

Cours n°2 – rappels régression

L. Carraro

Option MAFQ

Plan

- Rappel modèle linéaire
- Exemple : le cas « cartons »
- Choix de modèles : manuel, semi-automatique, automatique
- Résidus et validation
- Prévisions

Le modèle linéaire

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

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

$\varepsilon_1, \dots, \varepsilon_n$ sont des réalisations de v.a.

$E_1, \dots, E_n$ indépendantes et de loi $N(0, \sigma^2)$

Estimation et propriétés de $\hat{\beta}$

- Equation normale : $X'X \hat{\beta} = X'Y$
- $\hat{\beta}$ est de loi normale.
- $E(\hat{\beta}) = \beta$
- $\text{Cov}(\hat{\beta}) = \sigma^2 (X'X)^{-1}$
- $\hat{\beta}$ est BLUE
- A un facteur près, $\text{cov}(\hat{\beta})$ ne dépend que du plan d'expériences, pas des résultats.

Exemple

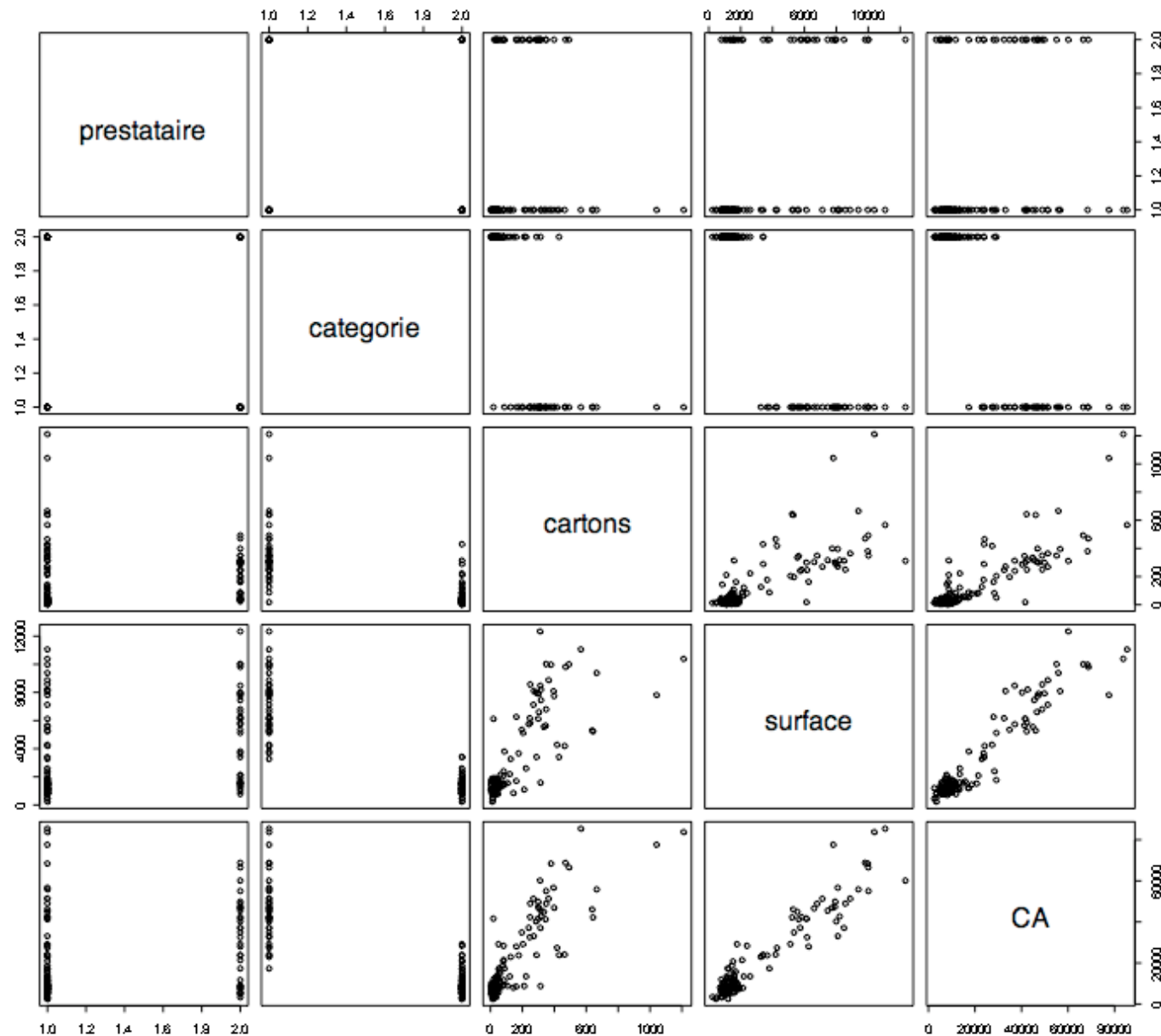
Récupération de cartons dans la grande distribution

- Réponse : tonnage de cartons annuel
- Prédicteurs :
 - CA HT
 - surface commerciale
 - prestataire (A ou B)
 - catégorie (super ou hyper)
- 137 observations

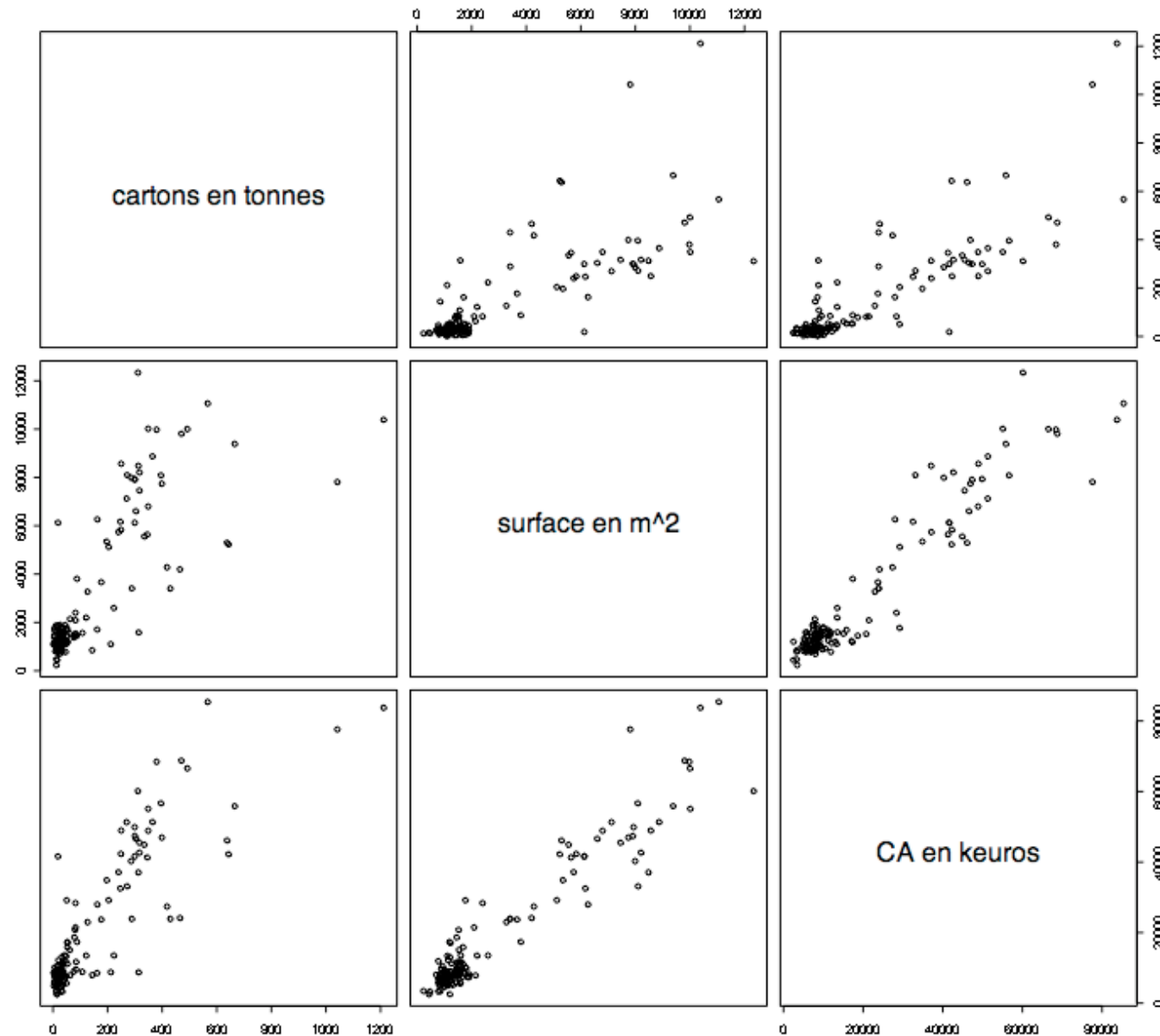
Questions du client

- Peut-on prévoir le tonnage cartons produit ?
 - Si oui, avec quelle précision ?
- Les deux prestataires ont-ils des résultats analogues ?
 - L'info tonnage vient des prestataires.
 - Approche « toutes choses égales par ailleurs ».

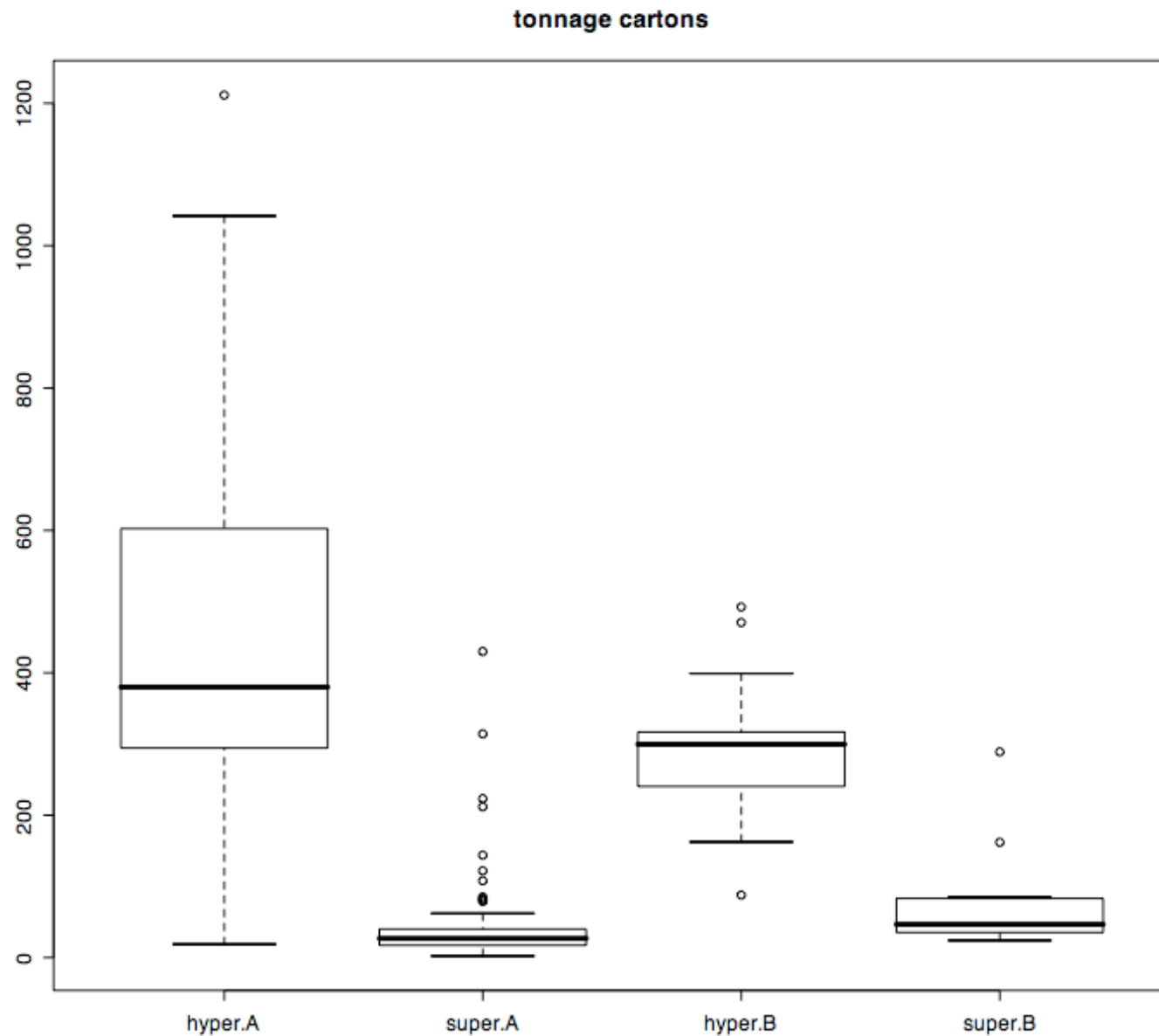
plot(dechets)



`pairs(cartons~surface+CA, data=dechets)`




```
boxplot(cartons~categorie+prestataire, data=dechets)
```



Premières observations

- CA et Surface semblent influentes
- Catégorie et Prestataire aussi
- Questions :
 - Faut-il considérer toutes les variables (on observe une corrélation entre CA et Surface) ?
 - Modélisation de quel type ?

Exemple

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

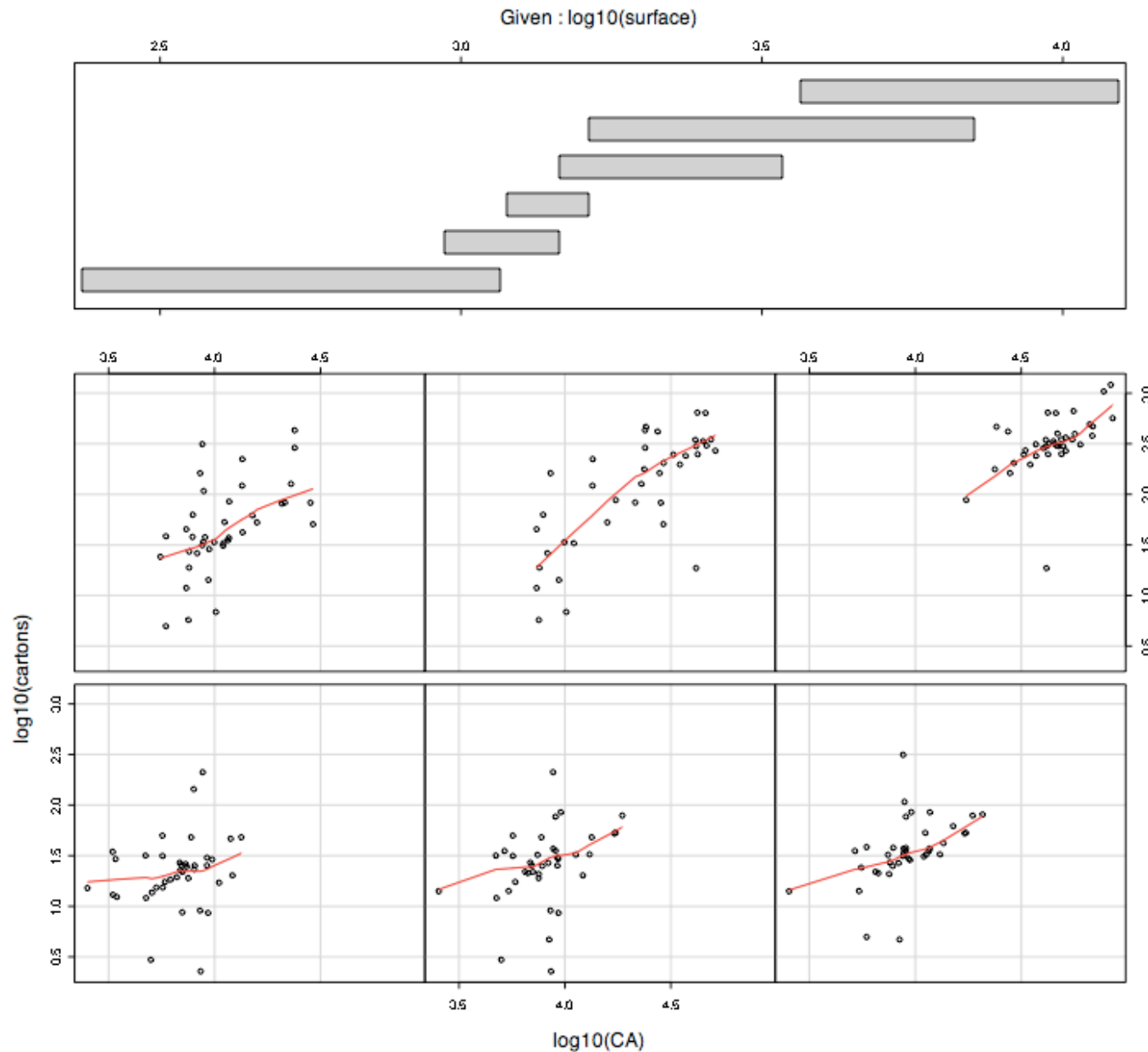
Exemple

changement de prédicteurs

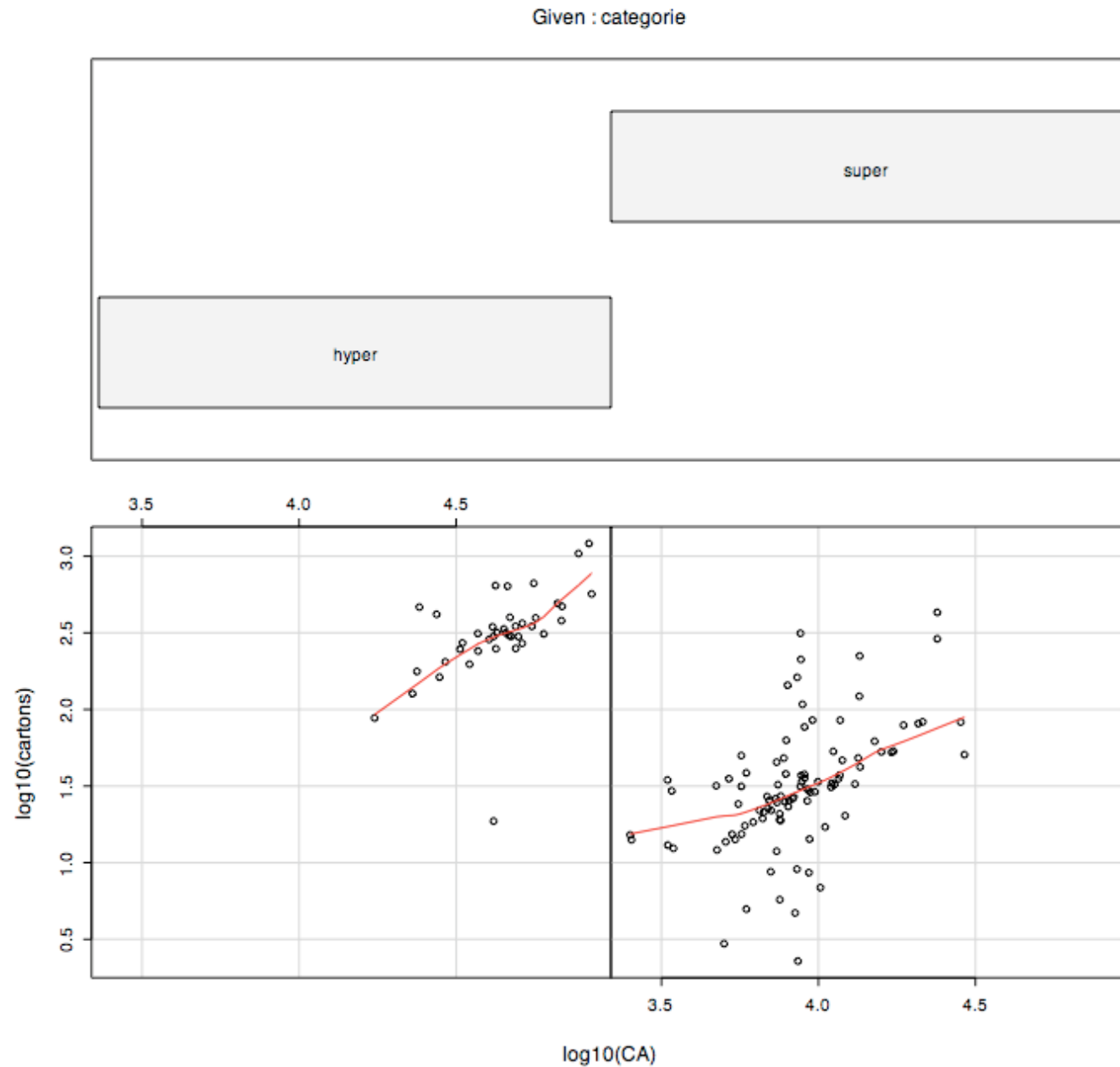
surface devient surface/CA

	$\log_{10}(\text{CA})$	$\log_{10}(\text{surface/CA})$	prestataire	catégorie
$\log_{10}(\text{CA})$	1.0000000	-0.2035992	0.3016459	0.86111312
$\log_{10}(\text{surface/CA})$	-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

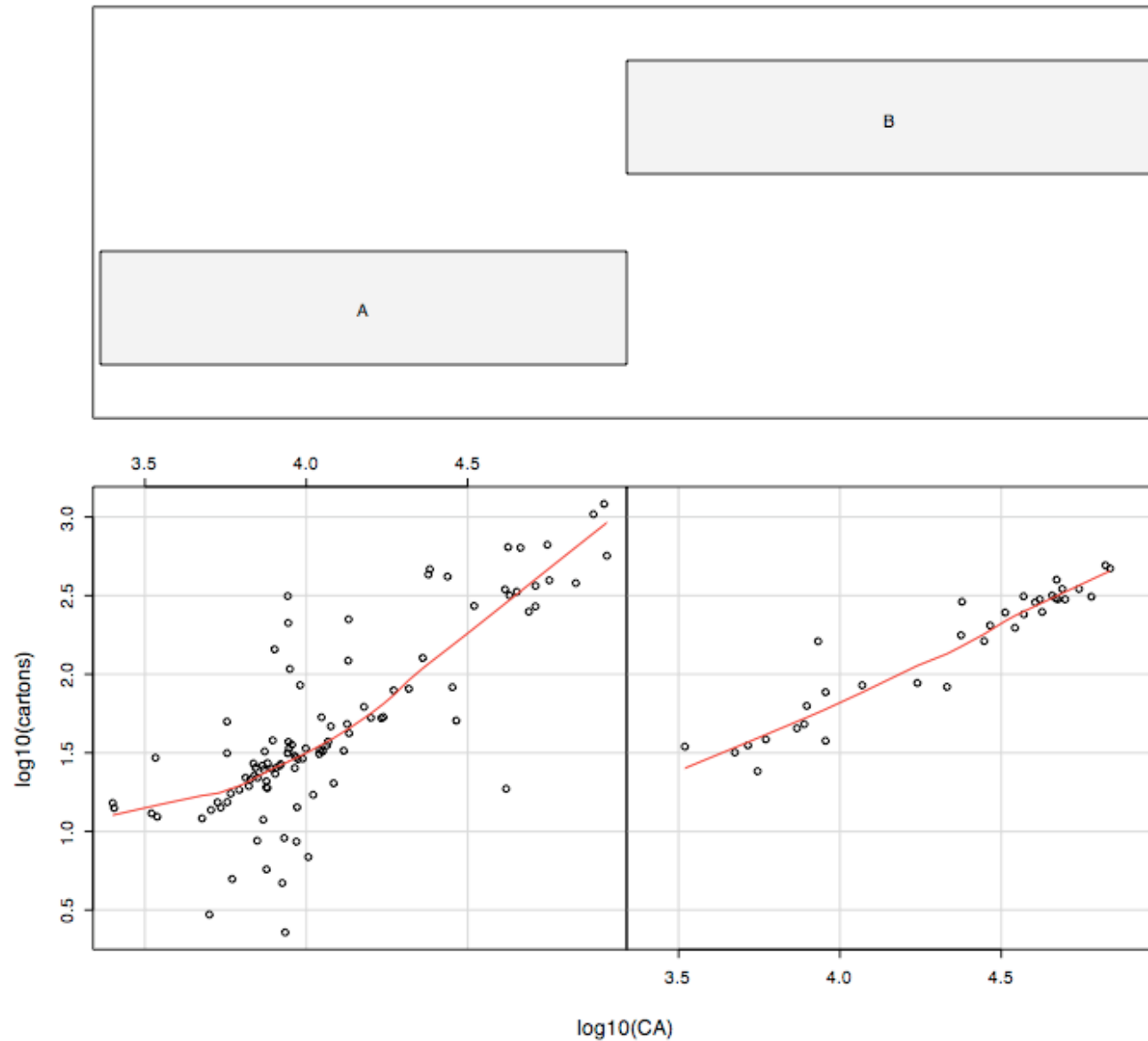


Cartons - interaction CA-catégorie

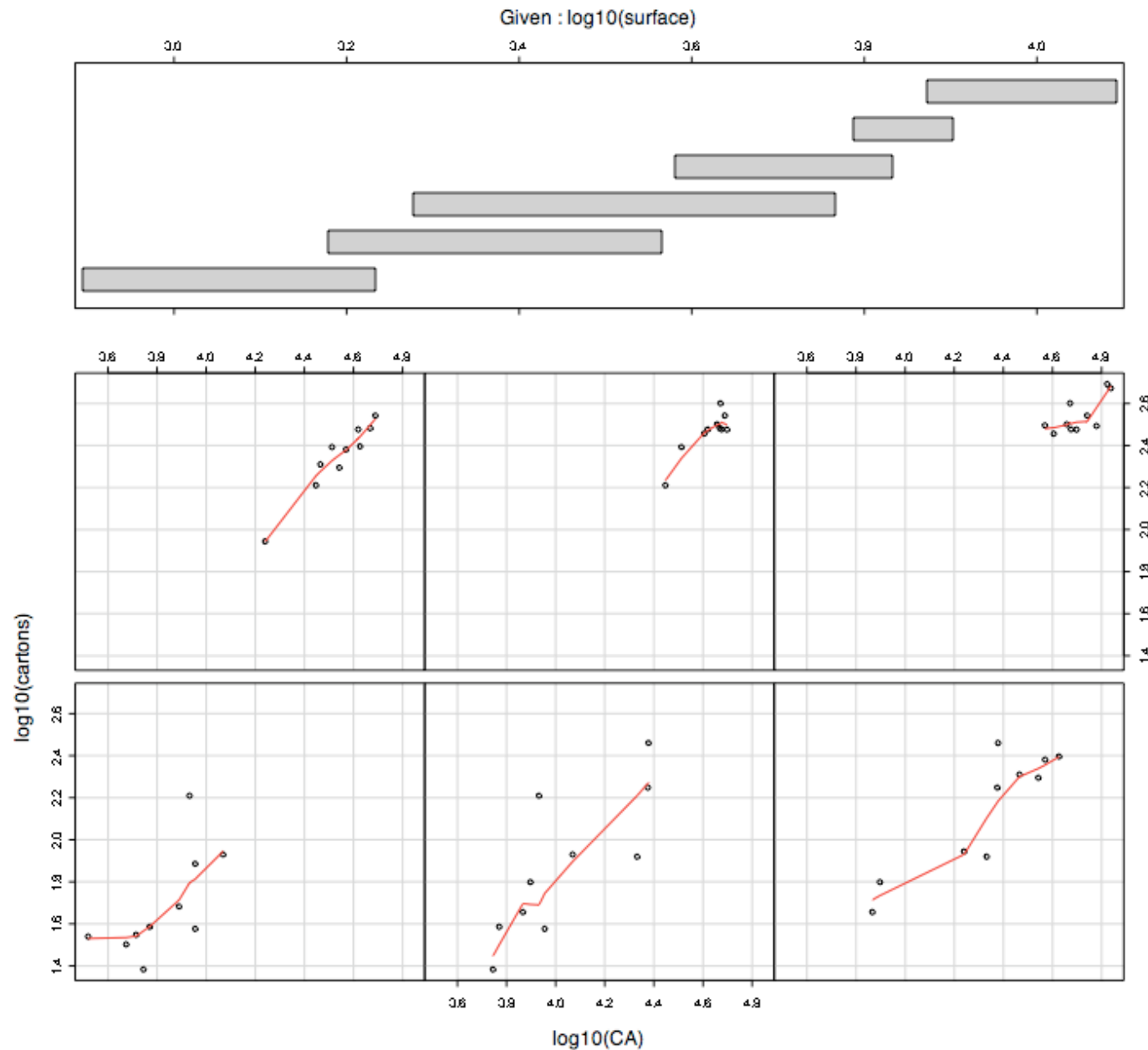


Cartons - interaction CA-prestataire

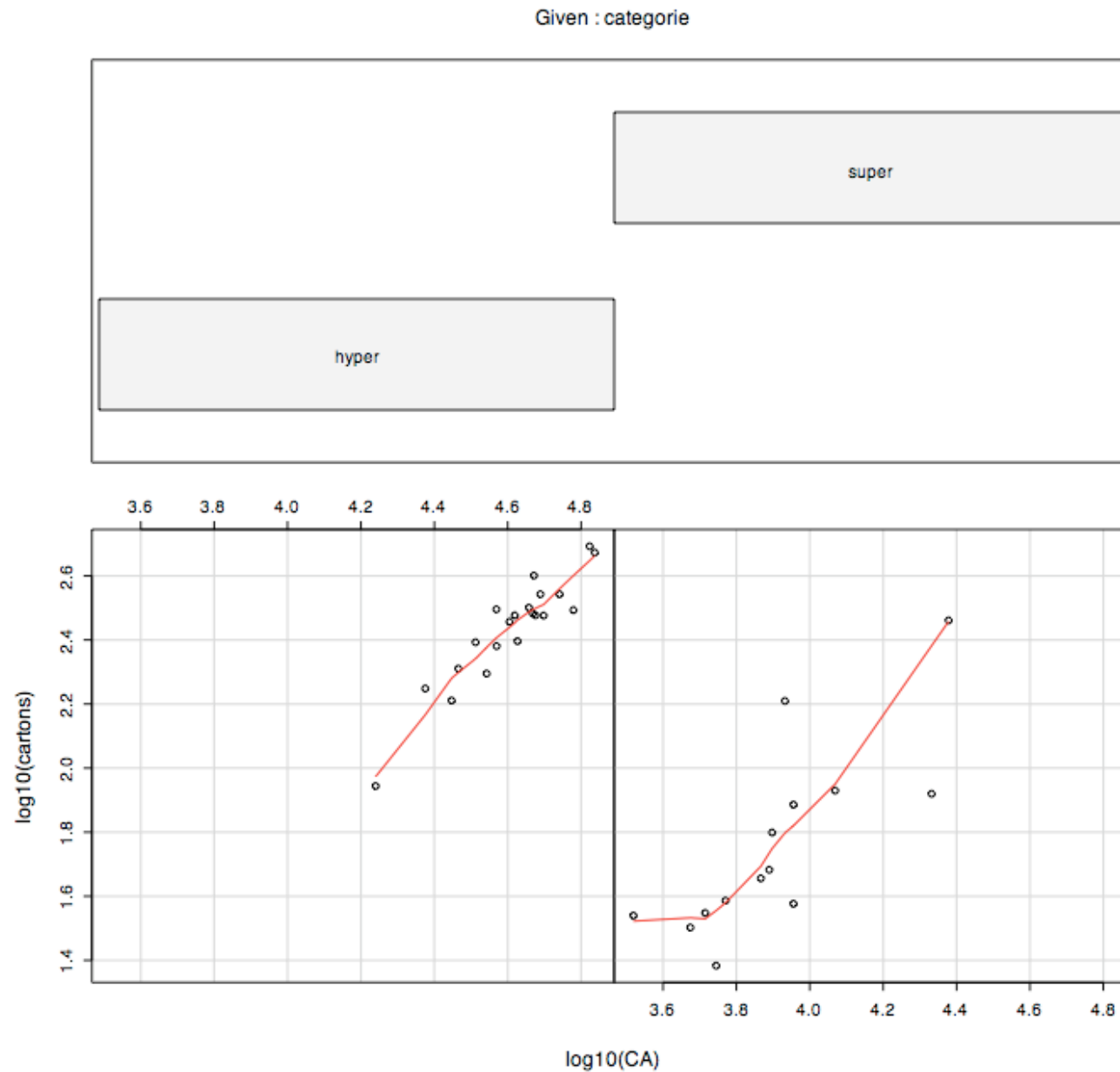
Given : prestataire

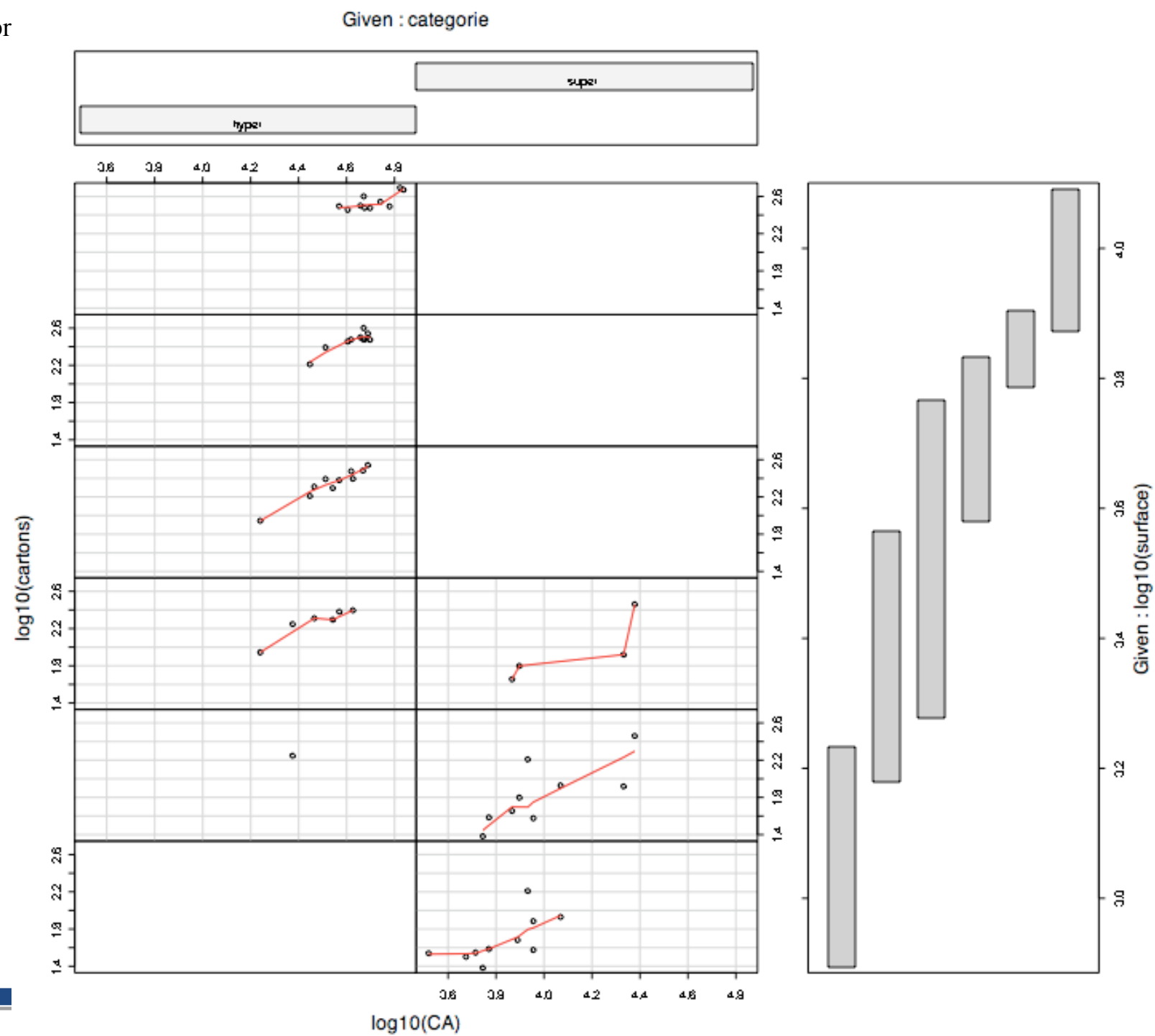


Prestataire B - interaction CA-surface



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

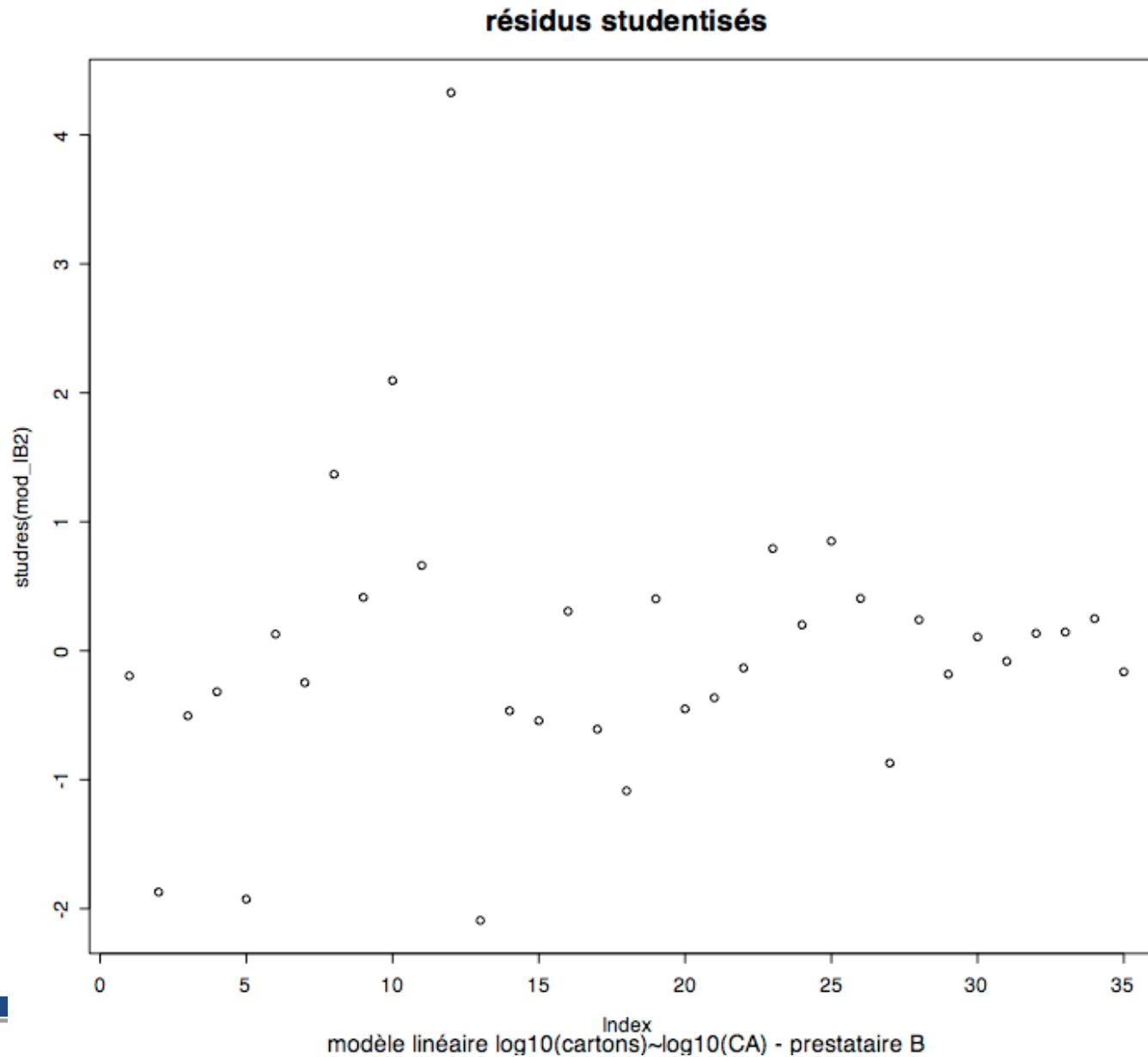
Validation - résidus

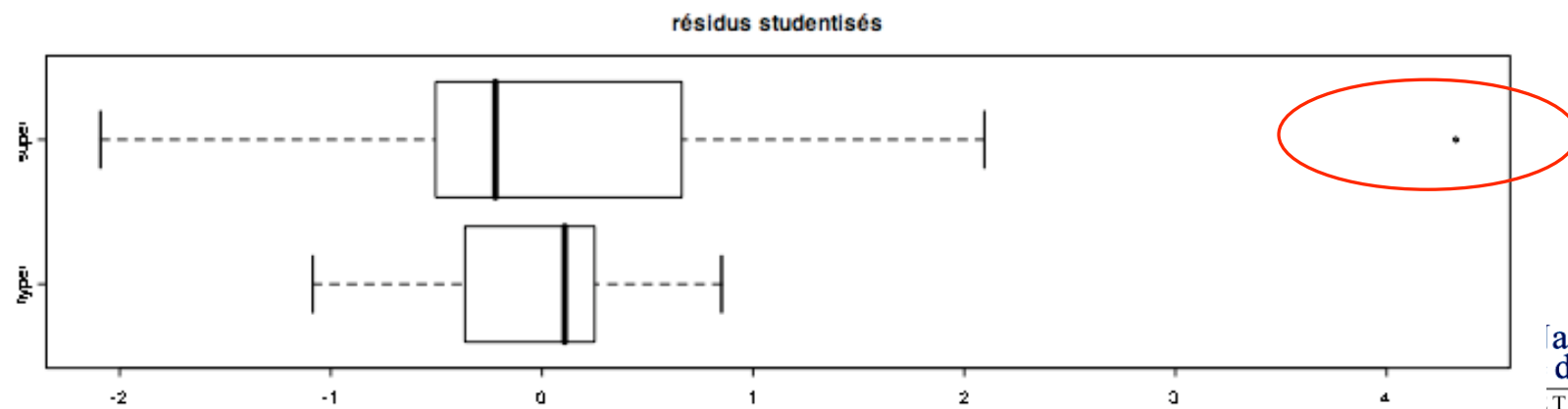
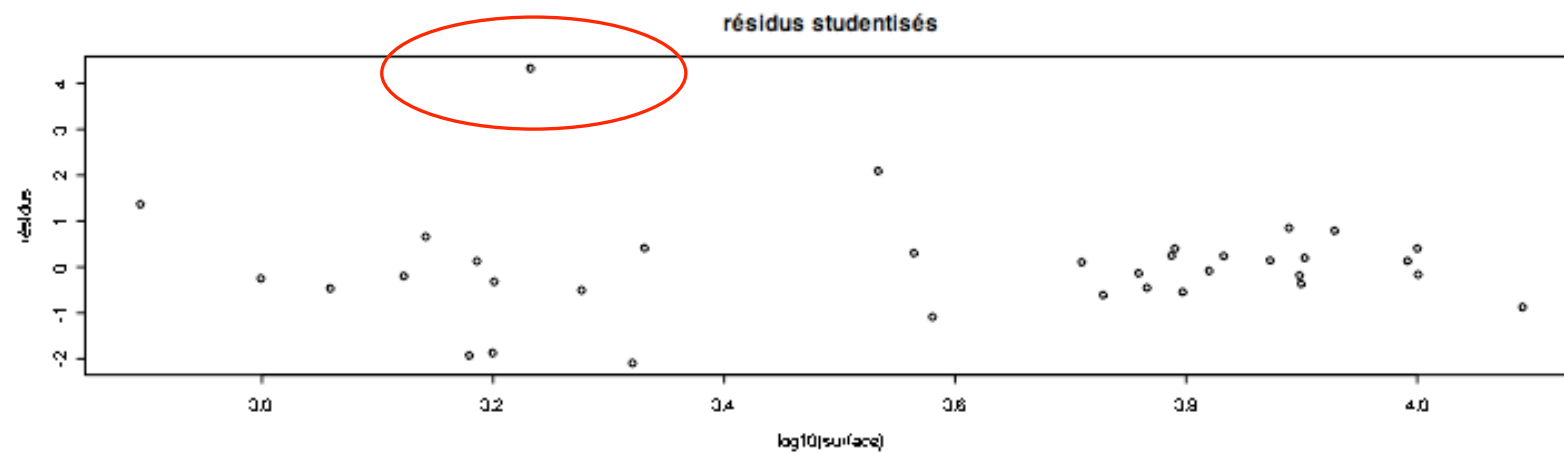
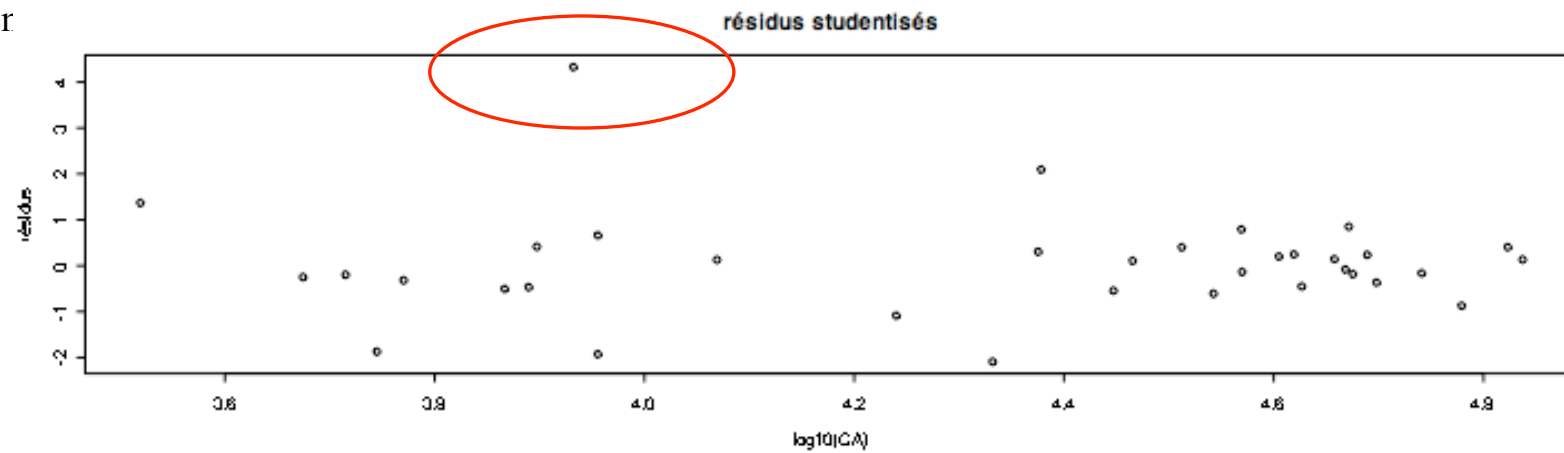
- Validation basée sur les résidus estimés $\hat{\varepsilon}$
- Tracés des résidus :
 - séquentiels
 - contre la réponse estimée
 - contre les prédicteurs
- Mais variance non constante !!

Transformations des résidus

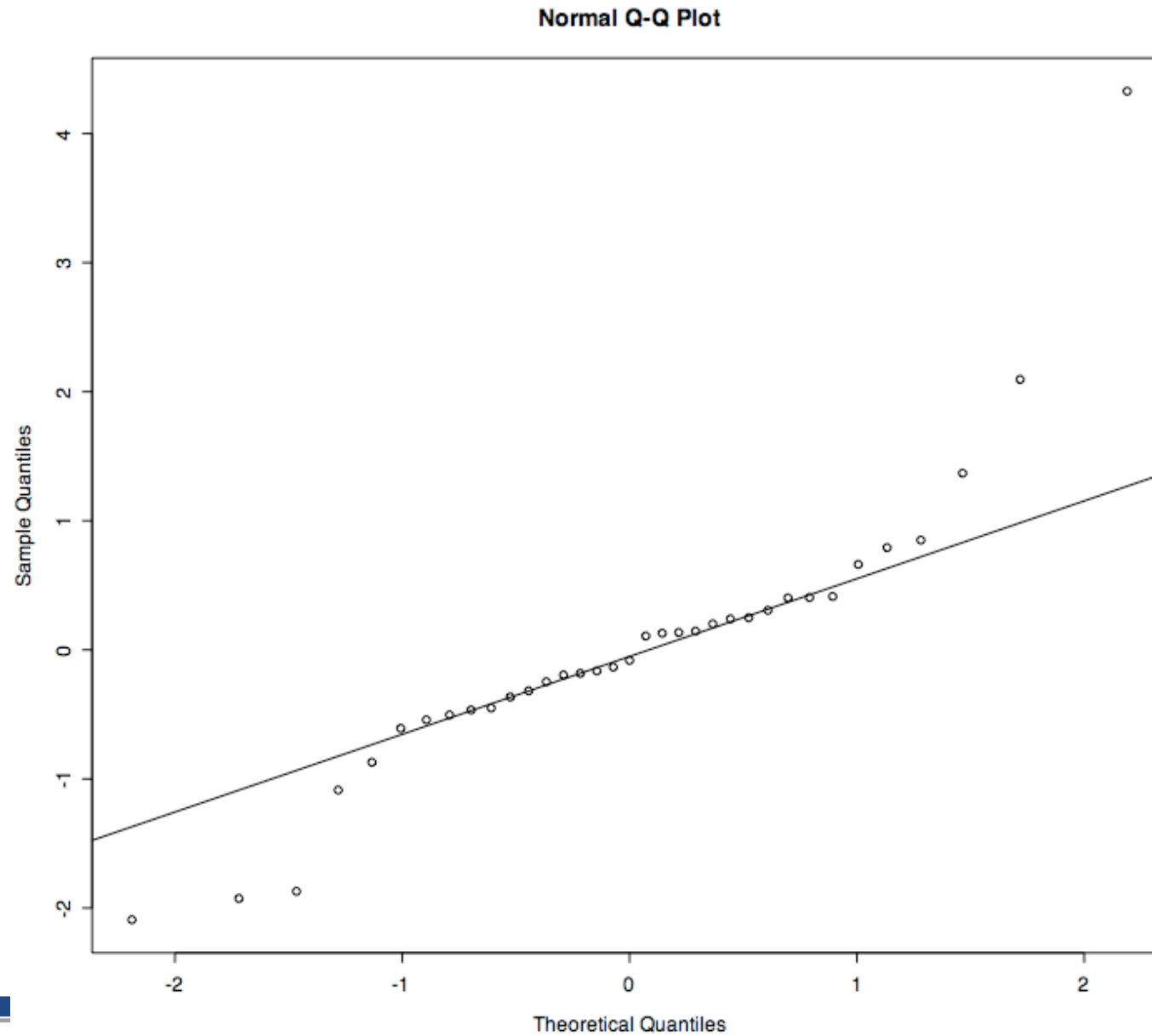
- $T_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma} \sqrt{1-h_{ii}}}$ est le i-ème résidu **standardisé**
- $T_i^* = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(i)} \sqrt{1-h_{ii}}}$ est le i-ème résidu **studentisé**
 - T_i^* est de loi de Student t_{n-2}
 - T_i^* est aussi le résidu par **validation croisée**

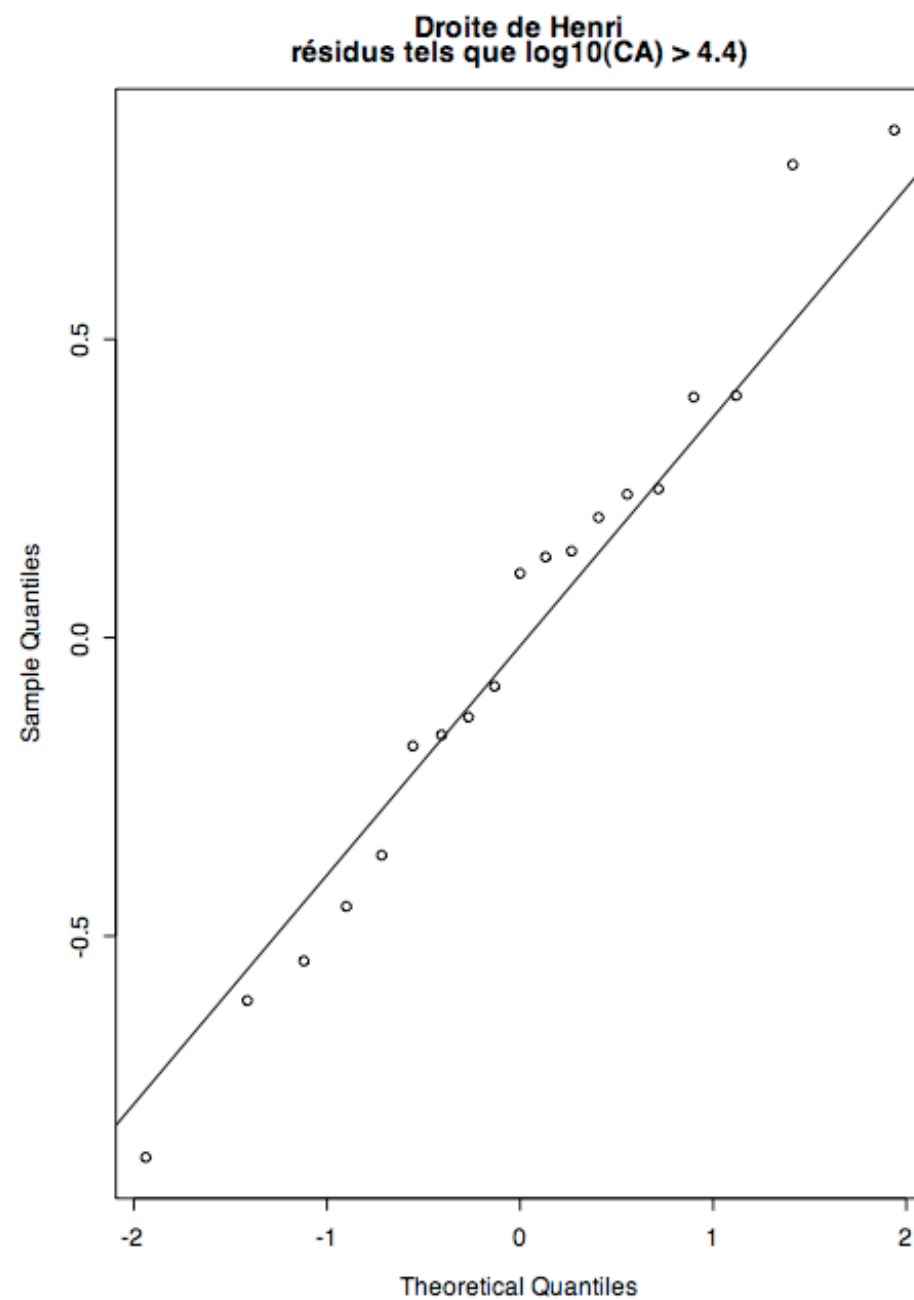
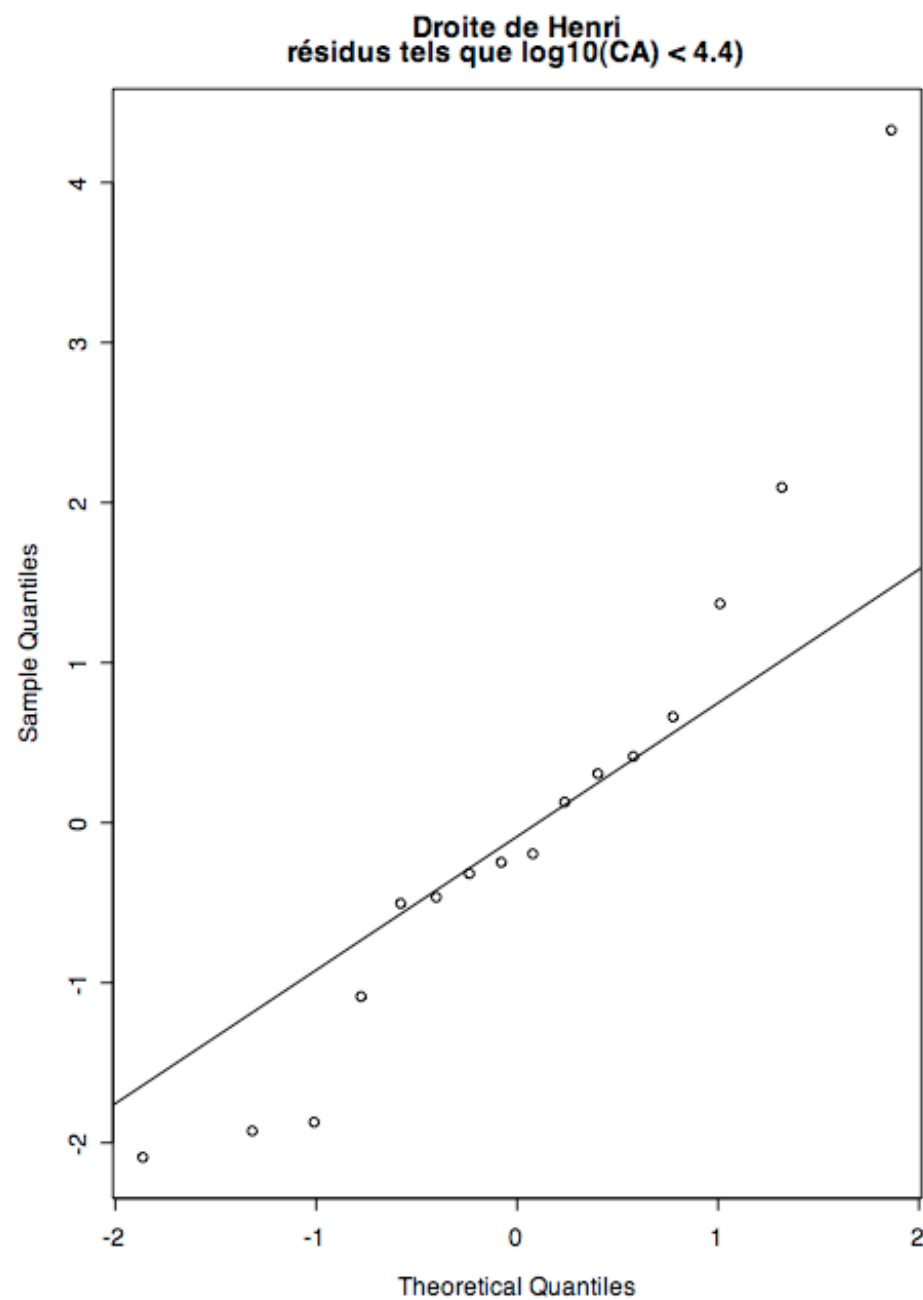
Régression pour le prestataire B





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:

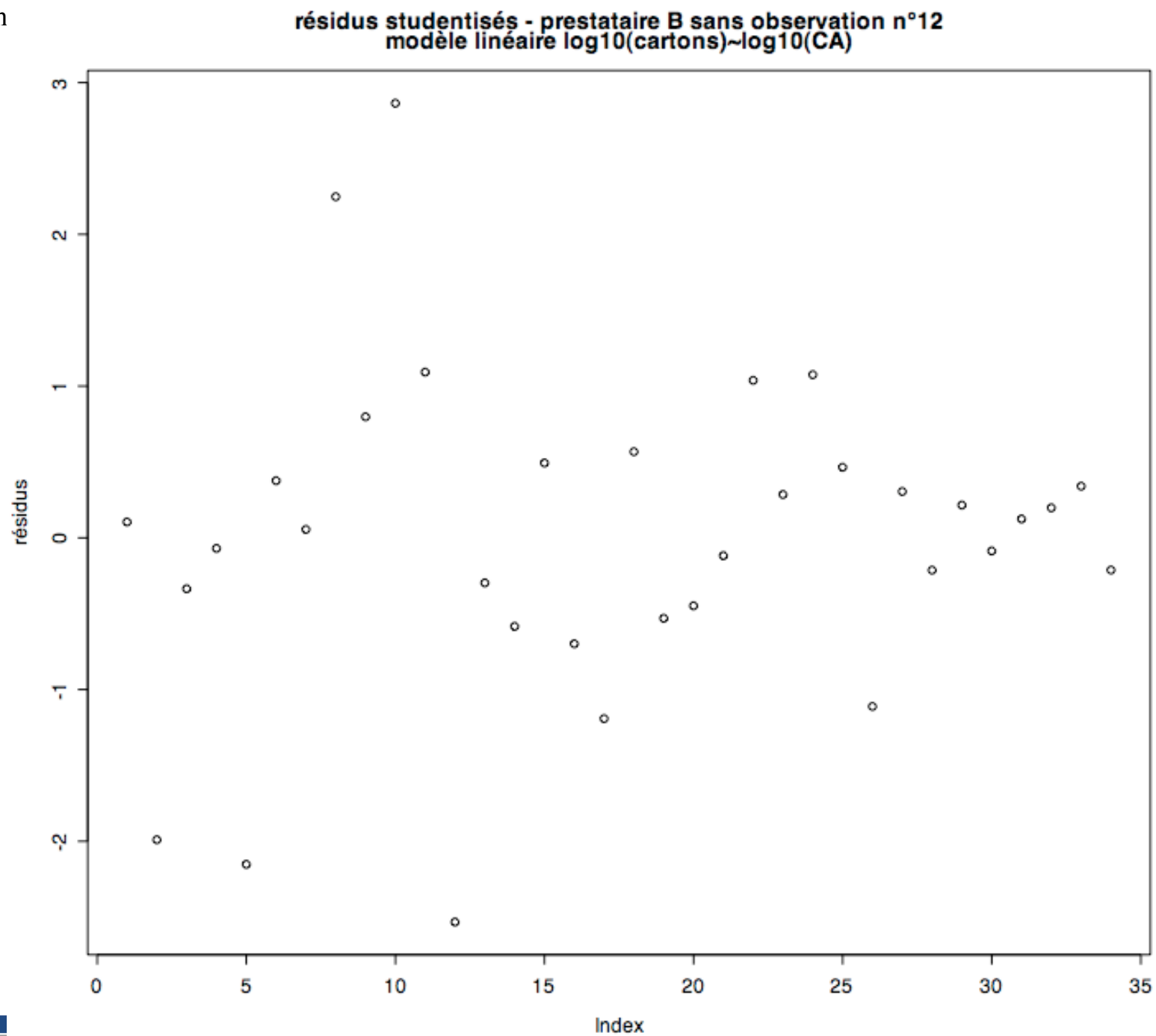
	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-2.17938	0.19594	-11.12	1.58e-12	***
log10(CA)	1.00050	0.04496	22.25	< 2e-16	***

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

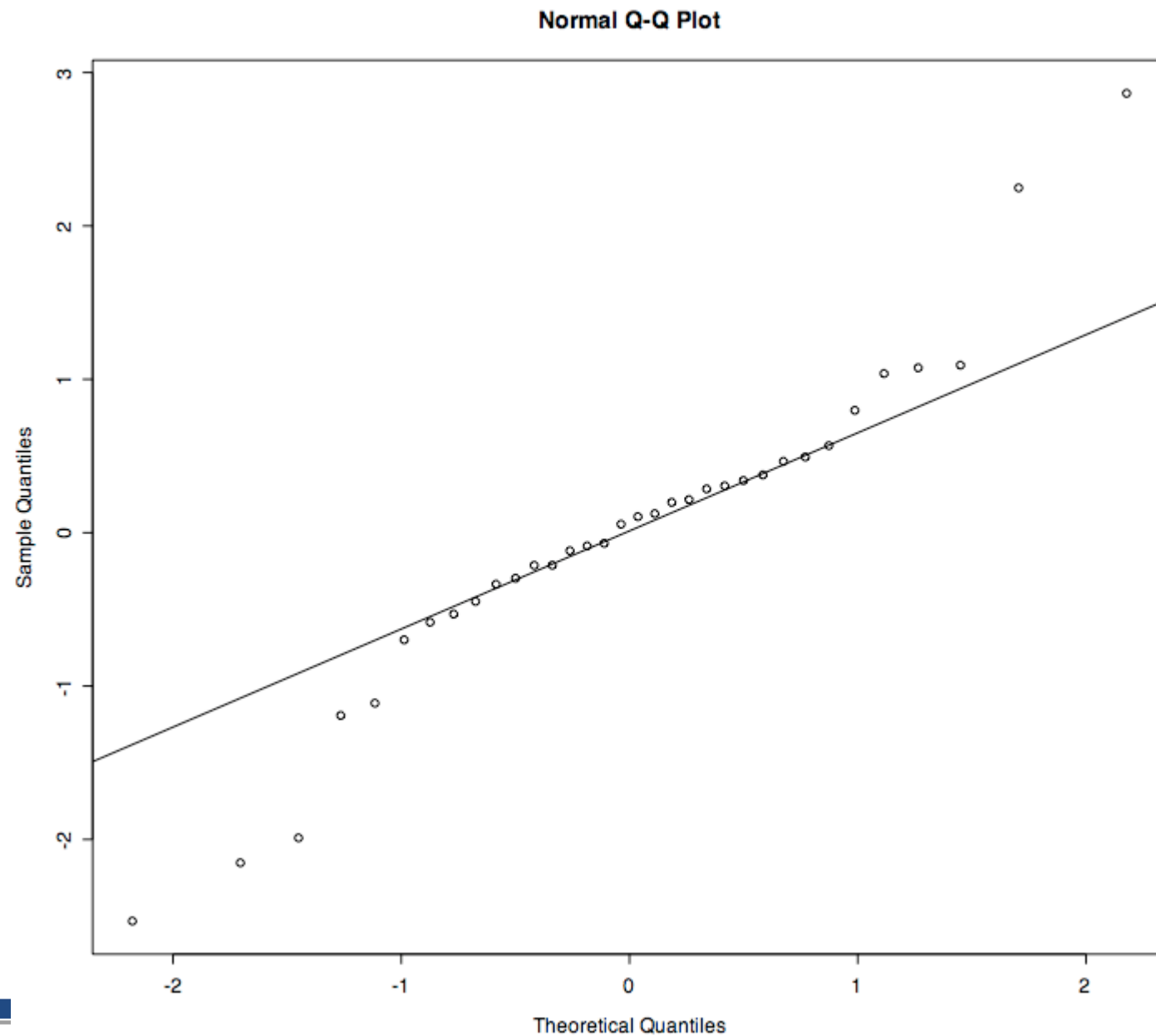
Residual standard error: 0.1019 on 32 degrees of freedom

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

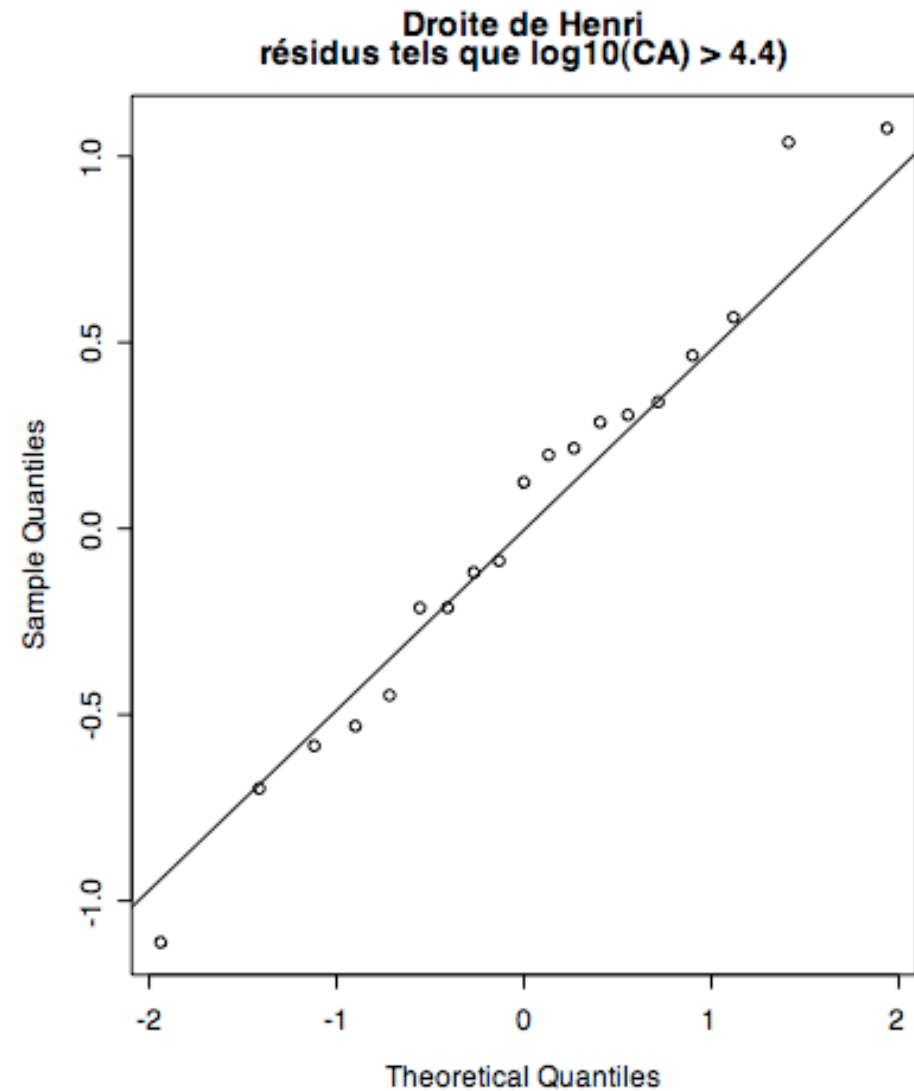
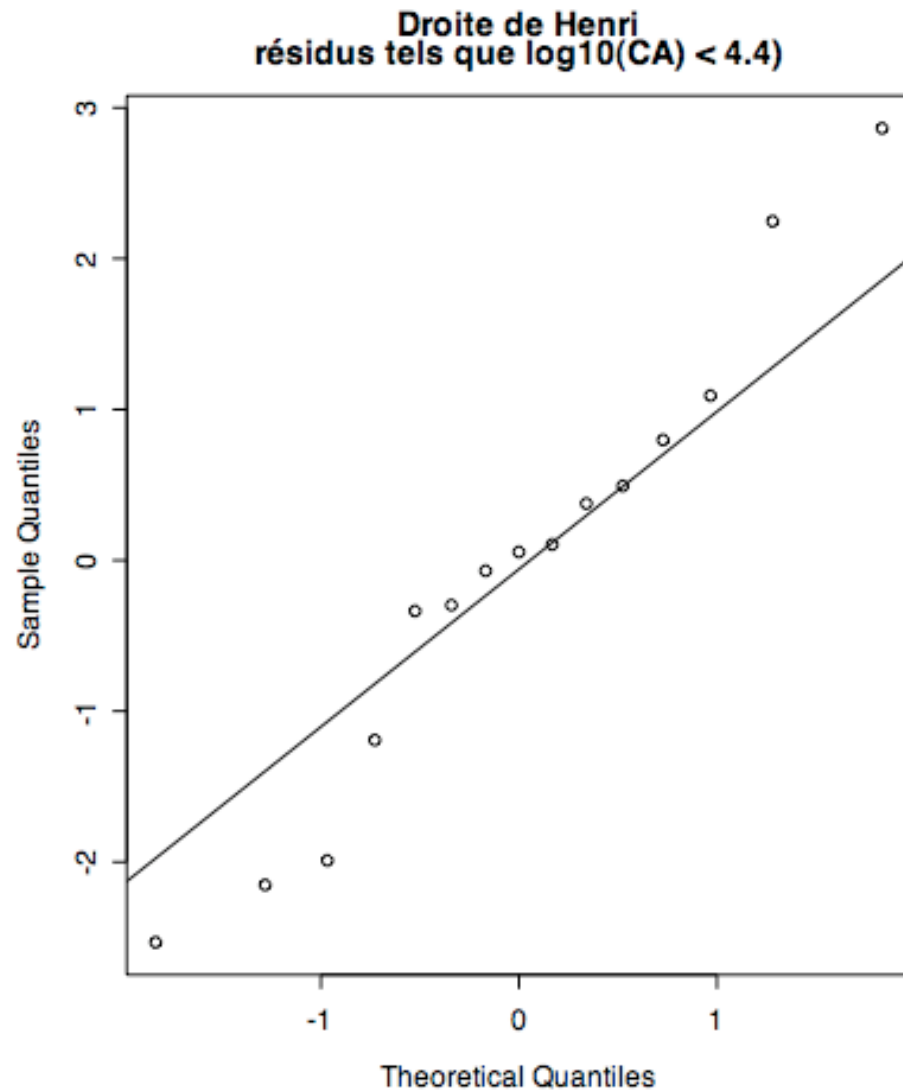
F-statistic: 495.2 on 1 and 32 DF, p-value: < 2.2e-16



Données sans observation n°12



Données sans observation n°12



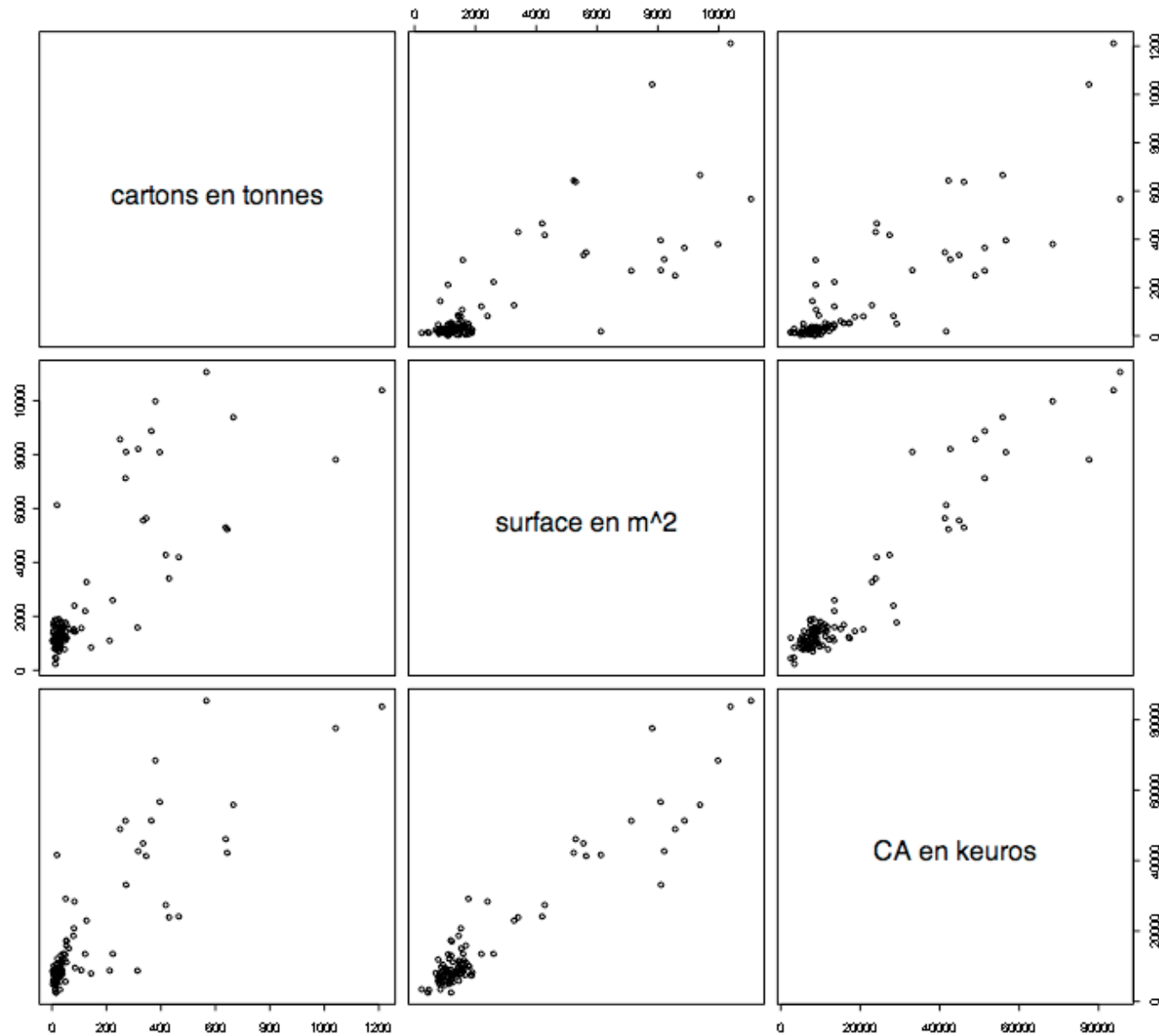
Tests d'adéquation

	Kolmogorov Student	Kolmogorov Gaussienne	Shapiro-Wilk Gaussienne
commandes	<code>ks.test(res, "pt",df)</code>	<code>ks.test(res, "pnorm",0,1)</code>	<code>shapiro.test (res)</code>
35 données	D = 0.17	D = 0.17	W = 0.86
	p-value = 0.24 OK	p-value = 0.25 OK	p-value = 0.0003 NON
sans obs. 12	D = 0.11	D = 0.11	W = 0.95
	p-value = 0.76 OK	p-value = 0.78 OK	p-value = 0.09 OK

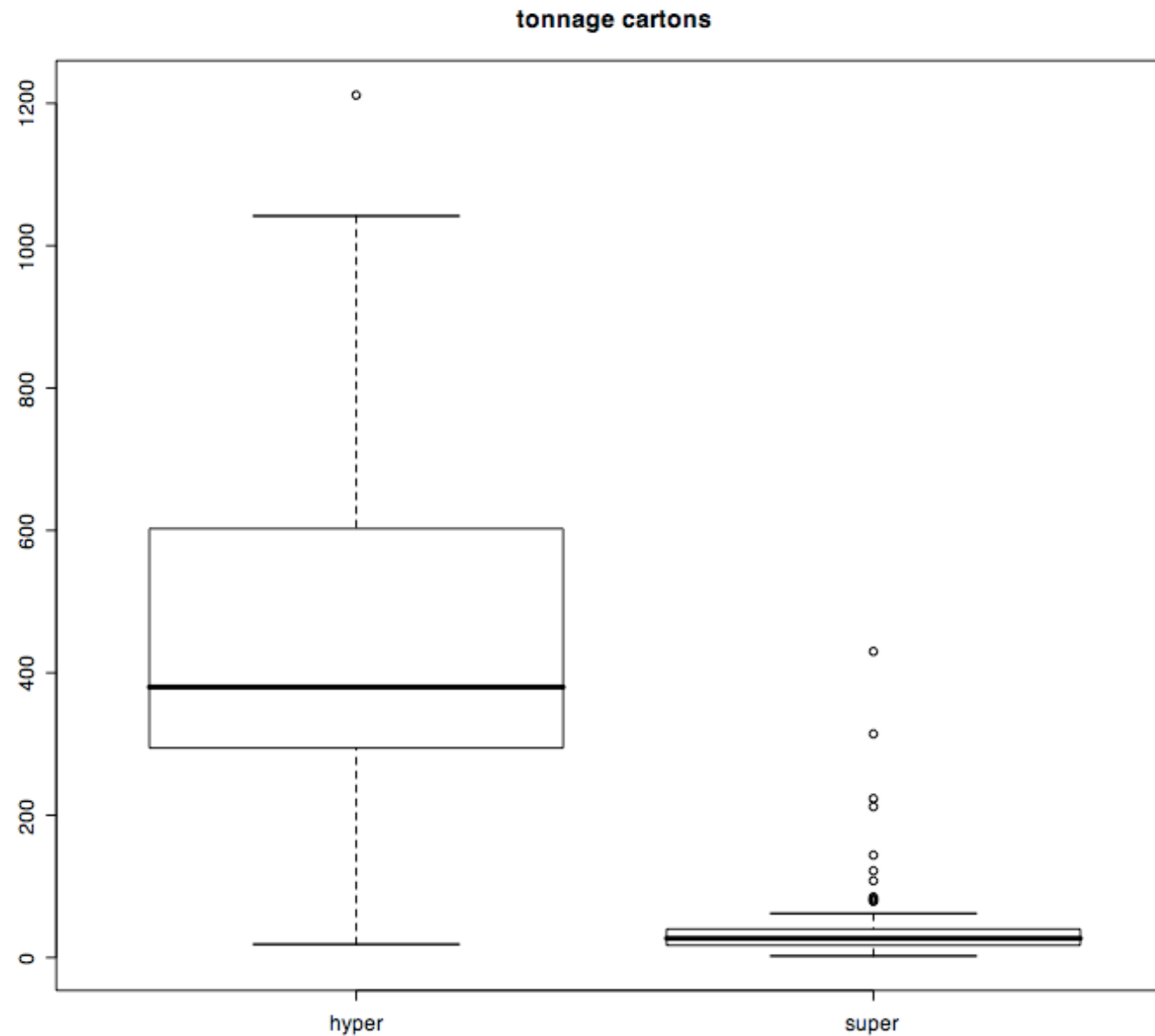
Remarques

- Résidus de variance dépendant du prestataire
- Problème de biais pour de petits CA
- Deux décisions :
 - Séparer les données par prestataire
 - Passage éventuel au log

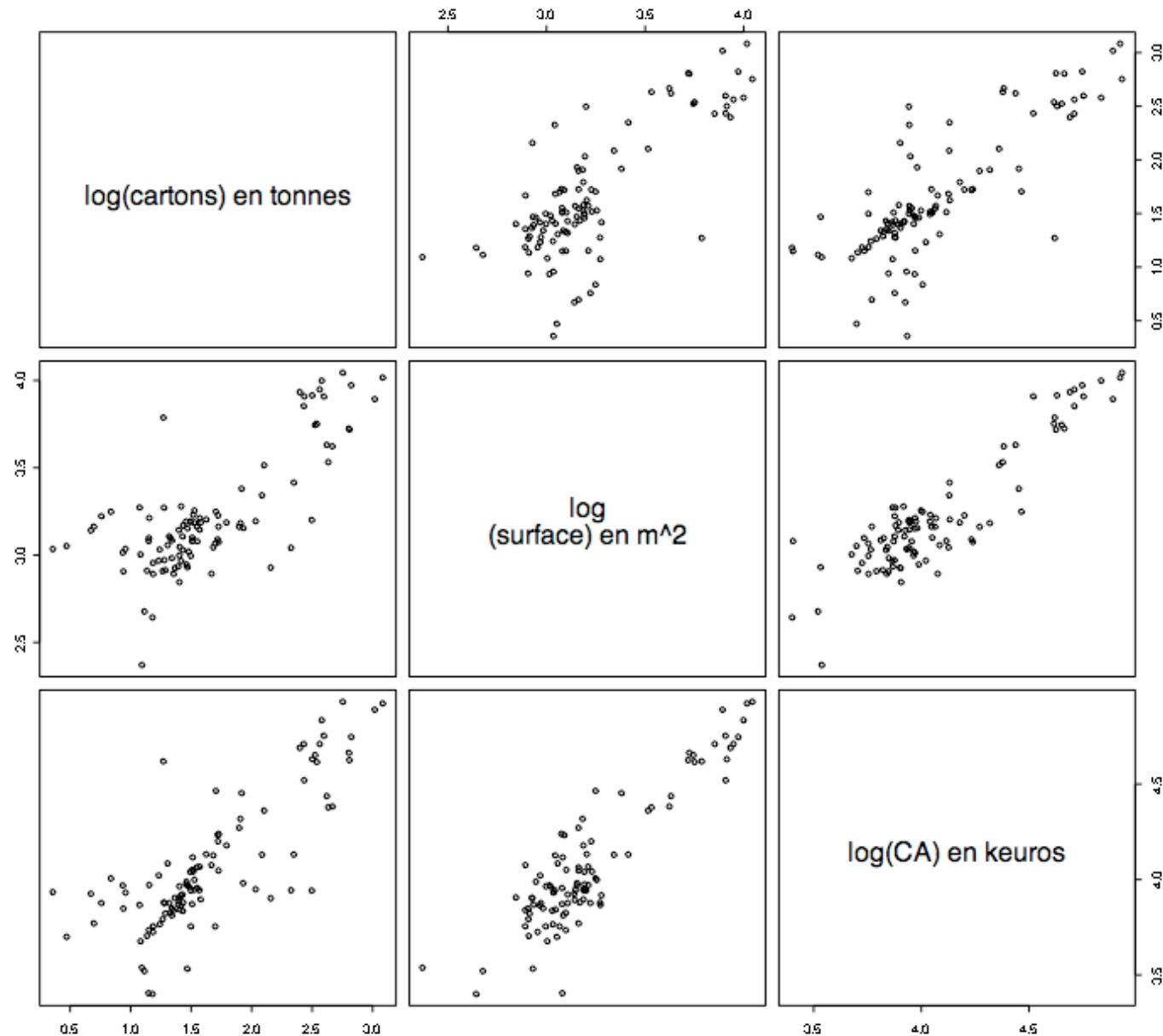
pairs(cartons~surface+CA) - prestataire A



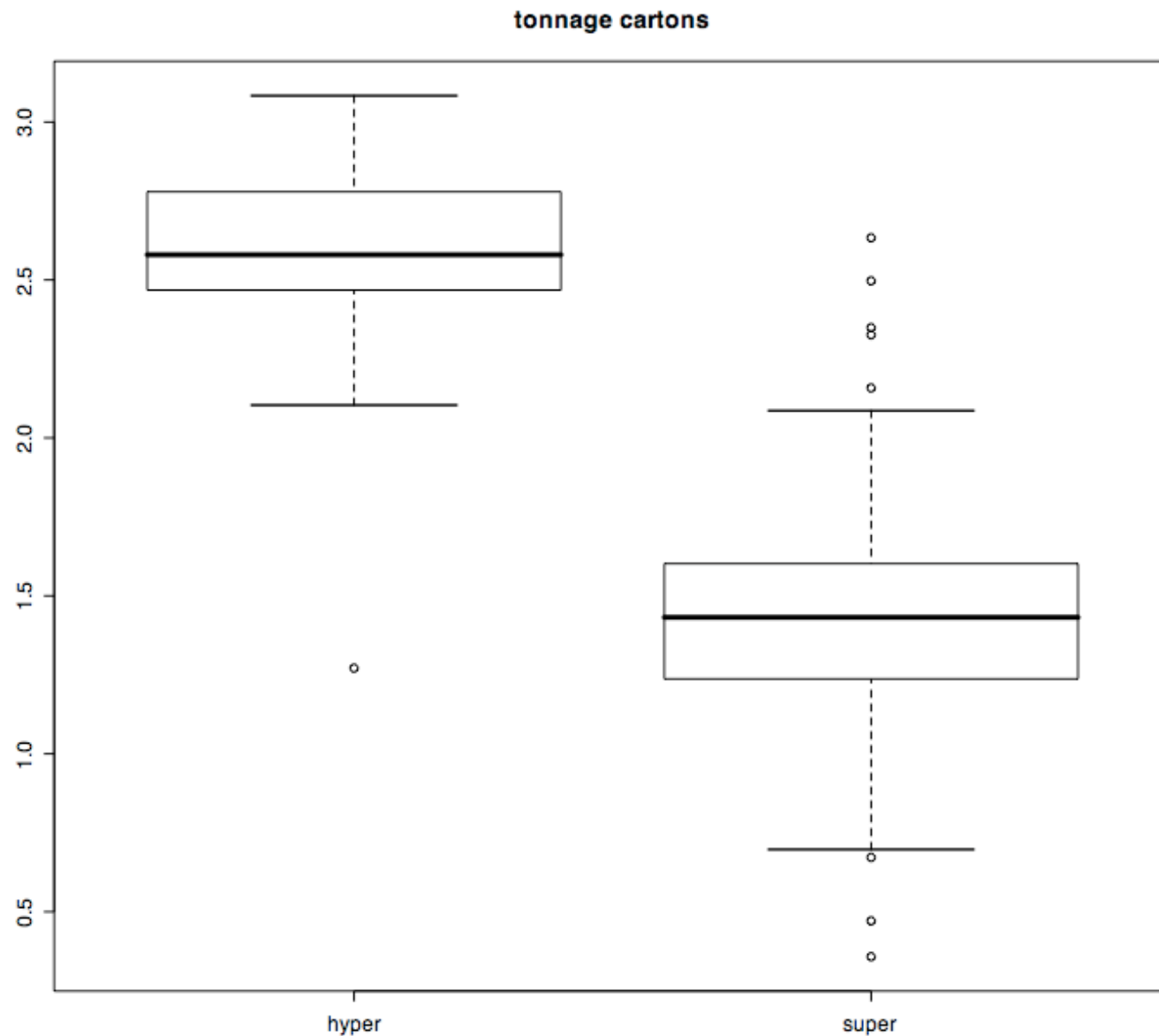
boxplot(cartons~categorie) - prestataire A



`pairs(cartons~surface+CA) - prestataire A- log10`



boxplot(log10(cartons)~categorie) - prestataire A



Modèle pour les log - prestataire A

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

Residuals:

	Min	1Q	Median	3Q	Max
	-1.22837	-0.07155	-0.01158	0.11457	1.02941

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-2.4924	0.9084	-2.744	0.00722	**
log10(CA)	0.9725	0.2382	4.082	9.1e-05	***
log10(surface)	0.1320	0.2809	0.470	0.63941	
categoriesuper	-0.2964	0.1857	-1.596	0.11372	

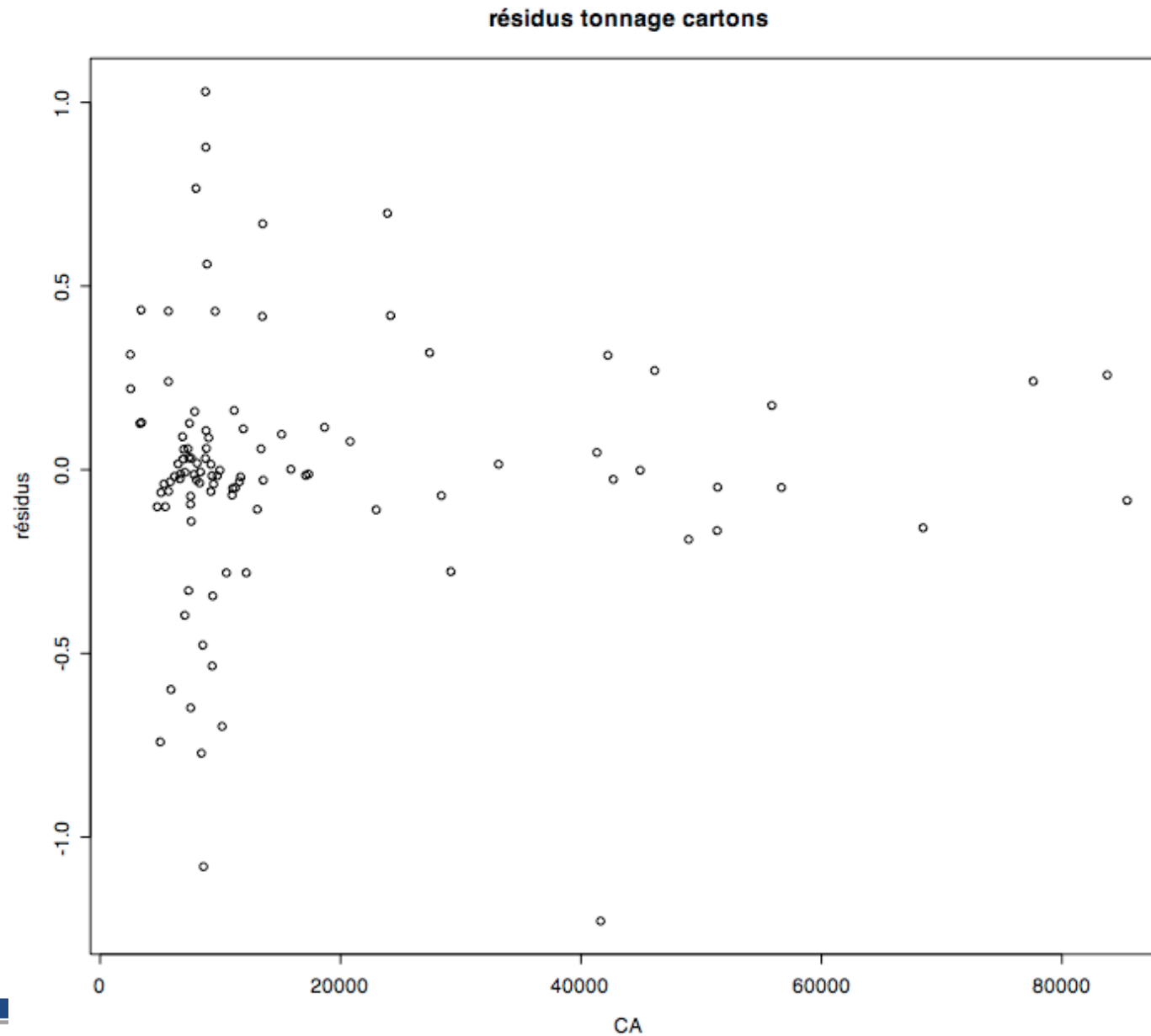
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3474 on 98 degrees of freedom

Multiple R-Squared: 0.6625, Adjusted R-squared: 0.6521

F-statistic: 64.12 on 3 and 98 DF, p-value: < 2.2e-16

résidus de ce modèle - prestataire A



Prévisions

- Nouvelle valeur des prédicteurs x_{new}
- Prédiction pour y :

$$\hat{y}_{\text{new}} = x_{\text{new}} \beta$$

- Intervalle de confiance pour $x_{\text{new}} \beta$ (réponse espérée) ?
- Intervalle de prévision pour la réponse y_{new} ?

Intervalles de confiance/prévision

- De la forme :

$$[x_{\text{new}} \hat{\beta} - s(x_{\text{new}}) t_{n-p-1}^{-1}(1-\alpha/2), x_{\text{new}} \hat{\beta} + s(x_{\text{new}}) t_{n-p-1}^{-1}(1-\alpha/2)]$$

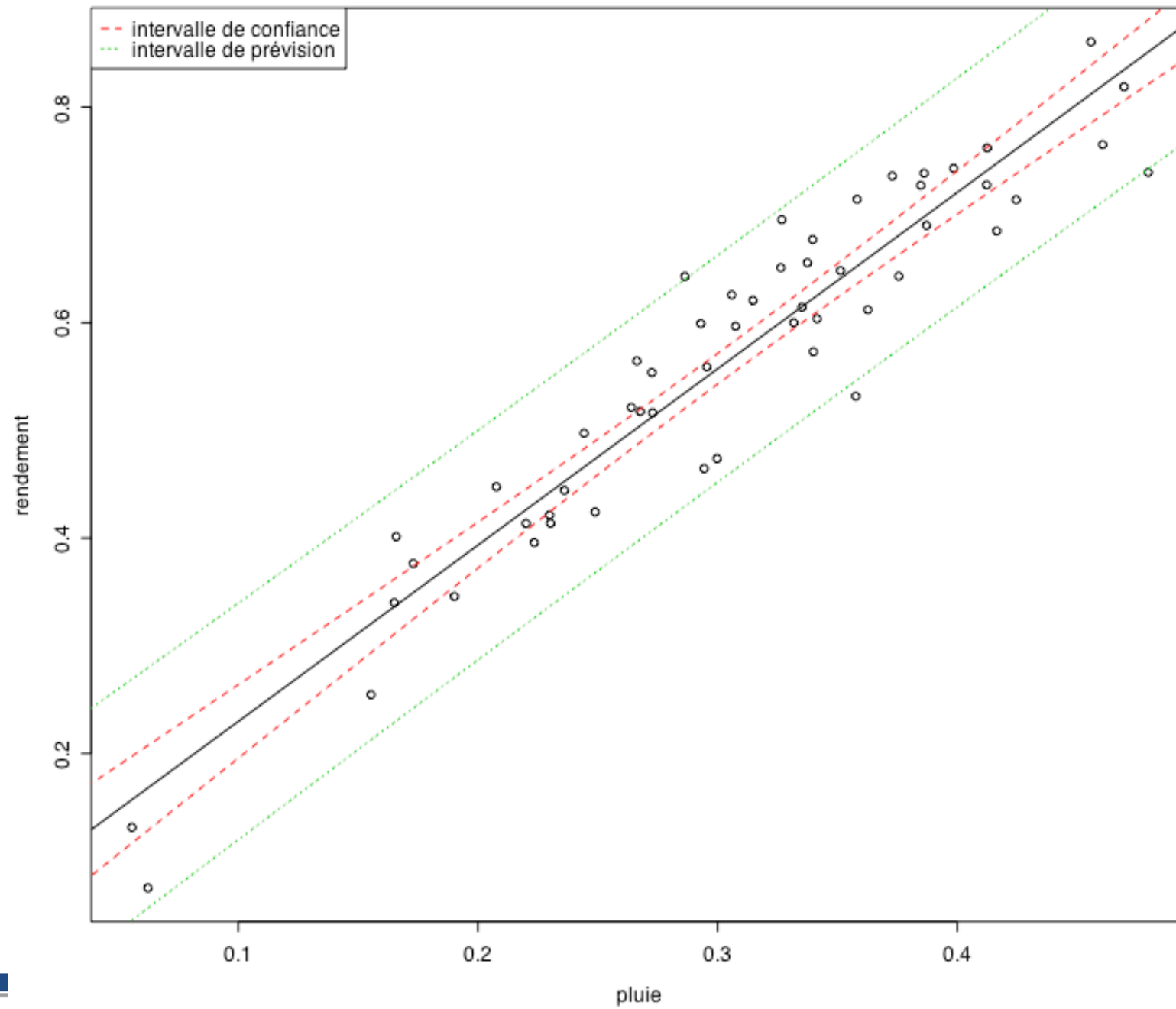
- Confiance :

La pente est-elle >1 ? la droite passe-t-elle par 0 ?

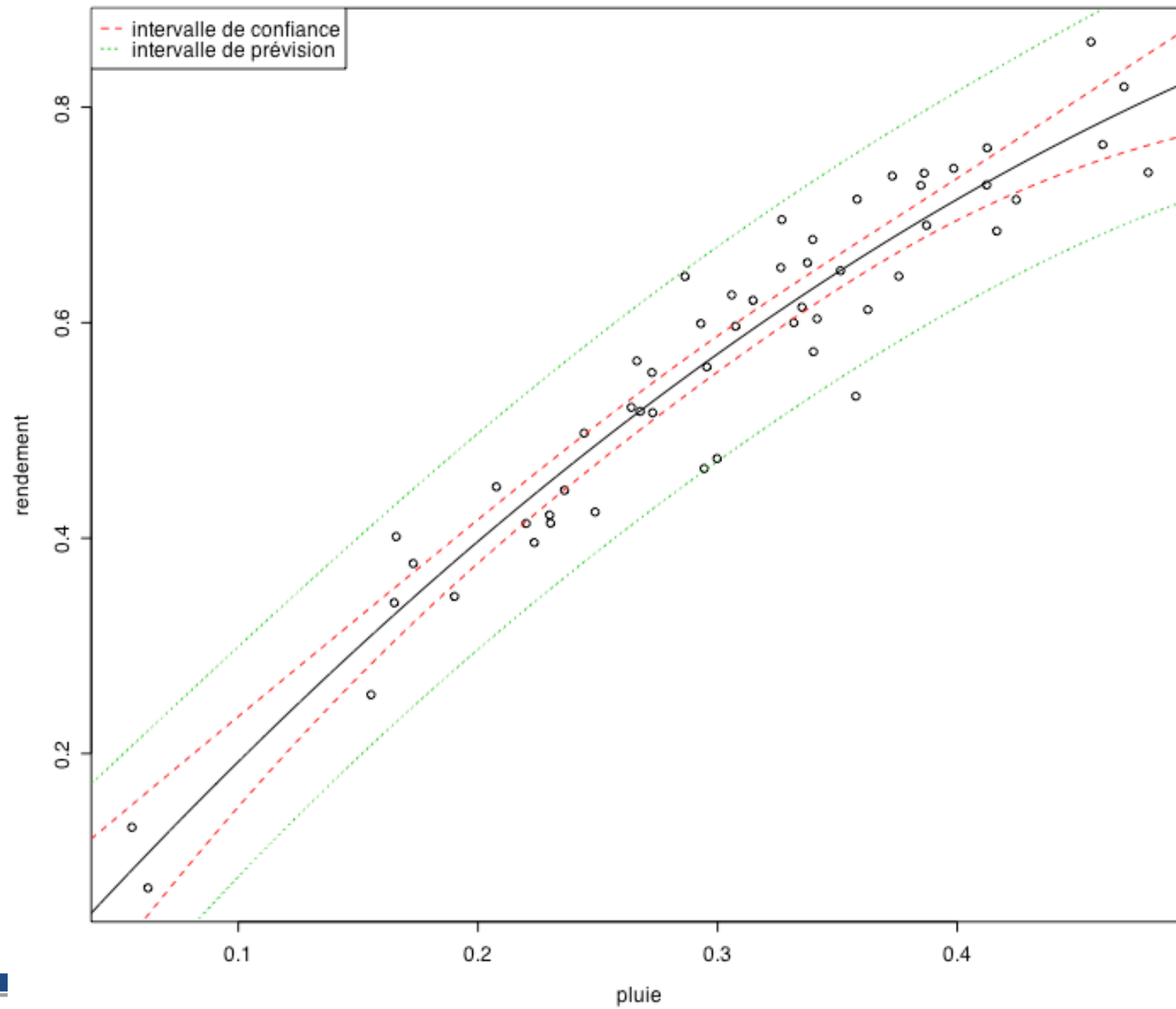
- Prévision :

A quel rendement s'attendre pour 20 mm de pluie ?

intervalles de confiance et de prévision pour le modèle de degré 1



intervalles de confiance et de prévision pour le modèle de degré 2



intervalles de confiance et de prévision pour le modèle de degré 7

