Régression linéaire multiple : lecture en pratique

Principes de la régression linéaire multiple

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

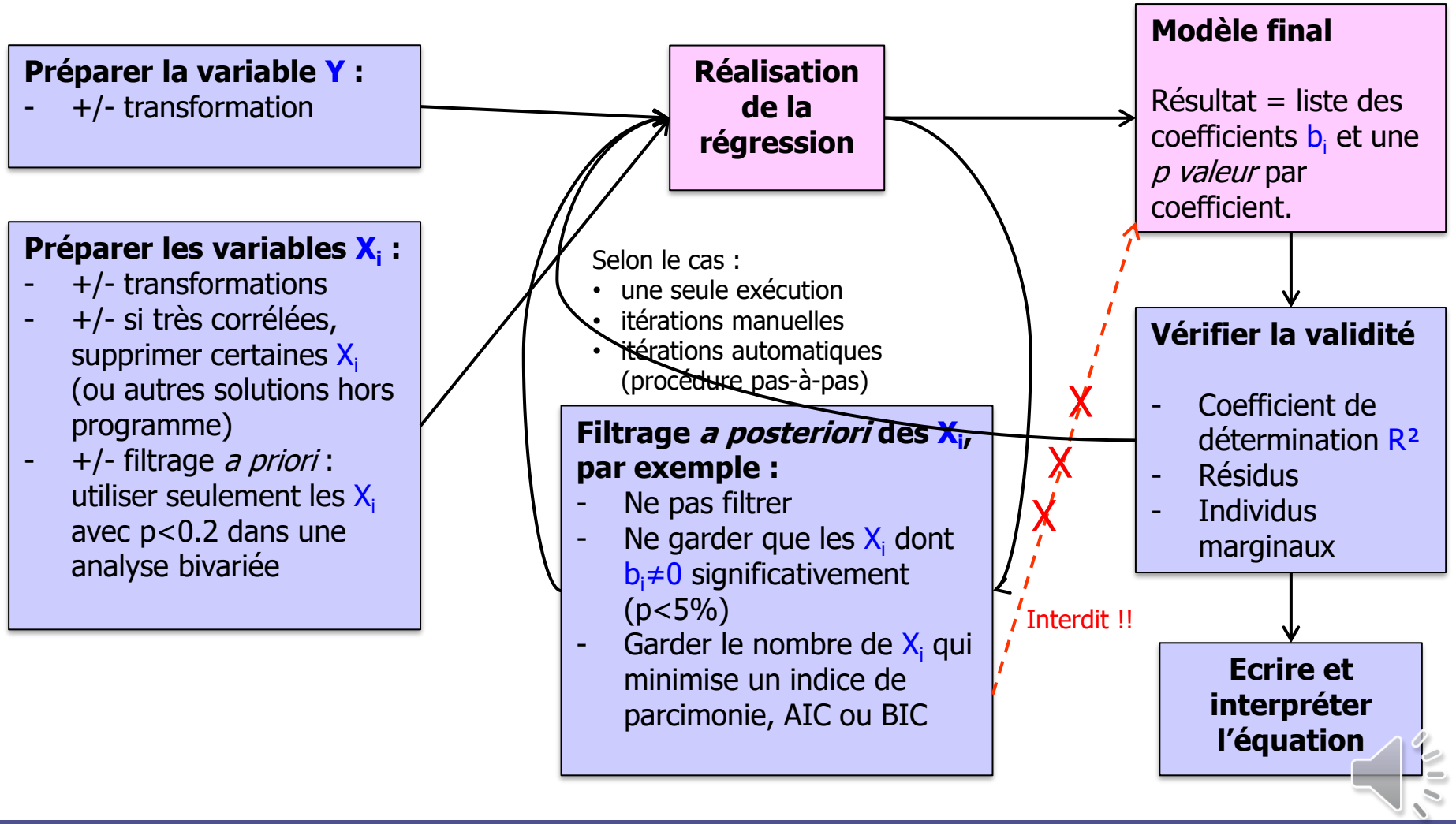
Principes de la régression linéaire multiple

- 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 = $\hat{Y}-Y$

Principes de 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 : un processus itératif !



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 pseudo-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)

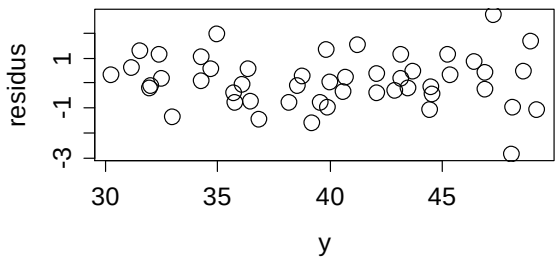
[1] ce n'est pas vrai pour toutes les régressions
[2] retenir « parcimonie » mais pas AIC et BIC

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

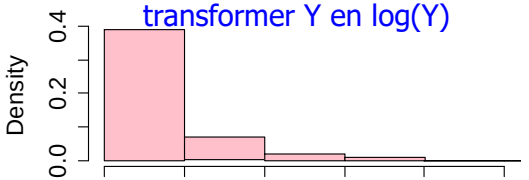
Exemple



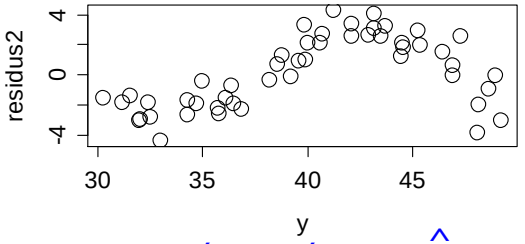
idem en fonction de X

Contrexemple

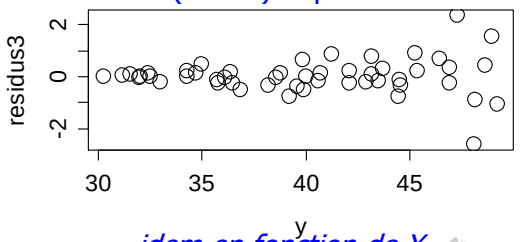
Résidu « lognormal » : essayer de transformer Y en log(Y)



E(résidu) dépend de $\hat{Y}$



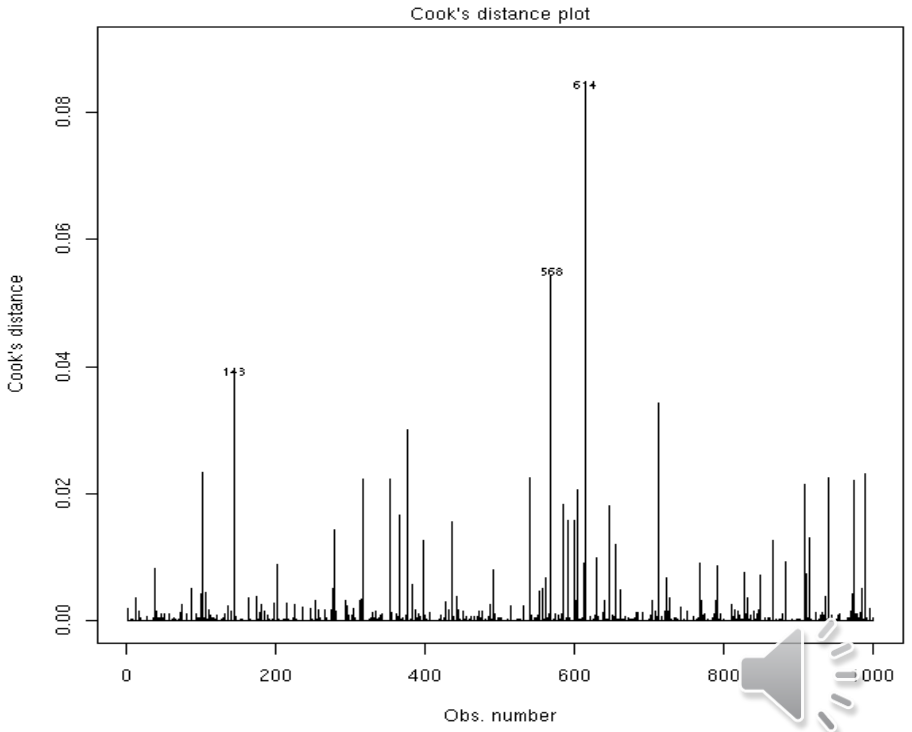
Var(résidu) dépend de $\hat{Y}$



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

Multicolinéarité

- Problème de multicolinéarité :
  - Lorsque plusieurs variables sont très corrélées entre elles (ex : tour de taille et tour de nombril)
  - 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 »

Procédure pas-à-pas ou « Stepwise »

- Exemple de pas-à-pas descendant sur une régression à 14 vairables :
  - Modèle avec 14 variables X => AIC=2254.16
  - *Essai de suppression de chacune des 14 variables (14 nouvelles régressions) => la suppression de la variable X_4 diminue le plus l'AIC*
  - Modèle sans X_4 => AIC=2252.16
  - *(idem sur les 13 variables restantes)*
  - Modèle sans X_4 et X_9 => AIC=2250.44
  - *(idem sur les 12 variables restantes)*
  - Modèle sans X_4, X_9 et X_7 => AIC=2248.76
  - *Ensuite, toute autre tentative de suppression d'une variable fait remonter l'AIC, la procédure s'arrête donc là.*
- Types de pas-à-pas
  - Ceci illustre le pas-à-pas descendant
  - Il existe aussi le pas-à-pas ascendant, et le bidirectionnel
- Avantage : choix automatique, le plus efficace en termes statistiques
- Inconvénients : pas forcément le choix le plus pertinent

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

Que regarder :

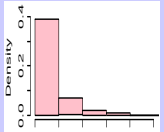
- Ajustement :
  - Souhaiter un R^2 élevé (0-1)
  - Discuter la parcimonie
- Ecrire le modèle
  - /!\ pour écrire, garder tous les coefficients, même les non-significatifs
  - Interpréter le modèle
  - /!\ ne pas conclure que les variables avec coefficients non-significatifs sont indépendantes de Y
  - 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 :

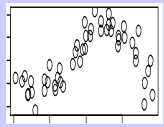
R^2 faible => régression peu utile

Trop de X_i => risque de surajustement

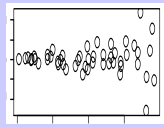
Variables X_i très corrélées (multicolinéarité) => pouvoir prédictif OK, mais mauvaise compréhension des relations entre X_i et Y



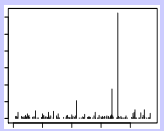
Distribution non-normale => transformer Y



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

/!\ 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

En synthèse

- Procédure statistique TRES utilisée en recherche médicale

- Attention toutefois :
  - Hypothèse d'additivité des effets pas toujours pertinente
  - Effets conditionnels plus ou moins pris en compte par des « interactions »
  - Penser aussi aux arbres de régression, qui ne s'appuient sur aucune hypothèse particulière