

Intervalles de confiance séparé et simultané : l'exemple du fichier "fines".

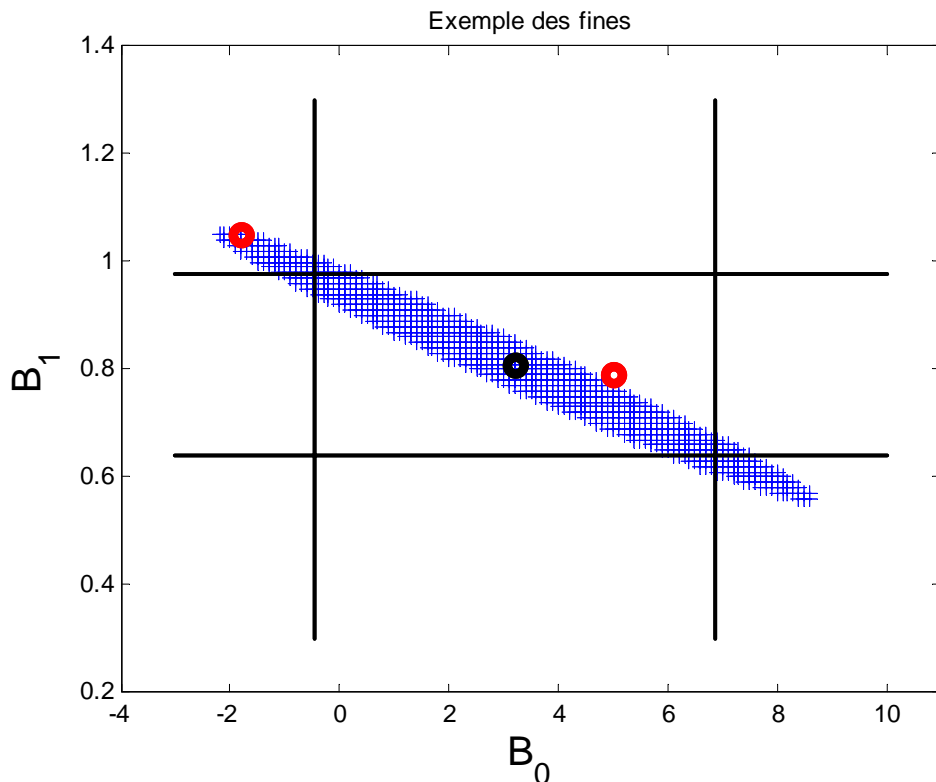
Données: fichier « fines.dat » 106 mesures de la quantité de fines dans un sable (Y : mesure par lavage au tamis; X : mesure par sédimentométrie)

Coefficients estimés: $b_0=3.2$, $b_1=0.81$

Écart-type sur $b_0=2.2$ -> intervalle de confiance $[3.2 \pm 1.66*2.2]=[-0.45 \ 6.85]$

Écart-type sur $b_1=0.10$ -> intervalle de confiance $[0.81 \pm 1.66*0.1]= [0.64 \ 0.98]$

Ellipse (ou ellipsoïde) de confiance: $(\beta-b)'X'X(\beta-b) \leq (p+1)s^2F_{p+1,n-(p+1),1-\alpha}$



L'ellipse et l'intervalle classique ne correspondent absolument pas! La raison est simple, la corrélation entre les coefficients b_0 et b_1 est de -0.98 , il faut en tenir compte.

A- Considérons le modèle théorique avec $B_0=-1.8$, et $B_1=1.05$. Les intervalles de confiance autour de b_0 et de b_1 séparément n'incluent pas ces valeurs. Pourtant le modèle est dans l'ellipse. Vérifions que c'est correct en effectuant la régression de:

$$Y - B_0 - B_1 * X = b_0 + b_1 * X + e$$

On trouve $SCR=91.8$ et $SCE=1621$.

$$H_0 \beta_0 = 0, \quad \beta_1 = 0$$

On calcule: $(91.8/2)/(1621/104)=2.94 < 3.08$ On ne peut rejeter H_0 donc on conclut que les données ne s'écartent pas significativement du modèle $B_0=-1.8$, et $B_1=1.05$.

B- Considérons le modèle théorique avec $B_0= 5$, et $B_1=0.79$. Les intervalles de confiance autour de b_0 et de b_1 séparément incluent ces deux valeurs. Pourtant le modèle n'est pas dans l'ellipse. Vérifions que c'est correct en effectuant la régression de :

$$Y - B_0 - B_1 * X = b_0 + b_1 * X + e$$

On trouve $SCR=210.9$ et $SCE=1621$.

$$H_0 \beta_0 = 0, \quad \beta_1 = 0$$

On calcule: $(210.9/2)/(1621/104)=6.8 > 3.08$ On doit rejeter H_0 donc on conclut que les données s'écartent significativement du modèle $B_0= 5$, et $B_1=0.79$.

Conclusion : Pour comparer les résultats d'un modèle de régression avec un modèle théorique, il faut considérer simultanément tous les coefficients du modèle et donc utiliser l'ellipsoïde de confiance simultané. L'intervalle de confiance peut être utilisé seulement pour juger si une variable donnée (une seule à la fois) ajoute de l'information significative par rapport au modèle qui l'exclurait.