Bandes de confiance simultanées par la méthode de Scheffé dans le cas de la régression linéaire simple

J.M. Azaïs *

21 mai 2012

Considérons un échantillon de femmes enceintes qui étaient parfaitement réglées. Par echographie on constitue un échantillon des deux variables suivantes

X l'âge du foetus mesuré en semaine d'aménorrhée Y la taille du fémur du foetus

Y	X	Y	X
178	48.2	152	39.9
185	47.4	168	44
157	37.7	106	17.1
226	58.8	119	21.6
227	58.3	186	46.7
153	38	175	45.4
184	52.6	125	26.2
227	59	139	33.5
140	30	148	35.6
139	31.6	155	38
230	62.4	150	36.8
122	24.1	154	38.4
180	45.2		

données diponibles à

`http://dl.dropbox.com/u/17864747/femur.m`

Il est clair que c'est la variable X qui a une influence sur Y et il est donc logique de faire la régression de Y sur X (on peut vérifier qu'une régression linéaire simple suffit).

Maintenant supposons qu'un femme qui ne connait pas l'âge de son foetus subit un échographie et la longueur du foetus est mesurée à 120 mm. Quel est l'âge estimé du foetus et comment construire un intervalle de confiance pour cet âge ? Comme nous allons le voir la réponse a cette question demande de

*Université de Toulouse, Statistique et Probabilités, IMT, Mél : jean-marc.azais@math.univ-toulouse.fr

construire un intervalle de confiance pour la droite de régression, un tel intervalle est dit simultané (voir définition plus bas).

1 Le modèle

On considère donc le modèle de régression linéaire simple.

$$Y_i = \mu + \beta X_i + \epsilon_i \quad i = 1, \ldots, n. \tag{1}$$

avec les hypothèses classiques : les ϵ_i sont i.i.d. de loi $N(0, \sigma^2)$.

Nous supposerons pour simplifier que le nombre d'observations n est grand ($n > 100$)(ce qui n'est pas le cas dans l'exemple numérique) Sans perte de généralité on peut également supposer que la variable X est centrée et d'inertie 1 : $\sum_{i=1}^n X_i = 0$; $\sum_{i=1}^n X_i^2 = 1$.

On sait bien qu'alors

$$\widehat{\mu} = \overline{Y} \qquad \text{est gaussien de variance } \sigma^2/n$$

$$\widehat{\beta} = \sum_{i=1}^{n} Y(t)X(t) \qquad \text{est gaussien de variance } \sigma^2,$$

et ces deux estimateurs sont indépendants.

La prédiction de la réponse au point x est donnée par

$$\widehat{Y}(x) = \widehat{\mu} + \widehat{\beta}x \quad \text{de variance } \sigma^2(\frac{1}{n} + x^2).$$

En centrant et réduisant on définit

$$\overline{Z(x)} := \frac{\widehat{Y}(x) - (\mu + \beta x)}{\sqrt{\sigma^2(x^2 + 1/n)}},$$

σ^2 qui est inconnu. Mais comme le nombre d'observations est grand, l' estimateur

$$\widehat{\sigma}^2 := 1/(n-2) \sum_{i=1}^{n} (Y_i - \widehat{\mu} - \widehat{\beta}X_i)^2.$$

est presque exact. et on peut remplacer sans conséquence σ^2 par $\widehat{\sigma}^2$ pour obtenir la quantité

$$Z(x) := \frac{\widehat{Y}(x) - (\mu + \beta x)}{\sqrt{\widehat{\sigma}^2(x^2 + 1/n)}},$$

qui est maintenant calculable.

Ce qui donne l'intervalle de confiance

$$Y(x) := (\mu + \beta x) \in IC(x) := \left[\widehat{Y}(x) \pm Z_\alpha \sqrt{\widehat{\sigma}^2(x^2 + 1/n)}\right], \tag{2}$$

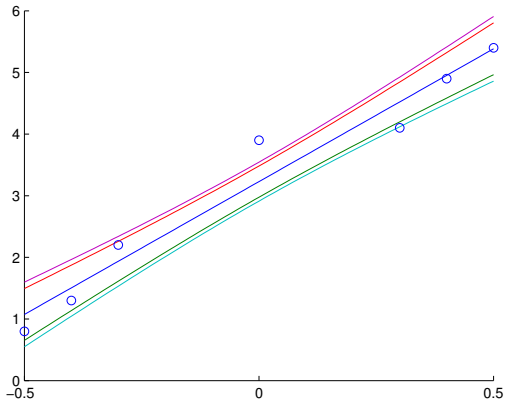

FIGURE 1 – Exemple de données avec la droite de régression, la bande de confiance non-simultanée, la bande de confiance simultanée (à l'extérieur)

où Z_α est la valeur critique au niveau α d'un loi normale centrée réduite, c'est à dire le fractile $1-\alpha/2$.

Si on fait varier x on obtient une bande autour de la droite de régression délimitée par une hyperbole.

Montrons qu'il s'agit bien d'une hyperbole : l'équation (2) peut se réécrire

$$y = A + Bx \pm C\sqrt{1+Dx^2}$$

ce qui se réécrit encore

$$(y - A - Bx)^2 = C^2 + C^2 D x^2.$$

C'est donc bien une conique et comme les coefficients de y et x dans le carré sont opposés, on voit que x et Y vont vaier de $-\infty$ à $+\infty$ c'est donc une hyperbole et non un ellipse ou une parabole.

La bande ainsi construite a la propriété **non simultanée** suivante. Pour x fixé

$$\mathrm{P}\{Y(x) \in IC(x)\} = 1-\alpha.$$

Mais si x varie dans un ensemble $\mathcal{X}$, la bande ci-dessus n'est pas simultanée dans le sens où en général

$\mathrm{P}\{\forall x \in \mathcal{X} : Y(x) \in IC(x)\}$ est en général beaucoup plus petite que $1-\alpha$.

Pour obtenir pour une nouvelle bande $B(x)$ **simultanée** (plus grande) vérifiant

$$\mathrm{P}\{\forall x \in \mathcal{X} : Y(x) \in B(x)\} = 1-\alpha, \tag{3}$$

on a recours à

2 La méthode de Scheffé

On sait que

$$\begin{bmatrix} \frac{\widehat{\mu}-\mu}{\sqrt{\sigma^2/n}} \\ \frac{\widehat{\beta}-\beta}{\sigma} \end{bmatrix} \sim N(0, I_2).$$

Donc en définissant, pour bien normaliser, $\nu := \sqrt{n}\mu$

$$\frac{(\widehat{\nu}-\nu)^2}{\widehat{\sigma}^2} + \frac{(\widehat{\beta}-\beta)^2}{\widehat{\sigma}^2} \sim \chi^2(2),$$

approximativement. On obtient donc une région de confiance pour (ν, β) qui est un disque en écrivant

$$(\widehat{\nu}-\nu)^2 + (\widehat{\beta}-\beta)^2 \leq \widehat{\sigma}^2 \chi^2_{1-\alpha}(2), \tag{4}$$

où $\chi^2_{1-\alpha}(2)$ est le fractile $1-\alpha$ de la loi $\chi^2(2)$.

Maintenant écrivons

$$\mu + \beta x = \sqrt{1/n + x^2} \;\; \langle (\nu, \beta), (\frac{n^{-1/2}}{\sqrt{1/n+x^2}}, \frac{x}{\sqrt{1/n+x^2}} \rangle.$$

Notez que le second vecteur dans le produit scalaire est de norme 1 et que dans (4) le rayon du disque vaut $\sqrt{\widehat{\sigma}^2 \chi^2_{1-\alpha}(2)}$. Il est facile d'en déduire que

$$|\mu + \beta x - (\widehat{\mu} + \widehat{\beta} x)| \leq \sqrt{(1/n + x^2)(\widehat{\sigma}^2 \chi^2_{1-\alpha}(2))},$$

avec proba $1-\alpha$, uniformément en x. La région est donc simultanée au sens de (3).

3 Suggestions

- Détailler toutes les étapes qui conduisent à la contruction de (2).
- Détailler soigneusement la preuve de la méthode de Sheffé.
- Étendre la méthode de Sheffé au cas où n n'est pas grand.
- Construire les deux région sur l'exemple et en déduire l'intervalle de confiance pour l'âge quand la longueur du fémur est 120 mm.
- Étendre la méthode à un modèle de régression multiple en utilisant un argument d'orthogonalisation de Gram-Schmidt. Plus précisément si

  $$Y_i = \beta_0 X_i^0 + \beta_1 X_i^1 + \cdots X_i^{p-1} + \epsilon_i \;\; i = 1, \ldots, n, \tag{5}$$

  est un modèle de régression où X^0 est le vecteur composé de 1, et X^1... X^{p-1} sont les autres régresseurs (l'exposant ne signifie pas un puissance !). On peut se ramener au ca s ou tous ces regresseurs sont orthogonaux et d'inertie 1. Dans ce cas, la méthode se généralise facilement.

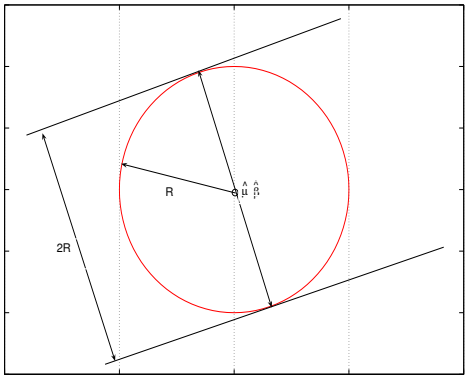

FIGURE 2 – Représentation du disque de confiance autour de la valeur $\widehat{\nu}, \widehat{\beta}$ et de sa projection sur une droite quelconque. On constate bien que cette projection a la même taille qu'un diamètre.