

### Cours 4 - Corrélations

#### 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 modèles

#### 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 estimateurs 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 à expliquer - ce que le modèle n'a pas réussi à expliquer)

### Cours 4 - Corrélations (cont)

ANOVA test stat F  
 Le résultat du test F permet de rejeter le modèle de base ( $Y = \bar{Y}$ ) au profit du modèle alternatif avec une probabilité d'erreur inférieure à 0.1% si H est vraie

Degrés de liberté (k=prédicte)  
 ddl 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 ( $Y = \bar{Y}$ ).

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 R ajusté est petite (corrigée)

#### Corrélation



### Cours 4 - Corrélations (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 cov	1. Plus intuitif (score z) 2. quantifiable

### Cours 5 - Regression linéaire multiple

#### Regression linéaire multiple

Représentation graphique  
 $b_0$  : val prédite de y quand  $x_1$  et  $x_2 = 0$  ;  $b_1$  correspond au nombre d'unités d'augmentation de la valeur prédite Y qnd la valeur du prédicteur  $X_1$  augmente d'une unité et que les valeurs des autres prédicteurs ne changent pas.

SCR  
 on a deux variables indépendantes qui ont chacune une 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 expliquer – ce qu'il reste à expliquer une fois qu'on a le modèle

Absence de Multicollinéarité  
 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  $r_{X1, X2} < 0.9$ )

### Cours 5 - Regression linéaire multiple (cont)

Corrélation simple  
 proportion de la variance totale en Y expliqué par x (ce que x explique / tout ce qu'il y a à expliquer)  
 on oublie que  $x_2$  existe

Corrélation semi partielle  
 proportion de la variance totale en y expliquée seulement par x (on élève seulement la partie expliquée --point commun avec la simple = ce qu'il y a à expliquer)

Corrélation partielle  
 proportion de la variance en Y qui n'est pas expliquée par les autres prédicteurs mais qui est expliquée par  $X_1$  (Partielle est tjrs plus grande ou égale à semi partielle)

On regarde les types de corrélations<sup>3</sup> pour éviter d'avoir des données redondantes

### Cours 5 - Regression linéaire multiple (cont)

différence entre les coefficients standardisés et non standardisés. Un coefficient non standardisé (ex.  $b_1$ ) permet de prédire le nombre d'unités de variation de la variable dépendante ( $y$ ) pour une variation d'une unité de la variable indépendante ( $x_1$ ). Par exemple, la note prédite à l'examen final monte de 10 points par heure d'étude (mesures brutes utilisées). Un coefficient standardisé (ex.  $b_1$ \_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 ( $x_1$ ). 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 corrélation entre  $x_1$  et  $x_2$ , mais la partie de  $y$  qui recoupe les deux n'est pas la même (-corrélation simple et semi = similaire) semi partielle nous permet de savoir si les deux expliquent la même chose ou pas (comparer avec simple)

il peut avoir une grande corrélation entre  $x_1$  et  $x_2$ , mais la partie de  $y$  qui recoupe les deux n'est pas la même

Stats de colinéarité VIF et tolérance

### Cours 5 - Regression linéaire multiple (cont)

Pour Calculer VIF1 :

1. On calcule SCT à partir du modèle  $X_1 = \bar{x}$
2. On trouve le modèle qui prédit le mieux  $X_1$  à partir des autres prédicteurs
3. On calcule SCR1.
4. On calcule  $SCM_1 = SCT - SCR$
5. On calcule  $2.6.VIF$  (La tolérance est simplement 1 sur VIF)

Si on a seulement 2 prédicteurs :  $\emptyset$  Alors,  $R^2$  correspond simplement au carré de la corrélation bivariée.

Critère VIF on veut une tolérance plus grande que 0,2 alors un  $r^2$  inférieur à 0.8 on veut pas être capable d'expliquer plus de 80% de la variabilité de  $x_1$  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  $SC_{Modèle}$  dans la prédiction de  $Y$ . Il augmente le nombre de degrés de liberté ( $k$ ) de  $SC_{Modèle}$  dans la prédiction de  $Y$ . DONC...  $C_{Modèle}$  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  $\bar{y}$ )



### 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.

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)

Si un score est extrême (extrême dans les résidus) ou a un potentiel d'influence (extrême dans les prédicteurs), alors la donnée semble peu représentative de la population. Si  $D_{cook}$  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 non extrême et non influent Semble représentatif mais ... Si  $D_{cook}$  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  $R^2$ . Pour ce faire, on utilise un test F de la « Variation du  $R^2$  ».

