Régression linéaire multiple : lecture en pratique

I. Rappel
II. Réalisation d'une régression multiple
III. Notions pour l'interprétation
IV. Exemples

Rappels sur la régression linéaire multiple

- Tableau de données :
  - Individus 1 à n (ex : j)
  - Variables Y, X_1 à X_k (dont X_i)

Variables / Individus	Y	X_1	X_2	…	X_i	…
1						
2						
…						
j	y_j	$x_{1,j}$	$x_{2,j}$		$x_{i,j}$	
…						
n						

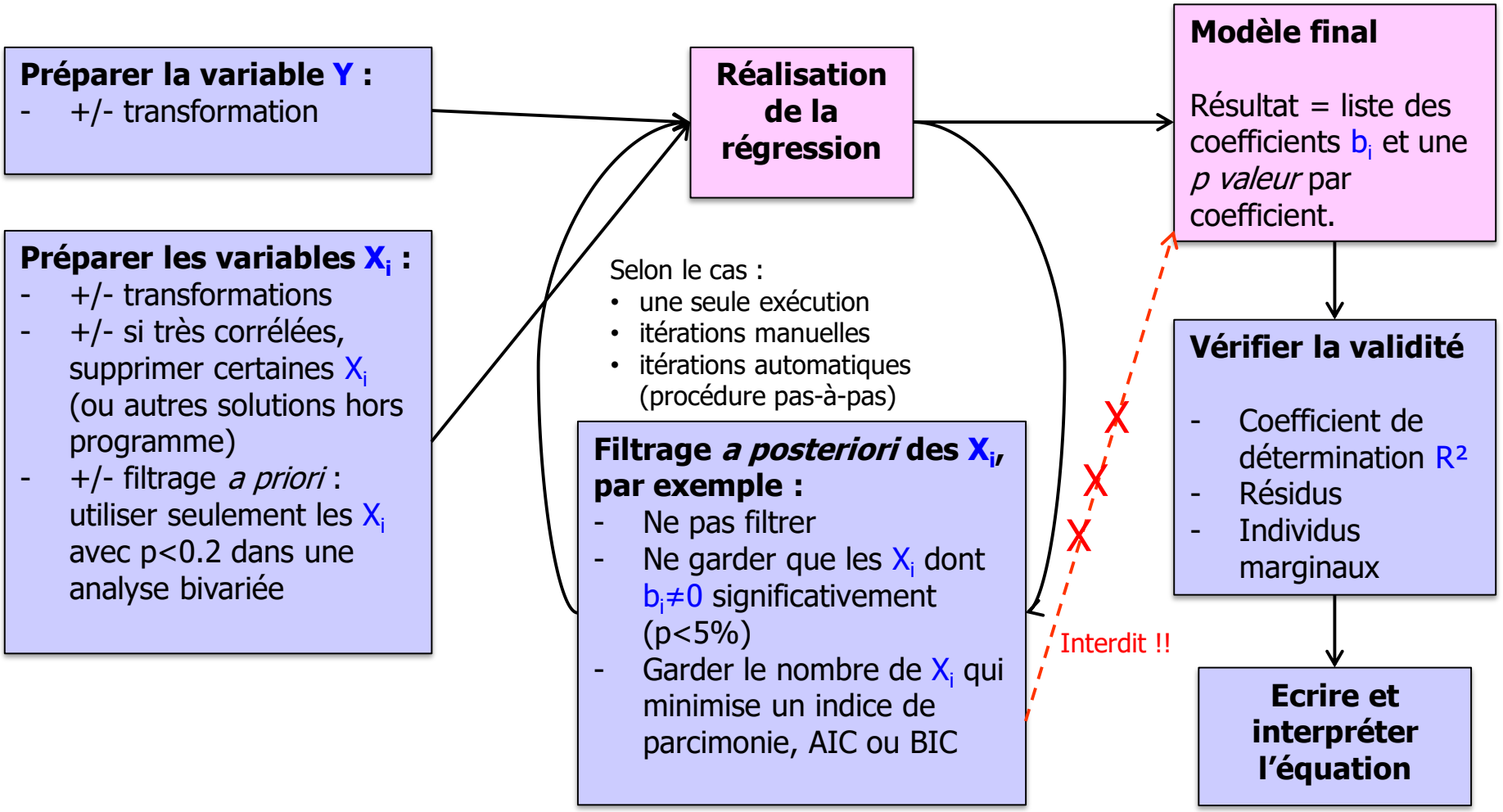
Rappels sur la régression linéaire multiple

- Méthode supervisée multivariée « phare » en médecine :
  - **Explication :** dans un jeu de données, expliquer une variable Y quantitative par des variables X_i quantitatives ou binaires
    Effet « ajusté » des variables X_i sur Y (isole l'effet propre de chaque X_i, sauf si les X_i sont fortement corrélées entre elles)
  - **Prédiction :** ensuite seulement, Y étant inconnue, prédire la valeur $\hat{y}_j$, avec intervalle de confiance, d'un nouvel individu j dont les valeurs $x_{i,j}$ sont connues
- Procédé (exemple avec 3 variables X_1, X_2 et X_3) :
  - Mise au point du modèle explicatif
    $Y=b_0 + b_1.X_1 + b_2.X_2 + b_3.X_3+\varepsilon$
  - Possibilité de prédiction avec la formule
    $\hat{Y}=b_0 + b_1.X_1 + b_2.X_2 + b_3.X_3$
  - Erreur de prédiction observée dans l'échantillon
    = résidu = $\hat{Y}-Y$

Rappels sur la régression linéaire multiple

- Variable Y, quantitative :
  - Dite « à expliquer » (ou « dépendante », « exogène », « réponse », « diagnostique »)
  - Fonctionne avec distribution quelconque, transformations possibles (ex : log)
- Variables X_i, quantitatives ou binaires :
  - Dites « explicatives » (ou « indépendantes », « endogènes », « prédicteurs »)
  - Fonctionne avec distribution quelconque, transformation aussi possible
- Risques de cette méthode (développés ci-après) :
  - Si Y peu lié aux X_i : faible pouvoir explicatif du modèle
    - Diagnostic : calculer le coefficient R^2
  - Si utilisation de trop de variables explicatives X_i : surajustement
    - Solution : calculer un indice de parcimonie et utiliser une procédure pas-à-pas
  - Si certaines variables X_i trop corrélées entre elles : modèle instable
    - Solution : explorer les corrélations entre variables
  - Si relation des X_i sur Y non linéaire : modèle inadapté
    - Diagnostic : regarder la distribution des résidus
  - Si présence d'individus trop influents : estimation du modèle faussée
    - Diagnostic : calculer la distance de Cook de chaque individu

Réalisation en pratique



Résultat d'une régression

- Les résultats sont simples :
  - Liste des coefficients b_i
  - Et pour chacun : p valeur du test de H_0 : $b_i=0$
  - Autrement dit, ces variables X_i ont un effet significatif si $p<5\%$ (l'équation n'est vraie qu'en les prenant tous, mais on peut très bien tenter une nouvelle régression en filtrant les variables X_i sur ce critère)
  - « intercept » = une variable qui vaudrait toujours « 1 », son coefficient est la constante b_0 du modèle

Paramètre	Coefficient	p valeur
Intercept	-114	0.005
X1	0.308	< 0.0001
X2	2.68	0.33

- Exemple :
  - Expliquer Y par X_1 et X_2
  - Modèle :
    $Y = -114 + 0.308*X_1 + 2.68*X_2$
  - X_1 est significativement associée à Y. Effet ajusté : en moyenne, chaque fois que X_1 augmente de 1, Y augmente de 0.308
  - X_2 n'est pas significativement associée à Y. Effet ajusté : en moyenne, chaque fois que X_2 augmente de 1, Y augmente de 2.68

Le coefficient de détermination R^2 Les indices de parcimonie

- Signification de R^2 :
  - = part de la variance de Y expliquée par le modèle
  - = part de la variance de Y retrouvée dans $\hat{Y}$
  - Qualité de l'ajustement, « goodness of fit »
- Interprétation :
  - Valeur de 0% (si modèle non explicatif) à 100% (si prédiction parfaite)
  - Dans le cas de modèle que nous étudions ici, $R^2=r^2=(Corr(Y, \hat{Y}))^2$ [1]
- Notion de parcimonie
  - En ajoutant des X_i, on améliorera souvent R^2 mais risque de surajustement
  - Critères de parcimonie AIC (Akaike information criterion) et BIC (bayesian information criterion) : traduisent la complexité du modèle par rapport à sa valeur explicative [2]
  - Pour choisir quelles X_i conserver : on peut minimiser AIC ou BIC (fait dans les procédures pas-à-pas, qui sélectionnent automatiquement les X_i à conserver, en les testant toutes)

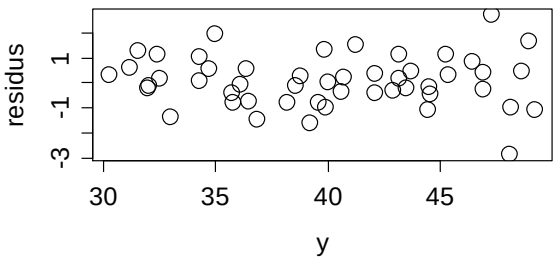
[1] ce n'est pas vrai pour toutes les régressions
[2] retenir « parcimonie » mais pas AIC et BIC

Visualisation des résidus =(Y prédit – Y observé)

Tracer les graphiques suivants :

- QQ-plot ou plus simplement histogramme des résidus
  - Moyenne nulle
  - Distribution d'allure normale
- Résidus en fonction de $\hat{Y}$
  - Moyenne ne dépend pas de $\hat{Y}$
  - Variance ne dépend pas de $\hat{Y}$ (homoscédasticité)
- (Résidus en fonction de chaque X)
  - Moyenne ne dépend pas de X
  - Variance ne dépend pas de X

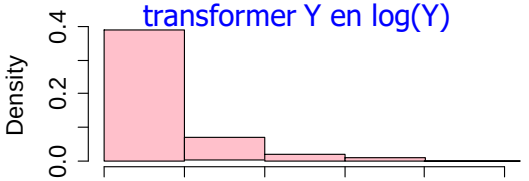
Exemple



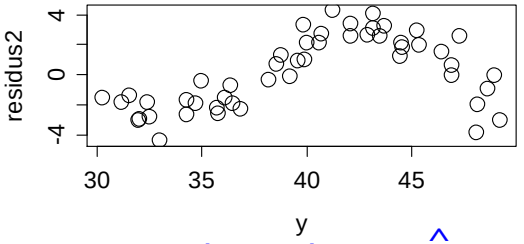
idem en fonction de X

Contrexemple

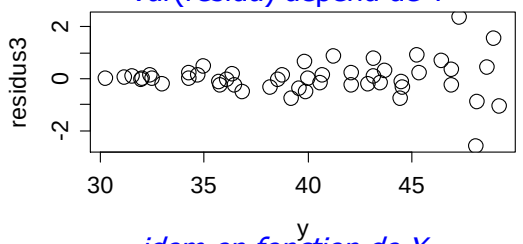
Résidu « lognormal » : essayer de transformer Y en log(Y)



E(résidu) dépend de $\hat{Y}$



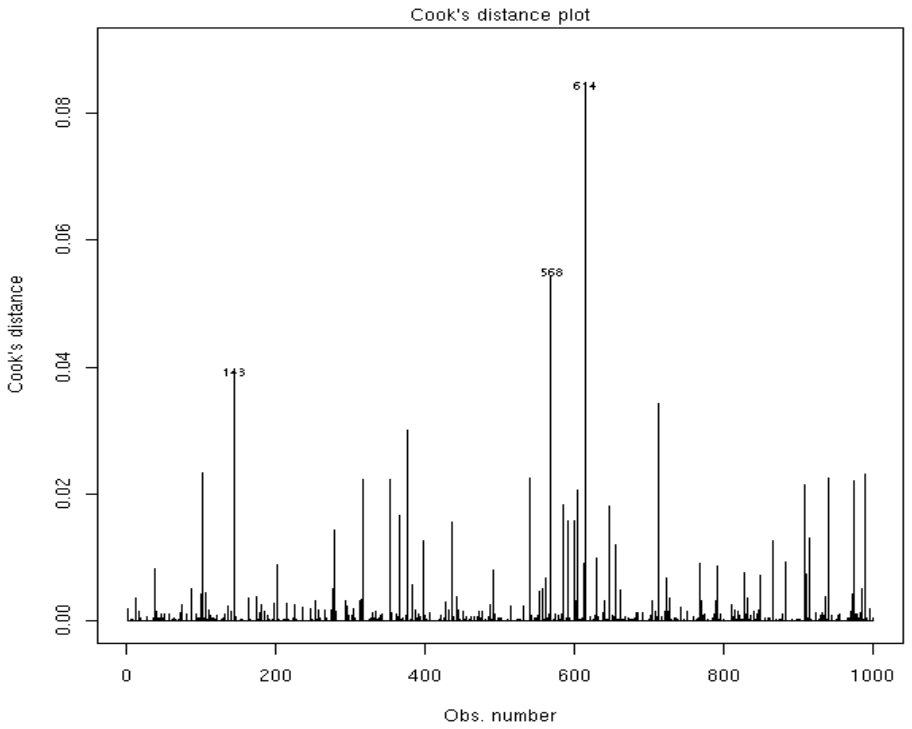
Var(résidu) dépend de $\hat{Y}$



idem en fonction de X

Identification d'individus trop influents

- Distance de Cook :
  - Calculée pour chaque individu (chaque point)
  - Evalue la différence entre la régression réalisée, et une régression réalisée en supprimant cet individu
    (en réalité, calculée sans refaire tourner la régression)
  - Distance élevée ⇔ point influent
  - On peut décider de supprimer ces individus trop influents



- Exemple :
  - X=numéro de l'individu
  - Y=distance de Cook
  - Quelques individus trop influents

Fiche d'interprétation d'une régression linéaire multiple déjà faite par un autre

Que regarder :

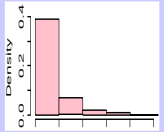
- Ajustement :
  - Souhaiter un R^2 élevé (0-1)
  - Discuter la parcimonie
- Ecrire le modèle
  - /!\ pour écrire, garder tous les coefficients, même les non-significatifs
  - Interpréter le modèle
  - /!\ ne pas conclure que les variables avec coefficients non-significatifs sont indépendantes de Y
  - Recherche multi-colinéarité
- Analyse des résidus :
  - Distribution normale
  - En fonction de $\hat{Y}$:
    - Moyenne indépendante de $\hat{Y}$
    - Variance indépendante de $\hat{Y}$ (homoscedasticité)
  - (parfois en fonction de chaque X)
- Recherche d'individus influents
  - Distance Cook des individus

Problèmes fréquents :

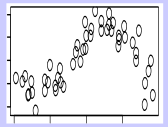
R^2 faible => régression peu utile

Trop de Xi => risque de surajustement

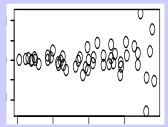
Variables Xi très corrélées (multicolinéarité) => pouvoir prédictif OK, mais mauvaise compréhension des relations entre X_i et Y



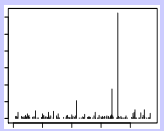
Distribution non-normale => transformer Y



Moyenne des résidus dépend de $\hat{Y}$ => relation non-linéaire



Variance des résidus dépend de $\hat{Y}$ (hétéroscédasticité) => relation non-linéaire

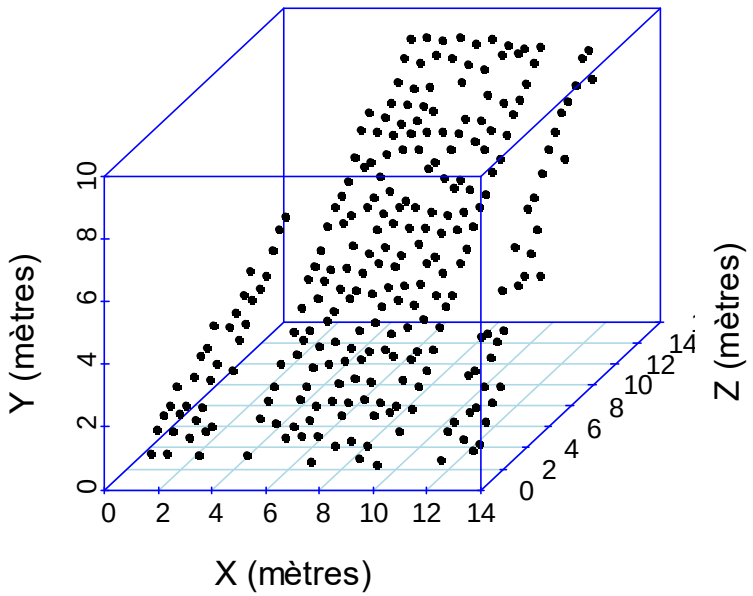


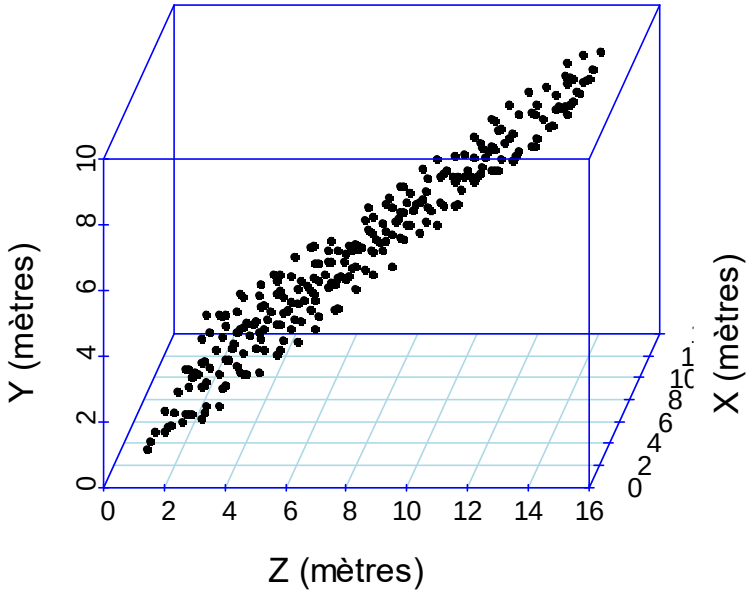
Quelques individus trop influents => discuter leur éviction

Exemple n°1
Etudiants dans un amphithéâtre

Données étudiées = position des étudiants (tête) dans un amphi

- Régression linéaire multiple, 256 étudiants
- Variable à expliquer :
  - Y=hauteur en mètres (hauteur du banc de la rangée + distance bassin-tête)
- Variables explicatives :
  - X= position gauche-droite en mètres
  - Z=position devant-derrière en mètres





Résultat de la régression Y ~ X+Z

- Résultat (R^2=0.997, AIC=-231):

```
Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept) -0.156585   0.029201   -5.364  2.03e-07 ***
x           -0.002853   0.000144   -0.907     0.365
z            0.670606   0.000480  270.425   < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Nom des variables (« intercept » représente la constante)

« estimate » =Coefficients

« Pr(…) » =Significativité (seuil : 5.0e-02)

- Interprétation
  - Sans surprise, excellent ajustement (R^2)
  - Modèle (arrondi) : Y= -0.157 -0.00285*X +0.671*Z
  - Le « résidu » est dans ce cas en rapport avec la hauteur bassin-tête des étudiants
  - Variable X non significative : il faudrait la supprimer => deuxième régression

Résultat de la régression Y ~ Z

- Résultat (R^2=0.997, AIC=-232):

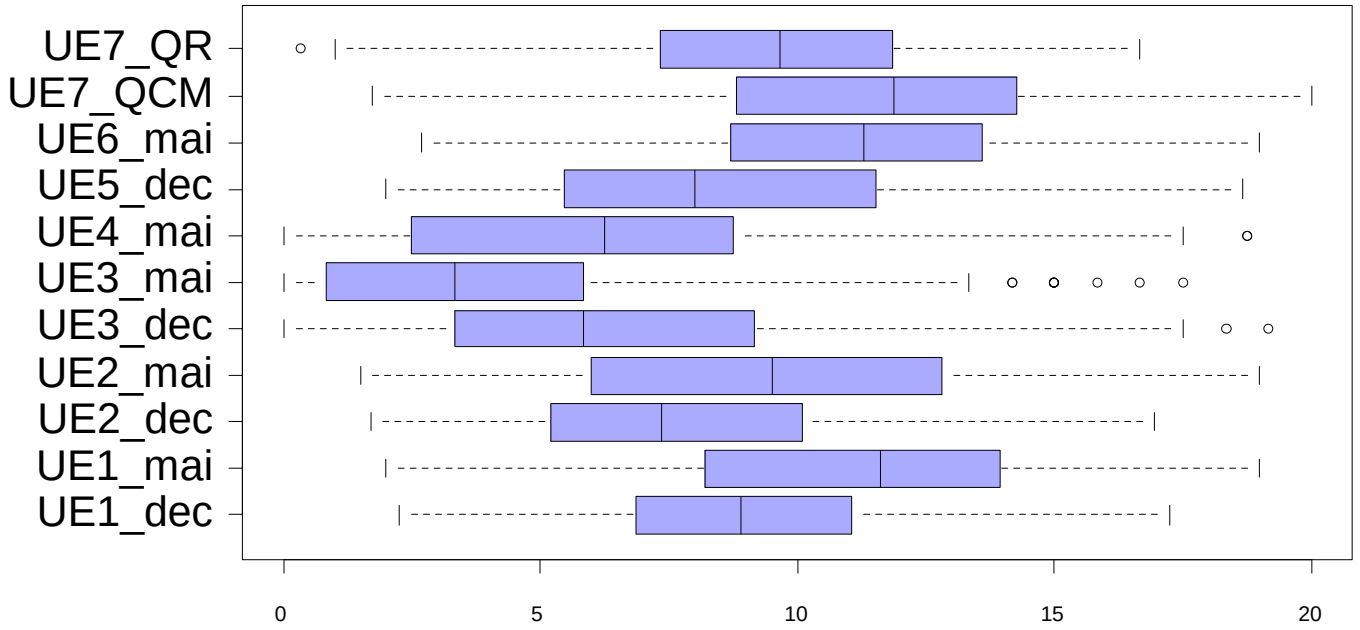
```
Coefficients:
             Estimate   Std. Error   t value   Pr(>|t|)
(Intercept) -0.175852     0.020184    -8.712    3.9e-16 ***
z            0.670290     0.002454   273.102    < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Interprétation
  - Qualité de l’ajustement inchangée (R^2 identique)
  - Critère de parcimonie amélioré (AIC plus bas)
  - Modèle (arrondi) : Y= -0.176 +0.670*Z
- Pour ce pseudo-exemple, les autres étapes de validation (résidus, Cook, corrélations) seront ignorées.

Exemple n°2
Notes au tronc commun de PACES

Données étudiées = notes aux épreuves du tronc commun PACES

- ~2500 étudiants ayant passé toutes ces épreuves
- Variable à expliquer :
  - Y=note finale au tronc commun
- Variables explicatives :
  - X_1 à X_{11} : 1 ou 2 notes sur 20 par UE du Tronc Commun PACES
- Finalité : retrouver les coefficients des épreuves



Résultat de la régression

- Résultat (R^2=1, AIC=-32800):

```
Coefficients:
             Estimate  Std. Error    t value  Pr(>|t|)
(Intercept) 9.364e-05   2.951e-05      3.173   0.00153 **
UE1_dec     1.177e-01   3.951e-06  29781.101   < 2e-16 ***
UE1_mai     5.882e-02   3.524e-06  16693.658   < 2e-16 ***
UE2_dec     7.843e-02   4.431e-06  17700.464   < 2e-16 ***
UE2_mai     3.921e-02   3.534e-06  11094.739   < 2e-16 ***
UE3_dec     5.882e-02   2.429e-06  24219.116   < 2e-16 ***
UE3_mai     5.883e-02   2.869e-06  20502.981   < 2e-16 ***
UE4_mai     5.882e-02   2.169e-06  27121.510   < 2e-16 ***
UE5_dec     1.765e-01   3.236e-06  54540.169   < 2e-16 ***
UE6_mai     1.177e-01   3.609e-06  32602.403   < 2e-16 ***
UE7_QCM     5.882e-02   3.248e-06  18111.220   < 2e-16 ***
UE7_QR      1.765e-01   3.034e-06  58165.932   < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Interprétation
  - On lit directement les coefficients des épreuves. Exemple : 5.8% pour l’UE4.
  - Le seul aléa est lié aux arrondis, négligeables => ajustement parfait (R^2=1)

Exemple n°3
Facteurs expliquant le poids des enfants

Données étudiées = poids des enfants

- Régression linéaire multiple, 237 enfants de 11 à 20 ans
- Variable à expliquer :
  - Y=poids en Kg
- Variables explicatives :
  - Sexe (binaire, 0=F, 1=H)
  - Age en mois
  - Taille en cm

```
     sexe             age              taille           poids       
 Min.   :0.0000   Min.   :139.0   Min.   :128.3   Min.   :22.73  
 1st Qu.:0.0000   1st Qu.:148.0   1st Qu.:149.4   1st Qu.:38.25  
 Median :1.0000   Median :163.0   Median :156.2   Median :45.45  
 Mean   :0.5316   Mean   :164.4   Mean   :155.9   Mean   :45.59  
 3rd Qu.:1.0000   3rd Qu.:178.0   3rd Qu.:163.3   3rd Qu.:50.40  
 Max.   :1.0000   Max.   :250.0   Max.   :182.9   Max.   :77.17  
```

Résultat de la régression poids ~ sexe+age+taille

- Résultat (R^2=0.631, AIC=1473.6):

```
Coefficients:
             Estimate  Std. Error  t value   Pr(>|t|)
(Intercept) -57.69399     5.51875   -10.45    < 2e-16 ***
sexe         -0.15188     0.72179    -0.21      0.834
age           0.10721     0.02522     4.25   3.09e-05 ***
taille        0.55005     0.04733    11.62    < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Interprétation
  - R^2=0.631 bonne qualité d'ajustement
  - On lit le modèle suivant :
    poids_kg = -57.7 - 0.152*sexe + 0.107*age_mois + 0.550*taille_cm
  - Tous ces coefficients sont significativement différents de zéro, sauf pour le sexe (p=83.4%>5%)
  - Notez que les garçons sont plus lourds que les filles, mais c'est parce qu'ils sont plus grands. A taille et âge égaux, le sexe n'a donc plus d'effet. Notion d'« ajustement ».
  - Il est approprié de relancer la régression sans le sexe

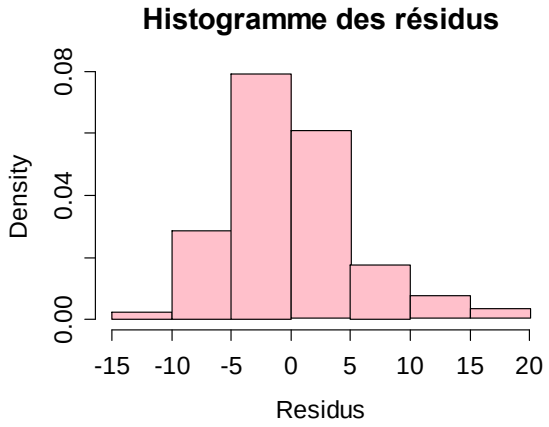
Résultat de la régression poids ~ age+taille

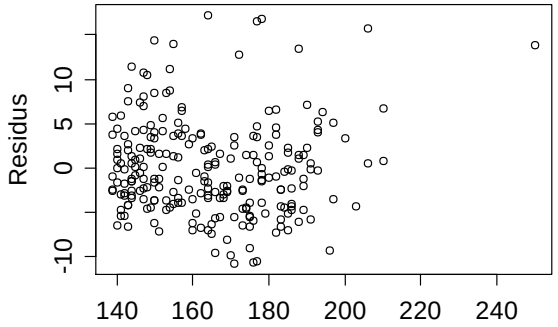
- Résultat (R^2=0.630, AIC=1471.6):

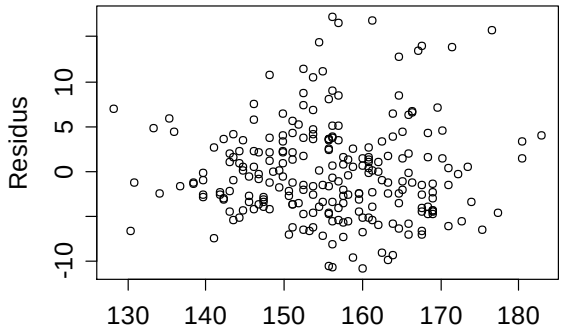
```
Coefficients:
             Estimate  Std. Error   t value    Pr(>|t|)
(Intercept) -57.51896     5.44455    -10.56     < 2e-16 ***
age           0.10812     0.02480      4.36    1.95e-05 ***
taille        0.54745     0.04559     12.01     < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

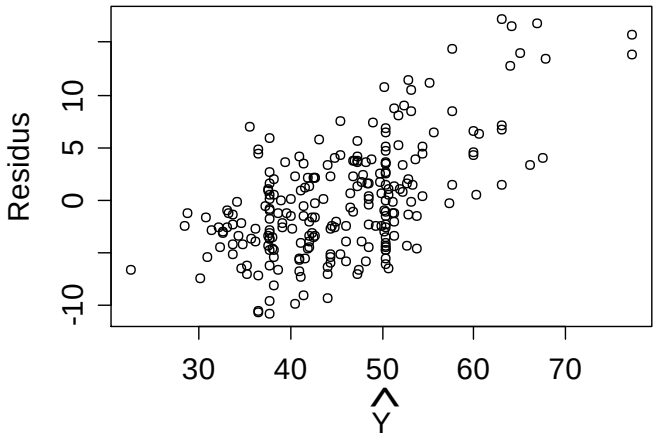
- Interprétation
  - On lit le modèle suivant (coefficients légèrement changés, tous signif. ≠0) : poids_kg = -57.5 + 0.108*age_mois + 0.547*taille_cm
  - R^2 à peine diminué, mais AIC également diminué : modèle plus parcimonieux
  - Âge significativement associé au poids. Toutes choses égales par ailleurs, pour chaque mois supplémentaire, poids augmenté en moyenne de 0.108kg. Donc pour chaque année supplémentaire, poids augmenté en moyenne de 1.30kg
  - Taille significativement associée au poids. Toutes choses égales par ailleurs, pour chaque cm supplémentaire, poids augmenté en moyenne de 0.547kg.

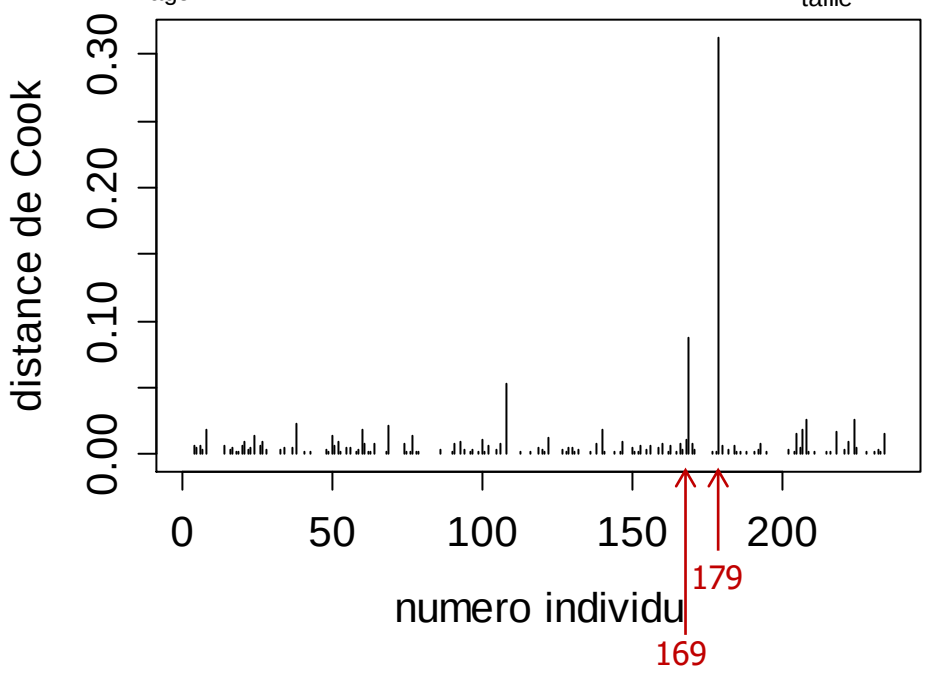
Diagnostic du deuxième modèle











Diagnostic du deuxième modèle

- Histogramme des résidus :
  - Limite mais acceptable
- Résidus en fonction de Y :
  - Homoscedasticité : OK
  - Les résidus augmentent avec la valeur de $\hat{Y}$. Le poids augmenterait moins pour les valeurs élevées des variables explicatives. Hypothèse de linéarité imparfaite donc.
- Résidus en fonction de l'âge :
  - OK dans l'ensemble
  - on note qu'un individu est nettement plus âgé que les autres.
- Résidus en fonction de la taille :
  - OK
- Distance de Cook :
  - les individus 179 et 169 sont trop influents
  - il est licite de relancer la régression sans eux

Exemple n°4
Je voudrais peser mes patients avec un mètre ruban !

Données étudiées = données anthropométriques

- 507 hommes et femmes
- Variables à expliquer :
  - Y = poids en kg
- 14 Variables explicatives :
  - Sexe (binaire, 0=F, 1=H)
  - Taille en cm
  - Tours en cm : épaule, poitrine, taille, nombril, hanche, cuisse, biceps, avant-bras, genou, mollet, cheville, poignet

Résultat de la régression

poids ~ tour_epaule + tour_poitrine + tour_taille + tour_nombril + tour_hanche + tour_cuisse + tour_biceps + tour_avantbras + tour_genou + tour_mollet + tour_cheville + tour_poignet + taille + sexe

- Résultat (R^2=0.973, AIC=2254):

```
Coefficients:
                Estimate Std. Error   t value   Pr(>|t|)
(Intercept)   -122.04195    2.66430   -45.806    < 2e-16 ***
tour_epaule      0.08450    0.02991     2.825   0.004914 **
tour_poitrine    0.19382    0.03566     5.435   8.63e-08 ***
tour_taille      0.36426    0.02741    13.291    < 2e-16 ***
tour_nombril    -0.01240    0.02396    -0.517   0.605075
tour_hanche      0.22961    0.04360     5.266   2.09e-07 ***
tour_cuisse      0.29047    0.05285     5.496   6.26e-08 ***
tour_biceps      0.07594    0.08579     0.885   0.376481
tour_avantbras   0.58569    0.13965     4.194   3.25e-05 ***
tour_genou       0.29186    0.07736     3.773   0.000181 ***
tour_mollet      0.39859    0.06992     5.700   2.06e-08 ***
tour_cheville   -0.00109    0.09976    -0.011   0.991290
tour_poignet    -0.10790    0.19547    -0.552   0.581181
taille           0.33890    0.01657    20.452    < 2e-16 ***
sexe            -0.99262    0.52583    -1.888   0.059654 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Interprétation
  - R^2=0.973 excellente prédiction du poids
  - Semble incohérent : par exemple, le tour de taille serait significatif mais pas le tour de nombril ?

Multicolinéarité

- Problème typique de multicolinéarité :
  - Plusieurs variables sont extrêmement corrélées entre elles
  - Le modèle « choisit » de faire porter toute la liaison à l'une d'entre elles
    - Aspect « tout ou rien » des p valeurs dans le tableau des coefficients
    - => pouvoir prédictif de la régression bien conservé
  - Ce choix est fluctuant au gré des échantillonnages
    - => intérêt explicatif de la régression remis en question : en multivarié, un coefficient non significatif ne veut pas dire absence de relation linéaire !
- Solutions à ce problème :
  - Diagnostic : matrice de corrélation 2 à 2 des variables X_i
  - Plusieurs solutions possibles (en choisir une seule) :
    - dans chaque groupe de variables similaires, n'en garder qu'une
    - laisser la machine choisir les variables en utilisant une procédure pas à pas « Stepwise »

Etude des corrélations entre X_i

- Les variables « non significatives » sont fortement corrélées à d'autres « significatives »
  - tour nombril : notamment corrélé à 85% au tour de hanche
  - tour biceps : notamment corrélé à 94% au tour d'avant-bras
  - tour cheville : notamment corrélé à 76% au tour de mollet
  - tour poignet : notamment corrélé à 90% au tour d'avant-bras
- Interprétation :
  - Ces variables ne sont plus nécessaires au modèle une fois les autres déjà prises en compte. Elles sont dans l'ombre d'autres variables. Elles sont « éclipsées ».
  - Attention : ces variables ne sont pas indépendantes de Y !
- Exemple de conduite à tenir :
  - On pourrait supprimer, dans chaque couple, le tour le moins fréquemment recueilli (c'est une option parmi d'autres)

Régression n°2 : filtrage humain

poids ~ tout sauf tour_nombril tour_biceps tour_cheville tour_poignet

- Résultat (R^2=0.973, AIC=2247):

```
Coefficients:
                Estimate Std. Error  t value   Pr(>|t|)
(Intercept)   -122.58889    2.55775  -47.928    < 2e-16 ***
tour_epaule      0.09026    0.02903    3.109   0.001983 **
tour_poitrine    0.19626    0.03409    5.758   1.50e-08 ***
tour_taille      0.35982    0.02452   14.675    < 2e-16 ***
tour_hanche      0.21615    0.03886    5.562   4.38e-08 ***
tour_cuisse      0.31408    0.04789    6.559   1.36e-10 ***
tour_avantbras   0.62535    0.09978    6.267   7.99e-10 ***
tour_genou       0.27480    0.07421    3.703   0.000237 ***
tour_mollet      0.38997    0.06356    6.136   1.74e-09 ***
taille           0.33535    0.01615   20.762    < 2e-16 ***
sexe            -0.89385    0.49751   -1.797   0.073002 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Interprétation
  - R^2=0.973 inchangée, AIC diminué => meilleure parcimonie
  - Cette option semble satisfaisante : la prédiction sera bonne
  - Ne surtout pas penser que les variables que nous avons supprimées ne sont pas utiles à la prédiction du poids ! Elles ne sont pas utiles lorsque les autres sont disponibles seulement.
  - Précisons ici que les étapes suivantes du diagnostic sont favorables (Cook, résidus)

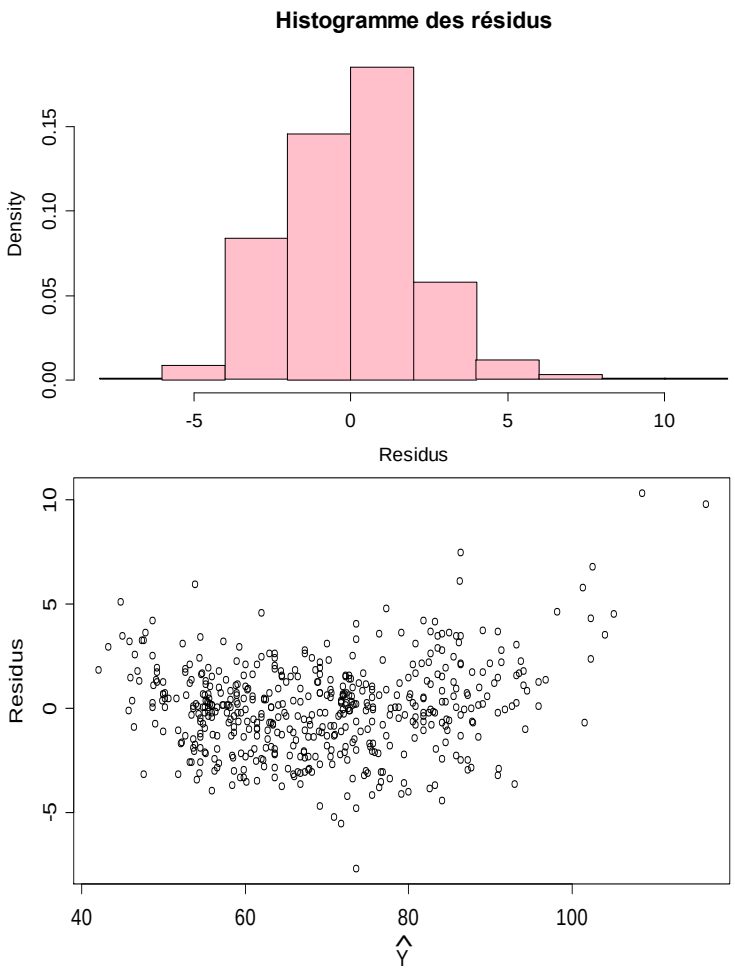
Régression n°3 : Procédure pas-à-pas, « Stepwise »

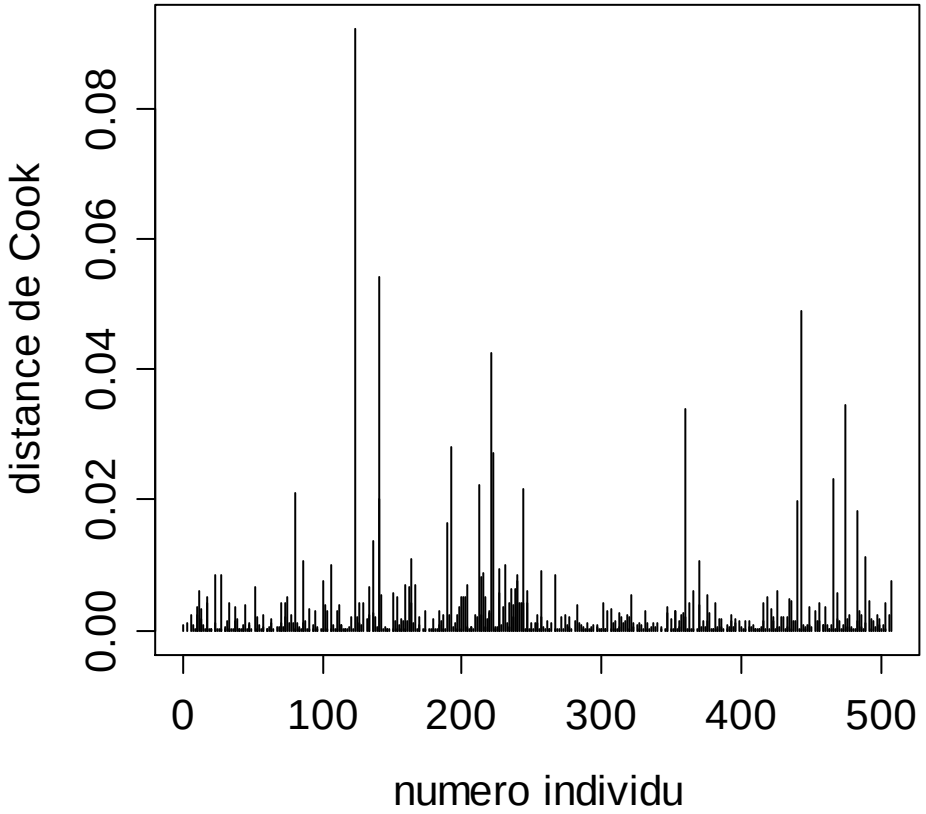
- Résultat (R^2=0.973, AIC=2247):

```
Coefficients:
                Estimate Std. Error  t value   Pr(>|t|)
(Intercept)   -122.58889    2.55775  -47.928    < 2e-16 ***
tour_epaule      0.09026    0.02903    3.109   0.001983 **
tour_poitrine    0.19626    0.03409    5.758   1.50e-08 ***
tour_taille      0.35982    0.02452   14.675    < 2e-16 ***
tour_hanche      0.21615    0.03886    5.562   4.38e-08 ***
tour_cuisse      0.31408    0.04789    6.559   1.36e-10 ***
tour_avantbras   0.62535    0.09978    6.267   7.99e-10 ***
tour_genou       0.27480    0.07421    3.703   0.000237 ***
tour_mollet      0.38997    0.06356    6.136   1.74e-09 ***
taille           0.33535    0.01615   20.762    < 2e-16 ***
sexe            -0.89385    0.49751   -1.797   0.073002 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Interprétation
  - Dans ce cas-là, par chance, on a retrouvé le même modèle qu’en supprimant à la main les 4 variables… avec l’embarras du choix en moins

Diagnostic des régressions n°2 et n°3 (identiques dans ce cas)





Bref, rien de rédhibitoire

Régression n°3 : Procédure pas à pas « Stepwise »

- Invocation
  - Pas-à-pas descendant
  - (il existe aussi le pas-à-pas ascendant, et le bidirectionnel)
- Synopsis exécuté par la machine :
  - Modèle avec 14 variables X => AIC=2254.16
  - *Essai de suppression de chacune des 14 variables (14 nouvelles régressions) => la suppression du tour de cheville diminue le plus l'AIC*
  - Modèle sans tour_cheville => AIC=2252.16
  - *(idem sur les 13 variables restantes)*
  - Modèle sans tour_cheville et tour_nombril => AIC=2250.44
  - *(idem sur les 12 variables restantes)*
  - Modèle sans tour_cheville, tour_nombril et tour_poignet => AIC=2248.76
  - *(idem sur les 11 variables restantes)*
  - Modèle sans tour_cheville, tour_nombril, tour_poignet et tour_biceps => AIC=2247.38
  - *Ensuite, toute autre tentative fait remonter l'AIC, la procédure s'arrête donc là.*

En synthèse

- Procédure statistique TRES utilisée en recherche médicale
- Deux intérêts :
  - Explication des relations : identifie l'effet des variables explicatives « ajusté » sur les autres variables explicatives
  - Prédiction d'une valeur Y inconnue
- Dangers à identifier et +/- contourner :
  - Multicolinéarité des variables Xi (corrélations, procédure pas-à-pas)
  - Faible pouvoir explicatif (coefficient de détermination)
  - Surajustement (procédure pas-à-pas)
  - Caractère inapproprié du modèle linéaire (transformation des variables, analyse des résidus)
  - Influence excessive de quelques individus (distance de Cook)

/!\ Association, causalité…

- Une variable peut être associée sans être causale
  - Ex : pointure de chaussure => quotient intellectuel
- Une variable avec un coefficient non-significativement différent de zéro n'est pas forcément indépendante
  - Ex : association non-linéaire
- En multivarié, une variable avec un coefficient non-significativement différent de zéro n'est pas forcément non-associée linéairement
  - Ex : variable éclipsée par une autre variable qui lui est très corrélée