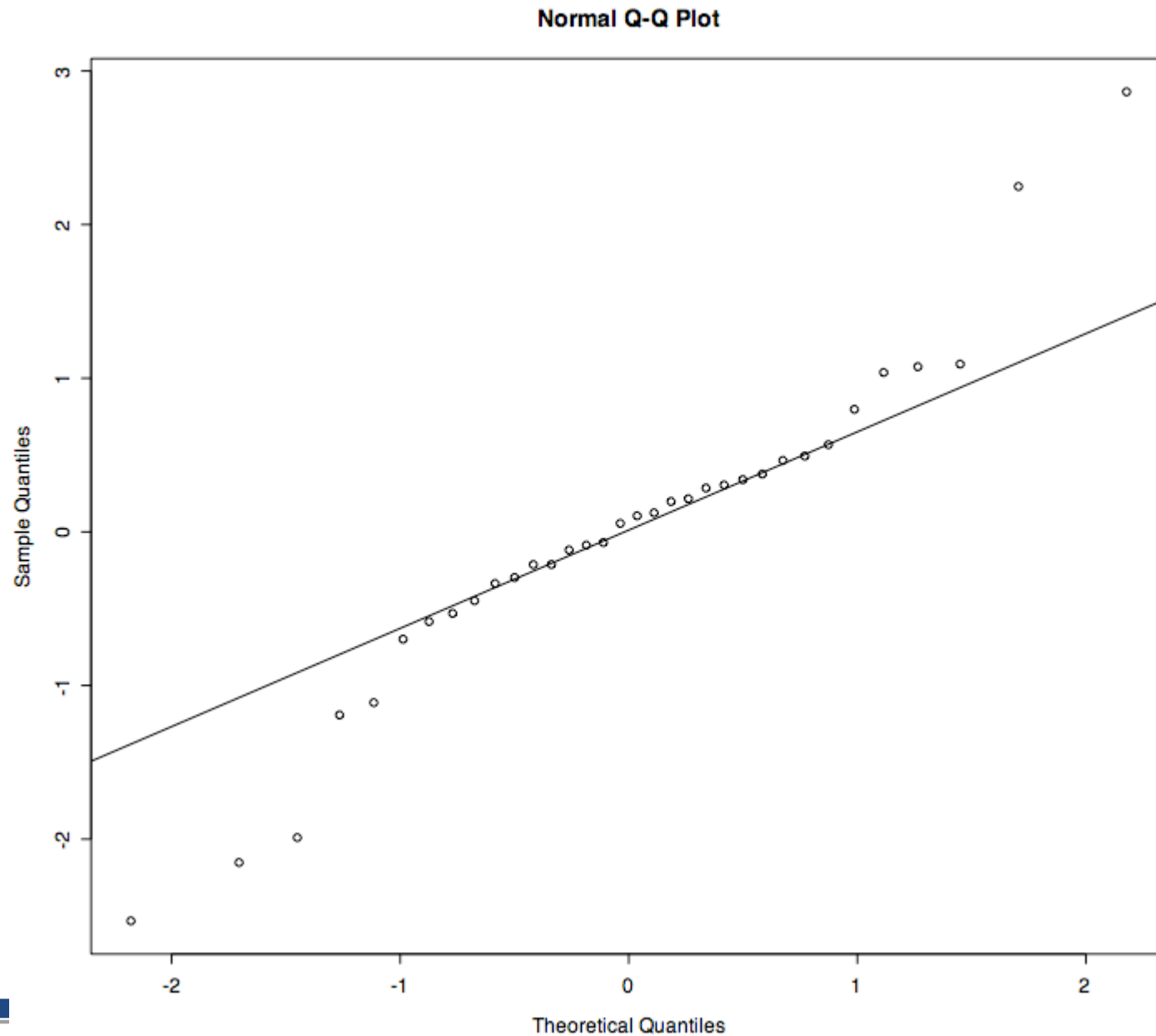
Residual standard error: 0.1019 on 32 degrees of freedom

Multiple R-Squared: 0.9393, Adjusted R-squared: 0.9374

F-statistic: 495.2 on 1 and 32 DF, p-value: < 2.2e-16



Données sans observation n°12



Données sans observation n°12

