Introduction à la régression
cours n°2

ENSM.SE – axe MSA

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

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

Exercice

- Représenter graphiquement les vecteurs $Y, \hat{Y}$ et $\bar{Y}$ 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 : $Y-\bar{Y}$, $\hat{Y}-\bar{Y}$ et $Y-\hat{Y}$

Formule d'ANOVA

- SST = SSR + SSE
- Coefficient de détermination :

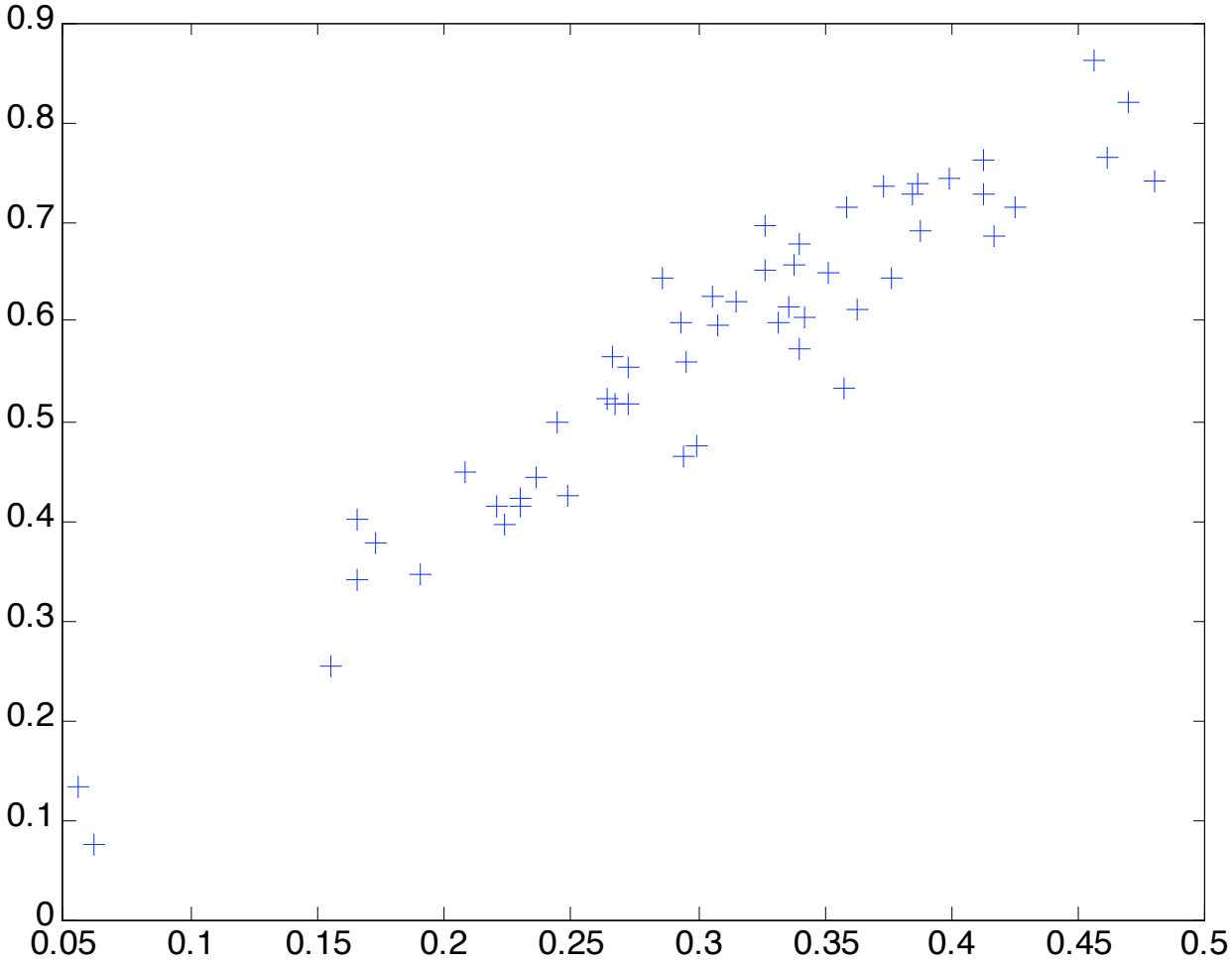
  $R^2 = SSR/SST = 1 - SSE/SST$
- Utilisations de R^2
- Coefficient de détermination ajusté :

  $R^2\text{-}adj = 1 - MSE/MST$

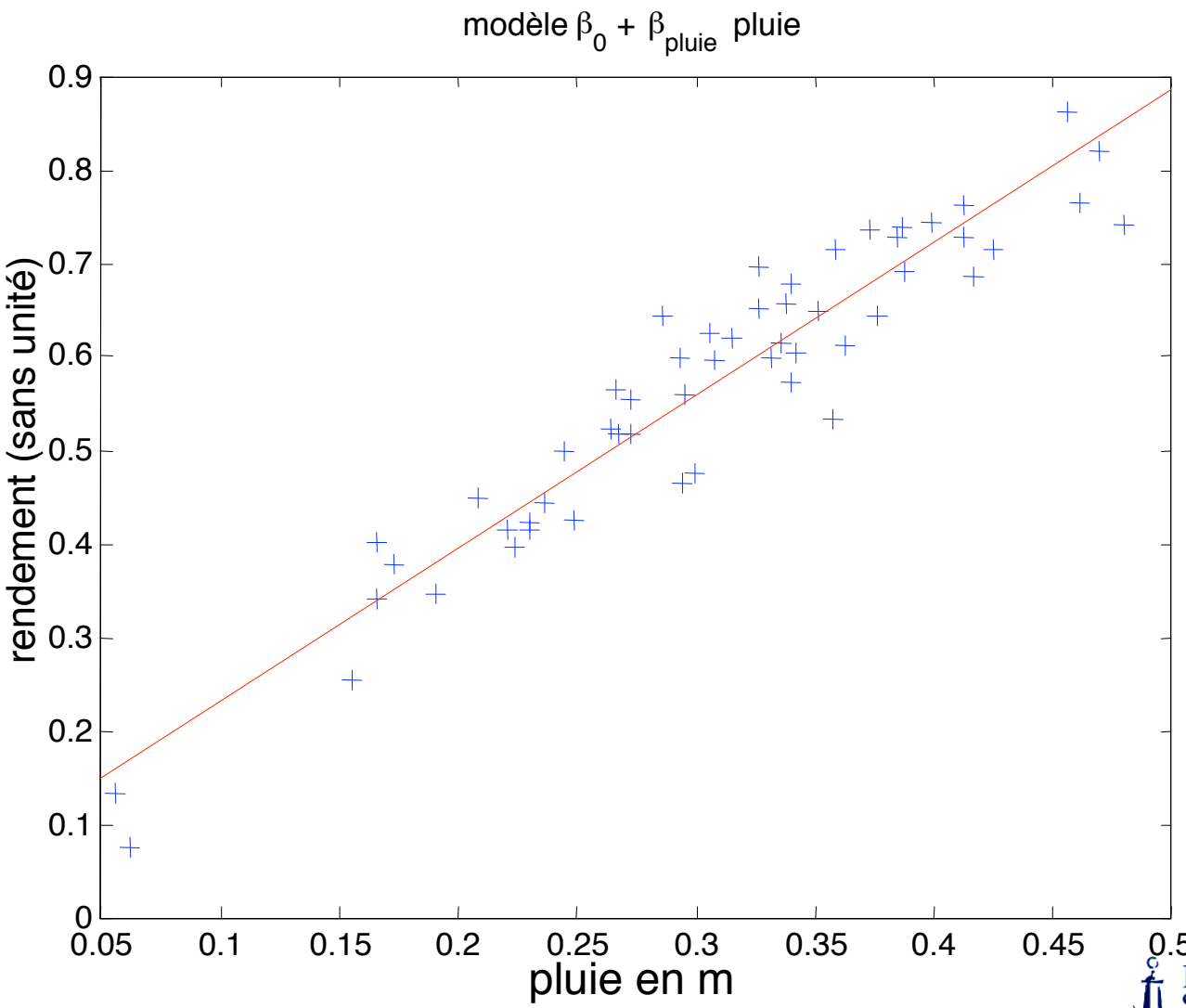
Exemple 2

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

Exemple 1



modèle rend=$\beta_0 + \beta_{pluie}$ pluie

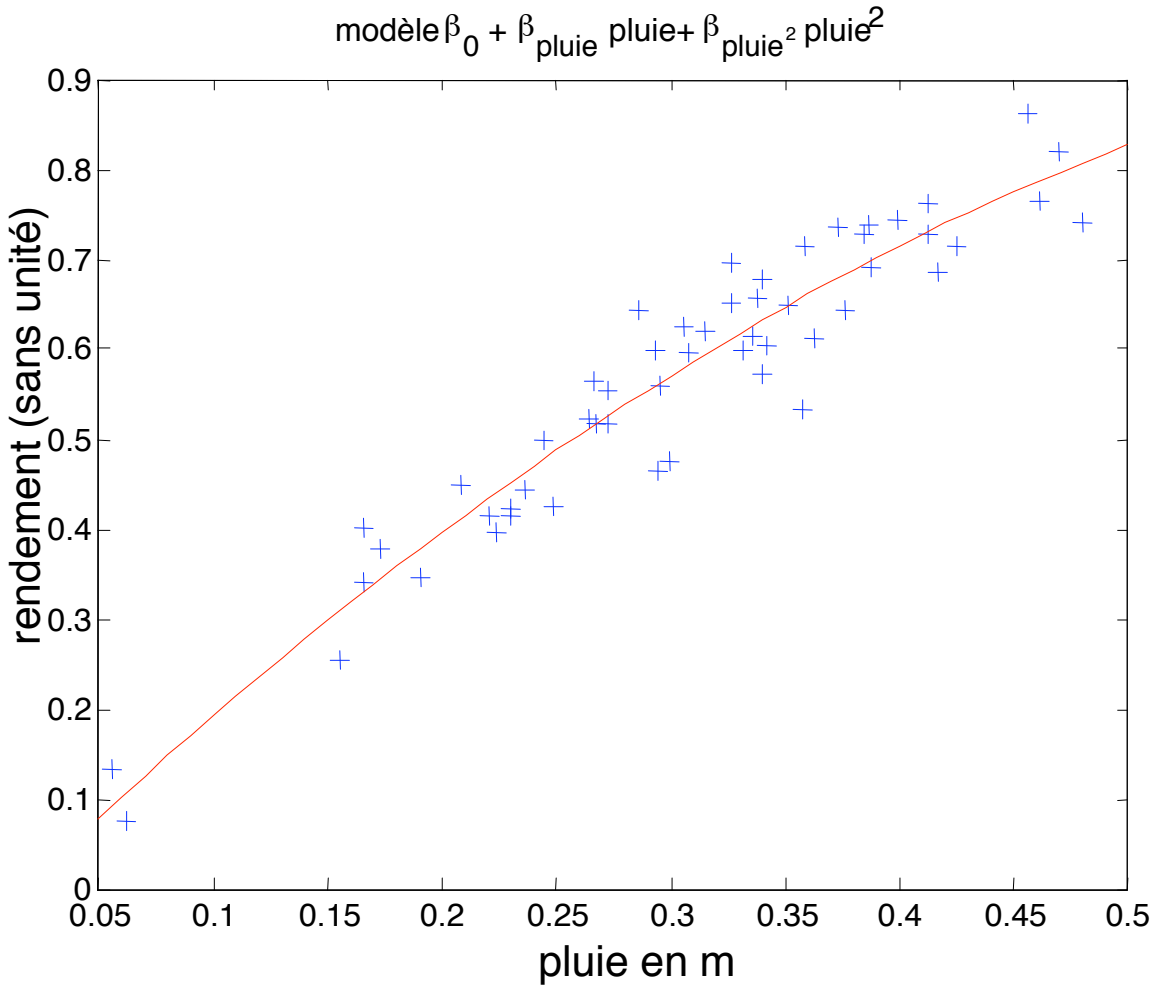


Exemple 2 – modèle de degré 1

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

Source	DF	SS	MS
Regression	1	1.279	1.279
Error	52	0.141	0.003
Total	53	1.420	
R?	0.901	R?-adj	0.899

modèle
$rend=\beta_0 + \beta_{pluie}\ pluie + \beta_{pluie\hat{}2}\ pluie^2$



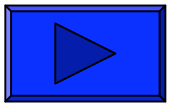
Exemple 2 – modèle de degré 2

Modèle rendement ≈ $\beta_0 + \beta_{pluie}$ pluie + $\beta_{pluie\hat{}2}$ pluie?

Source	DF	SS	MS
Regression	2	1.298	0.649
Error	51	0.122	0.002
Total	53	1.420	
R?	0.914	R?-adj	0.911

Questions de prédicteurs

- prédicteurs qualitatifs

  codage des niveaux 

- prédicteurs corrélés

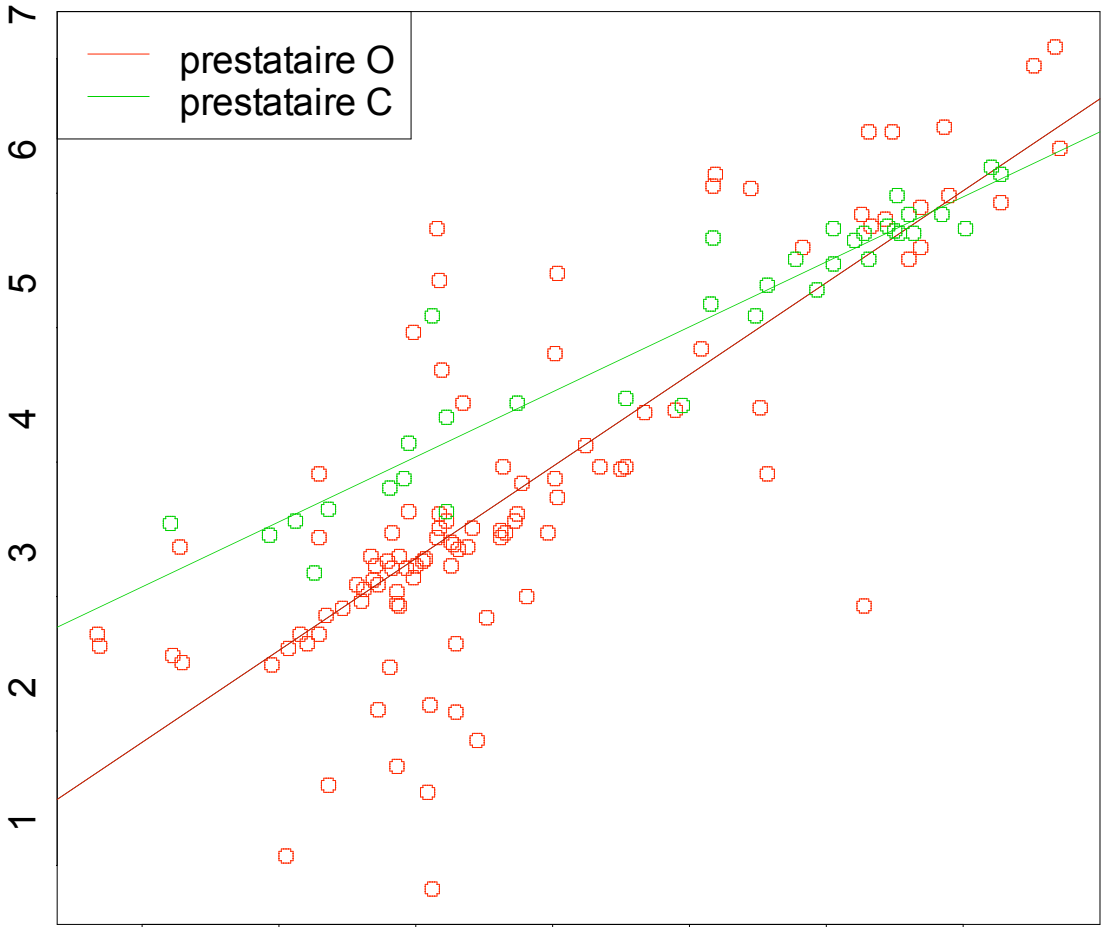
  problème d'identifiabilité

- interaction de prédicteurs
  - notion **liée à la réponse**
  - sans rapport avec la corrélation

Exemple 1

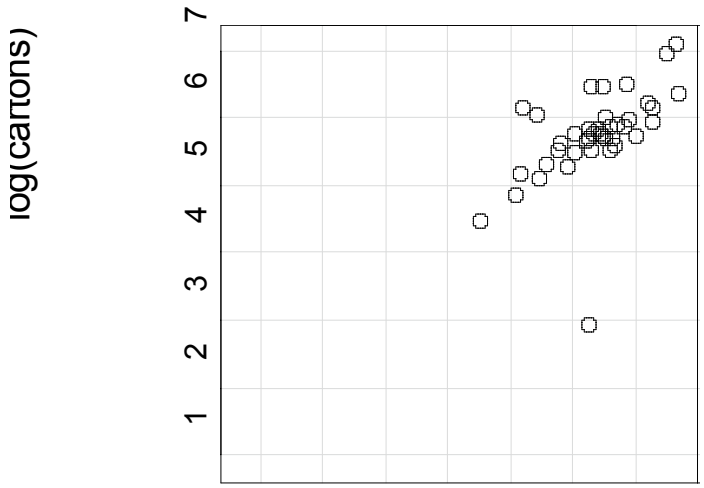
- Rappel étape précédente : passer au logarithme et modèle de degré 1.
- Autres prédicteurs :
  - Surface
  - Catégorie
  - Prestataire

Interaction CA*prestataire

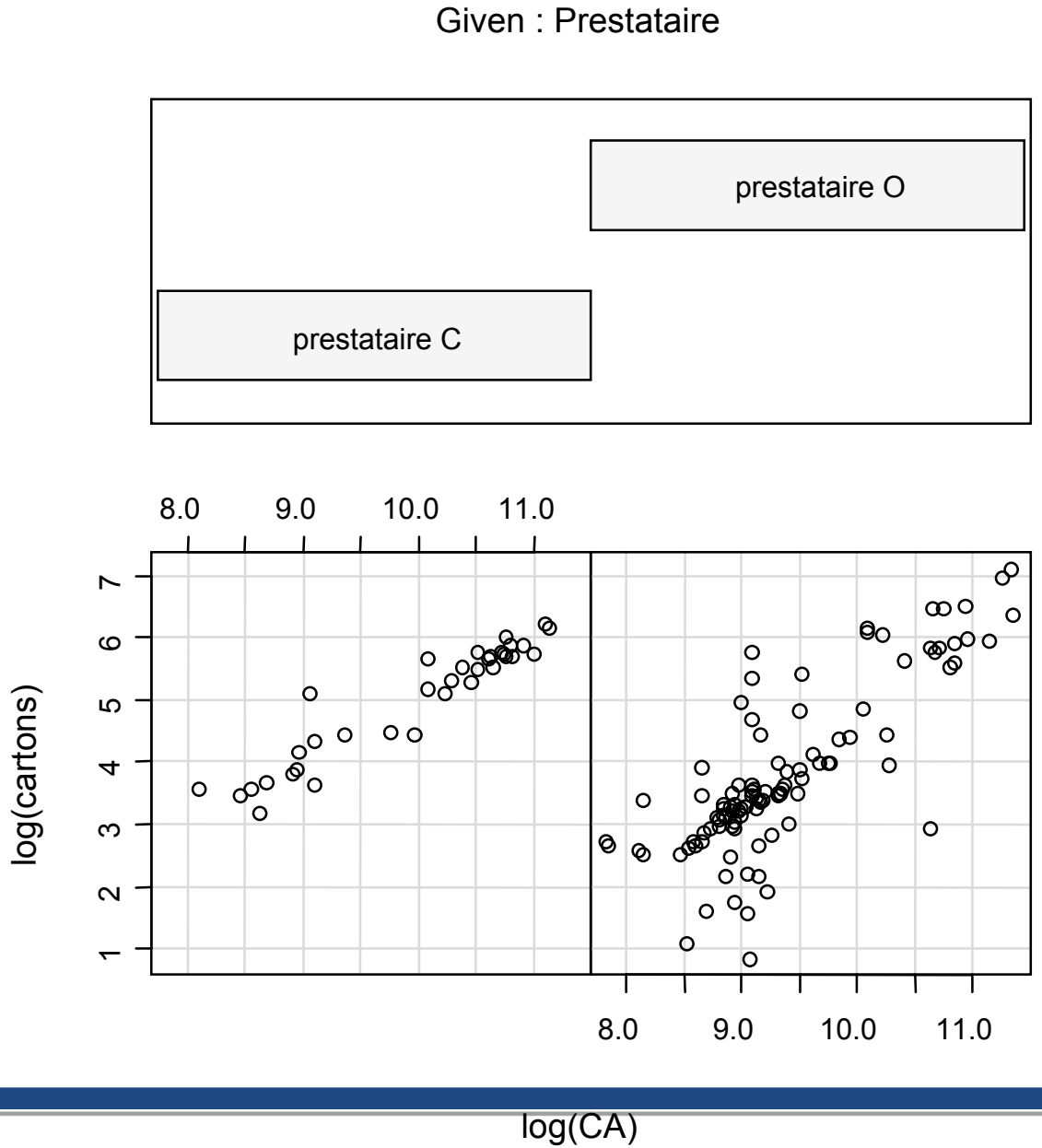


Interaction CA*surface

Interaction CA*catégorie



Interaction CA*prestataire



Le modèle linéaire

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

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

Hypothèses sur les résidus

- $\varepsilon_1, \ldots, \varepsilon_n$ sont des réalisations de v.a. $E_1, \ldots, E_n$
- Les E_i sont indépendantes
- Les E_i sont de loi $N(0,\sigma^2)$

Estimation de β

- Estimation par Maximum de vraisemblance
- résultat = moindres carrés !
- adaptation à des hypothèses différentes sur la loi des résidus
- à voir : régression logistique

Propriétés de $\hat{\beta}$

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

Exercice

Soit un modèle de dimension 1 avec un seul prédicteur :

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- Déterminer $\mathrm{cov}(\hat{\beta})$
- Les expériences sont menées avec $x \in [-1,1]$. Déterminer les niveaux des variables de façon à estimer « au mieux » β.