Cours 4 - Corréaltions

Regression Linéaire	
B0 et B1 dans le graphique	b0 : la valeur prédite de Y quand X vaut 0 b1 : # d'unités d'augmentation de Y quand X augmente d'une unité
Les 4 tableaux de sortie	1. var introduites et éliminés 2. Coefficient 3. ANOVA 4. Récap des moèles
Coefficients	
On réalise un test T pour chaque paramètre-paramètre	H0 de b1 et b2 = 0 n'indique pas si le modèle = bon Nous dit seulement si b0 et b1 sont différents de 0
ANOVA	
Pq réaliser ANOVA	Une ANOVA est réalisée pour tester si le modèle explique mieux les données que le modèle de base: la moyenne des valeurs de la variable Y. Ce modèle de base reflète l'hypothèse nulle de cette ANOVA
H0 de l'ANOVA	Meilleur prédicteur de Y = moyenne de y h0 = moy de Y calcul SC (pire qu'on puisse faire) Correspond à la somme des carrés total (SCT) , tout ce que le modèle doit expliquer
Hypothèse alternative de l'ANOVA	On utilise meilleures valeurs des éstimateurs b0 et b1 On SC du modèle (SCRésid) ce que le modèle n'a pas réussi à expliquer
SCM (régression)	SCT - SCR (ce qu'il y a a expliquer - ce que le modèle n'a pas réussi à expliquer)

Cours 4 - Corréaltions (cont)

ANOVA test stat F	Le résultat du test F permet de rejeter le modèle de base (Y = Ybarre) au profit du modèle alternatif avec une probabilité d'erreur inférieure à 0.1% si H est vraie
Degrés de liberté (k= prédict)	dll total= N – 1, ddl résid = N -1 -K, ddl modèle = k
Problème ANOVA	Le test est basé sur les carrés moyens (CM) du modèle et des résiduels Or, les CMR dépend de la taille de l'échantillon + N est grand, plus CMR diminue , plus CMR petit, plus F= grand (facile rejet H0)
Taille d'effet	Pour calculer R2, on calcule la proportion de la SCT expliquée par le modèle (SCM) - Le modèle permet d'expliquer approximativement 20 % de la variabilité totale dans le modèle de base (Y = Yl).
Prob de R2	+ k est grand par rapport à N, + on risque d'expliquer de la variabilité correspondant à de l'erreur d'échantillonnage, plutôt que d'expliquer un effet réel dans la population.- Donc, + k est grand par rapport à N, - l'explication des données se généralisera à l'ensemble de la population.
R2 ajusté	Plus k est grand par rapport à N, plus la valeur de Rajust est petite (corrigée)
Corrélation	

Cours 4 - Corréaltions (cont)

Corr	r tjrs positif dans tab (signe de b1) - racine de r2
Cov donne 2 infos	1. Sens d'inclinaison 2. degré d'aplatissement du nuage de point
Corrélation règle 2 probs de la cova	1. Plus intuitif (score z) 2. quantifiable

Cours 5 - Regression linéaire multiple

Regression linéaire multiple	
Représentation graphique	b0 : val prédite de y quand x1 et x2 =0 ; $b1$ correspond au nombre d'unités d'augmentation de la valeur prédite Y qnd la valeur du prédicteur X1 augmente d'une unité et que les valeurs des autres prédicteurs ne changent pas.
SCR (pente y pred et y)	on a deux variables indépendantes qui ont chacunes un pente : permet de prédire la variable y (deux droites)
SCT	Formule .
On peut ensuite faire test F et Rdeux	
SCM	Ce qui a été expliqué = var qu'il y ' a à expliqué – ce qu'il reste a expliquer une fois qu'on a le modèle
Absence de Multicolinéarité	On ne veut pas que les prédicteurs soient trop fortement corrélés Si les prédicteurs sont fortement corrélés, alors deux variables deviennent interchangeables et il devient difficile d'interpréter le modèle final (On souhaite avoir rX1,X2 < 0.9)

Cours 5 - Regression linéaire multiple (cont)

Corrélation simple	proportion dela variance totale en **Y** expliqué par x (ce que x explique / tout ce qu'il y a à expliquer) on oublie que x2 existe
Corrélation semi partielle	proportion de la variance totale en y expliquée seulement par x (on eleve seulement la partie exlpiquée --point commun avec la simple = ce qu'il y a a expliquer)
Corrélation partielle	proportion de la variance en Y qui n'est pas expliquée par les autres prédicteurs mais qui est explqiuée par X1 (Partielle est tjrs plus grande ou égale à semi partielle
On regarde les types de corrélations3 pour eviter d'avoir des données redondantes	

Cours 5 - Regression linéaire multiple (cont)

différence entre les coefficients standardisés et non standardisés.	Un coefficient non standardisé (ex. b1) permet de prédire le nombre d'unités de variation de la variable dépendante (y) pour une variation d'une unité de la variable indépendante (x1). Par exemple, la note prédite à l'examen final monte de 10 points par heure d'étude (mesures brutes utilisées). Un coefficient standardisé (ex. b1_standardisé) permet de prédire le nombre d'écarts types de variation de la variable dépendante (y) pour une variation d'un écart type de la variable indépendante (x1). Par exemple, la note prédite à l'examen final monte de 2 écarts types par augmentation d'une écart type d'heures d'études (mesures exprimées en écarts à la moyenne divisés par l'écart type).
il peut avoir une grande correlation entre x1 et x2, mais la partie de y qui recoupe les deux n'est pas la même	il peut avoir une grande correlation entre x1 et x2, mais la partie de y qui recoupe les deux n'est pas la même (-correlation simple et semi = similaire) semi partielle nous permet de savoir si les deux expliquent la même chose ou pas (comparer avec simple)
Stats de colinéarité	VIF et tolérance

Cours 5 - Regression linéaire multiple (cont)

Pour Calculer VIF1 :	1. On calcule SCT à partir du modèle $X1$ =xbarre 2. On trouve le modèle qui prédit le mieux X1 à partir des autres prédicteurs 3.On calcule SCR1. 4. On calcule SCM1 = SCT – SCR 5. On calcule R^2 6.VIF (La tolérance est simplement 1 sur VIF
Si on a seulement 2 prédicteurs : Ø Alors, R^2 correspond simplement au carré de la corrélation bivariée.	
Crièter VIF	on veut une tolerance plus garnde que 0,2 alors un r2 inférieur à 0.8 on veut pas etre capable d'expliquer plus de 80% de la variabilité de x1 en fonction des autres variables (On souhaite avoir chaque VIF < 5 (i.e. une Tolérance > 0.2).Sinon, on doit considérer éliminer le prédicteur correspondant.)
Colinéarité : Si un prédicteur peut être très bien prédit par les autres prédicteurs, alors il est inutile.	Il ne permet pas d'augmenter substantiellement SCModèle dans la prédiction de Y. Il augmente le nombre de degrés de libertés (k) de SCMModèle dans la prédiction de Y. DONC... CMModèle diminue !Et F aussi
Scores extrêmes	
Résidus standardisés	les scores z des résidus (i.e. de la variance non expliquée par le modèle) Les scores appelés « extrêmes », sont « extrêmes » dans la distribution des résidus (y et ybarre)

Cours 5 - Regression linéaire multiple (cont)	
Les « Valeurs influentes » (« leverage »)	un score élevé signifie que la donnée a le potentiel d'avoir une influence importante sur l'estimation des paramètres du modèle. Les scores appelés « valeurs influentes », ont en réalité simplement le POTENTIEL d'avoir une influence importante sur l'estimation des paramètres du modèle...
La distance de Cook	reflète l'influence réelle d'une observation sur l'estimation des paramètres.
L'objectif principal est de pouvoir généraliser les conclusions de l'analyse à la population.	
cook	Si le modèle estimé est trop sensible à certaines données de notre échantillon (ex. Distance de Cook élevée) , alors le modèle risque de varier beaucoup d'un échantillon à l'autre. On dira alors que le modèle est instable et est peu reproductible (et ne se généralise donc pas bien à la population).

Cours 5 - Regression linéaire multiple (cont)	
Zrésid et val influentes	Si un score est extrême (extrême dans les résidus) ou a un potentiel d'influence (extrêmedans les prédicteurs), alors la donnée semble peu représentative de la population. Si Dcook est néanmoins faible alors que le score est extrême (extrême dans les résidus) et/ou potentiellement influent (extrême dans les prédicteurs), alors garder le score risque d'augmenter artificiellement la puissance de l'analyse en augmentant le N à l'aide d'une valeur non représentative de la population.
Score nn extreme et non influent	Semble représentatif mais ...Si Dcook est néanmoins élevé alors que le score n'est ni extrême (non extrême dans les résidus) ni potentiellement influent (non extrême dans les prédicteurs), alors garder le score risque de rendre le modèle estimé instable et peu reproductible (et donc peu généralisable à la population de toute manière).
Les scores extrêmes (résidus) et les données avec potentiel d'influence (prédicteurs) posent un risque pour la représentativité du modèle et donc pour la généralisation des conclusions. Les données d'influence réelle (ex. distance de Cook) posent un risque pour la stabilité du modèle et donc pour la généralisation des conclusions.	
On doit donc toujours vérifier si le modèle avec plus de prédicteurs augmente « significativement » le R2.	Pour ce faire, on utilise un test F de la « Variation du R2 ».