

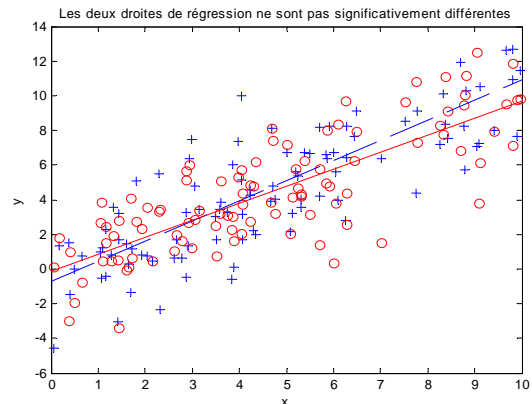
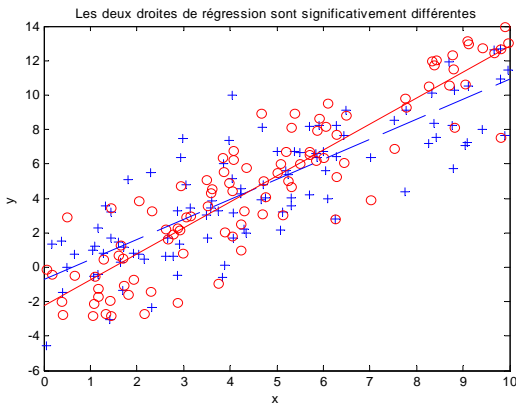
Deux jeux de données différents ont-ils même régression?

On compare le modèle réduit : $Y=b_0+b_1X+e$

au modèle complet :

$$Y=c_0+c_1X+c_2J+c_3JX+e$$

où J est une indicatrice qui prend la valeur 1 ou 0 selon le jeu de données.



À gauche, le test d'ajout sur la variable indicatrice J et JX est significatif ($F=7.95$, $F_{\text{table}}=3.09$). La simulation a effectivement été générée avec une pente de 1 pour les « + » et une pente de 1.5 pour les « o ». À droite, le test d'ajout sur la variable indicatrice J et JX n'est pas significatif ($F=2.43$, $F_{\text{table}}=3.09$). La simulation a été générée avec une pente de 1 pour les « + » et les « o ».

Pour le cas de droite, on a 5% des chances de déclarer les pentes différentes alors qu'en réalité elles sont égales (erreur de type α). De même, dans le cas de gauche, on aurait pu déclarer les pentes égales alors qu'en réalité elles sont différentes (erreur de type β). On ne peut dire avec quel risque cette éventualité se présente car elle dépend des valeurs inconnues des 2 droites et du nombre d'observations. Plus les droites seront différentes l'une de l'autre plus grande est la probabilité de le reconnaître et donc de prendre la bonne décision.

La capacité à reconnaître comme différentes des droites qui le sont vraiment est fortement tributaire du nombre d'observations. Ainsi avec 100 observations on pourra avoir du mal à distinguer une pente de 1 vs 1.2. Avec 1000 observations, ceci pourrait au contraire être facile comme le montrent les exemples suivants (à gauche, $F=0.81$, $F_{2,196}=3.04$ à droite $F=15.2$, $F_{2,1996}=3.00$).

