

Mes notes de cours

Thomas Husson

Table des matières

1 Mes notes de cours	4
2 Introduction	5
3 Résumé statistique	6
4 Position et dispersion	7
1 Position	7
1.A Moyenne	7
1.B Médiane	7
1.C Mode	7
2 Dispersion	7
2.A Étendue = empan	7
2.B Écart interquartile (IQR)	8
2.C Écart-type	8
2.D Variance	8
3 Exemple sur R	9
5 Analyses en sous groupe	11
1 Principe	11
2 Dans R	11
6 Dépendance, liaison et association	15
1 Variables quantitatives	15
1.A Dépendance	15
1.B Dépendance monotone ou linéaire	15
1.B.1 Exemple sur R	16
1.B.1.1 SMP	16
1.B.1.2 Données simulées	18
1.B.1.3 Matrice de corrélation	20
1.C Concordance	22
1.C.1 Principe	22
1.C.2 Exemple sur R	23
1.D Résumé des paramètres de dépendance entre deux variables quantitatives	25
2 Variables catégorielles	26
2.A Dépendance	26
2.A.1 Chi2 et associés	26
2.A.2 Odds-ratio et risque relatif	27
2.A.3 Exemple sur R	27
2.B Dépendance monotone	28
2.B.1 Exemple sur R	29
2.C Concordance	30
2.C.1 Coefficient kappa de Cohen	30
2.C.2 Sensibilité, spécificité, VPP, VPN	30
2.C.2.1 Exemple 1 sur R	31
2.C.2.2 Exemple 2 sur R	32
7 6 Effet centre	34
8 Introduction	35
9 Modèle linéaire	36
1 Analyse naïve sans prise en compte du centre	36
1.A Problèmes posés par cette analyse naïve	37
1.B Exemple R avec jeu de données fictif	38

2	Analyse avec prise en compte de l'effet centre	40
2.A	Exemple R	41
3	Analyse intra-centres (modèles conditionnels, mixtes ou non)	41
3.A	Modèle linéaire avec effet fixe par centre	42
3.B	Modèle linéaire avec pente commune (intercept variable par centre)	42
3.C	Modèle mixte avec effet aléatoire de centre	44
3.D	Exemple R	45
3.E	Conditions de validités des modèles mixtes	47
4	Modèle marginal = GEE (Generalized Estimating Equations)	47
4.A	Définition	47
4.B	Principe des GEE	48
10	Modèle linéaire généralisé	51
1	Principes généraux	51
2	Exemple R avec données simulées	51

1Mes notes de cours

2Introduction

Ceci est la page d'accueil du document. Le contenu pourra être complété plus tard.

3Résumé statistique

4Position et dispersion

1 Position

Paramètres de position = valeurs qui résument la tendance centrale d'une distribution.

- Moyenne
- Médiane
- Mode

1.A Moyenne

Moyenne = somme des valeurs divisée par le nombre de valeurs.

$$\frac{1}{n} \sum_{i=1}^n x_i$$

Correspond au centre de gravité des points si on les représente sur une droite.

Hypothèses :

- Les valeurs sont indépendantes
- Équivalence de la quantité (1 euro vaut 1 euro quelque soit sa position sur la droite des réels) : donc les notes c'est pas top en vrai !!
- Les valeurs sont continues

1.B Médiane

Signification plus directe : valeur qui partage la distribution en deux parties égales.

Si la distribution est symétrique, la moyenne et la médiane sont égales.

1.C Mode

Mode = valeur la plus fréquente.

2 Dispersion

Mesures de dispersion = valeurs qui résument la variabilité d'une distribution.

- Étendue = empan

2.A Étendue = empan

Correspond à la différence entre la valeur maximale et la valeur minimale.

2.B Écart interquartile (IQR)

$$= Q3 - Q1$$

2.C Écart-type

écart type = écart “le plus typique” par rapport à la moyenne.

Si m est la moyenne des observations x_i , l’écart type s est défini par la racine carrée de la variance :

$$s = \sqrt{[(x_1 - m^2) + \dots + (x_n - m)^2]/(n - 1)}$$

La variance correspond à la moyenne des carrés des écarts par rapport à la moyenne.

Donc en gros : écart type = racine carrée des carrés des écarts par rapport à la moyenne.

Pourquoi ajouter des carrés ?

- Positive les écarts négatifs
- Accentue les écarts importants
- Et pour une super propriété de l’écart-type :
 - La variance de deux variables indépendantes est égale à la somme de leurs variances.

Comment l’interpréter ?

- Dans le cas d’une distribution normale,
 - environ 2/3 des observations se situent à moins d’un écart-type de la moyenne.
 - environ la moitié des observations se situent à $[m - (2/3)s; m + (2/3)s]$

Avantages et inconvénients :

- Avantage :
 - utilise toutes les valeurs de la distribution
 - propriétés mathématiques intéressantes : la variance de la somme de deux variables indépendantes est égale à la somme de leurs variances.
 - s’utilise dans de nombreux tests statistiques (t-test, ANOVA, régression linéaire, etc.)
 - Facile à manipuler mathématiquement
 - Interprétation claire dans le cas de distributions normales
- Inconvénient :
 - sensible aux valeurs extrêmes (outliers)
 - ne s’interprète pas facilement dans le cas de distributions asymétriques
 - Pas adapté aux variables ordinaires ou catégorielles

2.D Variance

Variance = écart type au carré

ou moyenne des carrés des valeurs moins le carré de la moyenne.

3 Exemple sur R

Utilisation du jeu de données `smp.d` (version réduite de `smp`) et de la fonction `summary()`

```
summary(smp.d)
```

```
age                  profession      nb.enfants      depression
Min.   :19.00        ouvrier       :228    Min.   : 0.000  Min.   :0.0000
1st Qu.:28.00        sans.emploi  :221    1st Qu.: 0.000  1st Qu.:0.0000
Median :37.00        employé       :136    Median : 1.000  Median :0.0000
Mean   :38.94        commerçant   : 91    Mean   : 1.572  Mean   :0.3917
3rd Qu.:48.00        intermédiaire: 57    3rd Qu.: 2.000  3rd Qu.:1.0000
Max.   :84.00        (Other)       : 60    Max.   :14.000  Max.   :1.0000
NA's   :2            NA's         : 6     NA's   :26
schizophrenie      gravite       recherche.nouv evit.danger
Min.   :0.0000        Min.   :1.000  Min.   :1.000  Min.   :1.000
1st Qu.:0.0000        1st Qu.:2.000 1st Qu.:1.000  1st Qu.:1.000
Median :0.0000        Median :4.000  Median :2.000  Median :2.000
Mean   :0.0801        Mean   :3.635  Mean   :2.058  Mean   :1.865
3rd Qu.:0.0000        3rd Qu.:5.000 3rd Qu.:3.000  3rd Qu.:3.000
Max.   :1.0000        Max.   :7.000  Max.   :3.000  Max.   :3.000
NA's   :4            NA's   :104   NA's   :108
dep.recompense
Min.   :1.000
1st Qu.:1.000
Median :2.000
Mean   :2.152
3rd Qu.:3.000
Max.   :3.000
NA's   :114
```

Deux inconvénients à la fonction `summary()`:

- Ne donne pas l'écart-type
- La disposition des résultats n'est pas très claire.

On peut utiliser la fonction `describe()` du package `prettyR` pour un résumé plus complet.

```
describe(smp.d)
```

Description of `smp.d`

Numeric

	mean	median	var	sd	valid.n
age	38.94	37	175.72	13.26	797
nb.enfants	1.57	1	3.42	1.85	773
depression	0.39	0	0.24	0.49	799
schizophrenie	0.08	0	0.07	0.27	799
gravite	3.64	4	2.72	1.65	795
recherche.nouv	2.06	2	0.77	0.88	695
evit.danger	1.87	2	0.76	0.87	691
dep.recompense	2.15	2	0.69	0.83	685

Factor

profession	ouvrier	sans.emploi	employé	commerçant	intermédiaire	autre	cadre
Count	228.00	221.00	136.00	91.00	57.00	31.00	24
Percent	28.54	27.66	17.02	11.39	7.13	3.88	3

profession <NA> agriculteur

Count	6.00	5.00
Percent	0.75	0.63

Mode ouvrier

5Analyses en sous groupe

1 Principe

Dans un essai thérapeutique, il faut décrire les caractéristiques des patients inclus dans chaque groupe de traitement.

Il faut donc les décrire en fonction de différentes modalités (groupes de traitement, sexe, âge, etc.)

2 Dans R

On peut aussi utiliser la fonction `table()` pour faire des tableaux de contingence.

```
table(  
  smp.d$profession,  
  smp.d$depression,  
  deparse.level=2, # deparse.level fait apparaître les noms des variables dans  
  ↳ le tableau  
  useNA="ifany")
```

```
          smp.d$depression  
smp.d$profession  0   1  
  agriculteur     3   2  
  commerçant      65  26  
  cadre           16  8  
  intermédiaire   31  26  
  employé          81  55  
  ouvrier          132 96  
  autre            22  9  
  sans.emploi     132 89  
  <NA>             4   2
```

Si on voulait les pourcentages plutôt :

La fonction `prop.table()` permet de calculer des pourcentages à partir d'un tableau de contingence.

L'option `margin` permet de choisir si on veut les pourcentages par ligne (`margin=1`) ou par colonne (`margin=2`).

```
options(digits=3) # pour afficher 3 décimales  
prop.table(  
  table(  
    smp.d$profession,  
    smp.d$depression,  
    deparse.level=2, # deparse.level fait apparaître les noms des variables  
    ↳ dans le tableau
```

```

    useNA="ifany"),
margin=1) # margin=1 pourcentage par ligne ; margin=2 pourcentage par colonne

```

```

smp.d$depression
smp.d$profession      0      1
agriculteur     0.600 0.400
commerçant       0.714 0.286
cadre            0.667 0.333
intermédiaire   0.544 0.456
employé          0.596 0.404
ouvrier          0.579 0.421
autre             0.710 0.290
sans.emploi     0.597 0.403
<NA>              0.667 0.333

```

Mais encore une fois, je trouve personnellement que le top est d'utiliser `tbl_summary` du package `gtsummary`.

Il va falloir me convaincre de ne pas utiliser ce banger absolu : je ne vois pas pourquoi.

A la limite, pourquoi pas `tableone` aussi.

avec `tableone` : (fait des tests t pour les variables continues et Chi2 / Fisher pour les catégorielles par défaut)

```

library(tableone)
vars <- c("age","profession","nb.enfants", "gravite","recherche.nouv",
         "evit.danger","dep.recompense")
catVars <- c("profession","gravite","recherche.nouv",
            "evit.danger","dep.recompense")
table1 <- CreateTableOne(vars = vars, data = smp.d, factorVars = catVars, strata =
                           "depression")
print(table1, showAllLevels = TRUE, formatOptions = list(digits = 2))

```

Stratified by depression					
	level	0	1	p	test
n		486	313		
age (mean (SD))		39.93 (13.69)	37.41 (12.42)	0.009	
profession (%)	agriculteur	3 (0.6)	2 (0.6)	0.333	
	commerçant	65 (13.5)	26 (8.4)		
	cadre	16 (3.3)	8 (2.6)		
	intermédiaire	31 (6.4)	26 (8.4)		
	employé	81 (16.8)	55 (17.7)		
	ouvrier	132 (27.4)	96 (30.9)		
	autre	22 (4.6)	9 (2.9)		
	sans.emploi	132 (27.4)	89 (28.6)		
nb.enfants (mean (SD))		1.57 (1.92)	1.58 (1.73)	0.936	
gravite (%)	1	100 (20.7)	6 (1.9)	<0.001	

	2	111 (23.0)	18 (5.8)	
	3	82 (17.0)	33 (10.5)	
	4	89 (18.5)	74 (23.6)	
	5	68 (14.1)	114 (36.4)	
	6	26 (5.4)	55 (17.6)	
	7	6 (1.2)	13 (4.2)	
recherche.nouv (%)	1	168 (38.5)	81 (31.3)	0.005
	2	107 (24.5)	50 (19.3)	
	3	161 (36.9)	128 (49.4)	
evit.danger (%)	1	229 (52.8)	86 (33.5)	<0.001
	2	109 (25.1)	45 (17.5)	
	3	96 (22.1)	126 (49.0)	
dep.recompense (%)	1	121 (28.1)	71 (28.0)	<0.001
	2	148 (34.3)	49 (19.3)	
	3	162 (37.6)	134 (52.8)	

avec `gtsummary` :

- (attention il a tendance à faire des tests de Wilcoxon par défaut pour les variables continues, il faut lui dire de faire des t-tests si on veut ça)
- Pour les variables catégorielles, il fait par défaut des tests du Chi2 (ou Fisher si effectifs petits) donc autant ne pas lui donner d'instructions

```
# utiliser smp.d.bis avec la variable depression en facteur recodé en "Dépressif"
#   / "Non dépressif"
smp.d.bis <- smp.d
smp.d.bis$depression <- factor(smp.d.bis$depression, levels=c(0,1), labels=c("Non
#   dépressif","Dépressif"))

tableau <- smp.d.bis %>%
 tbl_summary(
  by = depression,
  statistic = list(
    all_continuous() ~ "{mean} ({sd})",
    # all_continuous() ~ "{median} [{p25}, {p75}]",
    all_categorical() ~ "{n} / {N} ({p}%)"
  ),
  digits = all_continuous() ~ 2,
  missing = "no"
) %>%
  modify_header(label = "***Caractéristiques**") %>%
  bold_labels() %>%
  add_overall() %>%
  add_p(
    # test = list( (rajouter si tests spécifiques pr les 2)
    all_continuous() ~ "t.test" # ou "wilcox.test" pour test non
    # paramétrique
    # all_categorical() ~ "chisq.test" # ou "fisher.test" si effectifs
    # petits
  )

```

```
)
```

```
# ajouter les p-values pour les comparaisons entre groupes
```

Juste un peu chiant pour avoir un bel affichage en pdf mais franchement...

```
tableau %>%
  # Conversion en objet kable (LaTeX standard)
  as_kable_extra(booktabs = TRUE, longtable = TRUE) %>%
  kableExtra::column_spec(1, width = "6cm") %>%
  # (Optionnel) Ajuste la taille de la police si le tableau est encore trop
  #   large
  kableExtra::kable_styling(latex_options = c("repeat_header"), font_size = 9)
```

Caractéristiques	Overall N = 799	Non dépressif N = 486	Dépressif N = 313	p-value
age	38.94 (13.26)	39.93 (13.69)	37.41 (12.42)	0.007
profession				
agriculteur	5 / 793 (0.6%)	3 / 482 (0.6%)	2 / 311 (0.6%)	
commerçant	91 / 793 (11%)	65 / 482 (13%)	26 / 311 (8.4%)	
cadre	24 / 793 (3.0%)	16 / 482 (3.3%)	8 / 311 (2.6%)	
intermédiaire	57 / 793 (7.2%)	31 / 482 (6.4%)	26 / 311 (8.4%)	
employé	136 / 793 (17%)	81 / 482 (17%)	55 / 311 (18%)	
ouvrier	228 / 793 (29%)	132 / 482 (27%)	96 / 311 (31%)	
autre	31 / 793 (3.9%)	22 / 482 (4.6%)	9 / 311 (2.9%)	
sans.emploi	221 / 793 (28%)	132 / 482 (27%)	89 / 311 (29%)	
nb.enfants	1.57 (1.85)	1.57 (1.92)	1.58 (1.73)	>0.9
schizoprenie	64 / 799 (8.0%)	29 / 486 (6.0%)	35 / 313 (11%)	0.008
gravite				<0.001
1	106 / 795 (13%)	100 / 482 (21%)	6 / 313 (1.9%)	
2	129 / 795 (16%)	111 / 482 (23%)	18 / 313 (5.8%)	
3	115 / 795 (14%)	82 / 482 (17%)	33 / 313 (11%)	
4	163 / 795 (21%)	89 / 482 (18%)	74 / 313 (24%)	
5	182 / 795 (23%)	68 / 482 (14%)	114 / 313 (36%)	
6	81 / 795 (10%)	26 / 482 (5.4%)	55 / 313 (18%)	
7	19 / 795 (2.4%)	6 / 482 (1.2%)	13 / 313 (4.2%)	
recherche.nouv				0.005
1	249 / 695 (36%)	168 / 436 (39%)	81 / 259 (31%)	
2	157 / 695 (23%)	107 / 436 (25%)	50 / 259 (19%)	
3	289 / 695 (42%)	161 / 436 (37%)	128 / 259 (49%)	
evit.danger				<0.001
1	315 / 691 (46%)	229 / 434 (53%)	86 / 257 (33%)	
2	154 / 691 (22%)	109 / 434 (25%)	45 / 257 (18%)	
3	222 / 691 (32%)	96 / 434 (22%)	126 / 257 (49%)	
dep.recompense				<0.001
1	192 / 685 (28%)	121 / 431 (28%)	71 / 254 (28%)	
2	197 / 685 (29%)	148 / 431 (34%)	49 / 254 (19%)	
3	296 / 685 (43%)	162 / 431 (38%)	134 / 254 (53%)	

¹ Mean (SD); n / N (%)

² Welch Two Sample t-test; NA; Pearson's Chi-squared test

6 Dépendance, liaison et association

Deux variables sont dépendantes si une valeur donne une information sur l'autre.

Par exemple, le poids et la taille sont dépendantes : connaître la taille d'une personne permet d'avoir une idée de son poids.

1 Variables quantitatives

3 types de liaisons entre variables quantitatives :

- Dépendance : connaître X permet de mieux estimer Y (par exemple tabac et maladie respiratoire, taille et poids, etc.)
- Dépendance monotone : X et Y varient dans le même sens (par exemple âge et pression artérielle)
 - Dépendance linéaire : relation linéaire entre X et Y (par exemple taille et poids chez les adultes)
 - Corrélation
 - Variance partagée = proportion de la variance de Y expliquée par X dans une relation linéaire entre les deux variables
- Concordance : si X est plus grand pour un individu que pour un autre, alors Y est aussi plus grand pour le premier individu (par exemple taille et poids)
 - pour une variable quantitative : coefficient de corrélation intraclassé (ICC)

1.A Dépendance

Il n'existe pas de paramètre estimant parfaitement la dépendance ou l'indépendance entre deux variables quantitatives.

L'idéal serait d'avoir un paramètre $\delta(X, Y)$ valant 0 quand X et Y sont indépendantes et 1 quand elles sont parfaitement dépendantes.

1.B Dépendance monotone ou linéaire

Le coefficient de corrélation de Pearson r mesure la dépendance linéaire entre deux variables quantitatives X et Y.

Il est désigné par les lettres r ou ρ (rho), en référence à Karl Pearson qui l'a introduit en 1895.

Il varie entre -1 (les deux variables X et Y sont parfaitement linéairement dépendantes de façon négative) et +1 (les deux variables X et Y sont parfaitement linéairement dépendantes de façon positive).

Corrélation nulle ($r = 0$) signifie que les deux variables sont linéairement indépendantes.

NB : le coefficient ne matérialise pas la **force** de la dépendance, mais seulement son **type** !

Pour matérialiser la force de la relation, on utilise le **coefficient de détermination r^2** .

Le paramètre r^2 représente approximativement la **proportion de la variance de Y** expliquée par la variance de X dans une relation linéaire entre les deux variables.

C'est à dire :

- X est le nombre d'heures de révision
- Y est la note obtenue à un examen

On trouve un coefficient de corrélation $r = 0.8$ entre X et Y , alors $r^2 = 0.64$ soit 64%.

Donc : 64% de la variabilité (de la variance) des notes Y s'expliquer par le modèle linéaire basé sur le nombre d'heures de révision X .

Ce n'est pas la même chose que “ $r^2\%$ des notes s'expliquent par le nombre d'heures de révision”.

! Important

Coefficient de corrélation r :

- mesure la **direction** de la relation linéaire entre deux variables quantitatives.
- varie entre -1 et +1 (avec 0 = pas de relation linéaire).
- en pratique : mesure un peu la force quand même... mais il n'y a pas vraiment d'unité pour l'exprimer

Coefficient de détermination r^2 :*

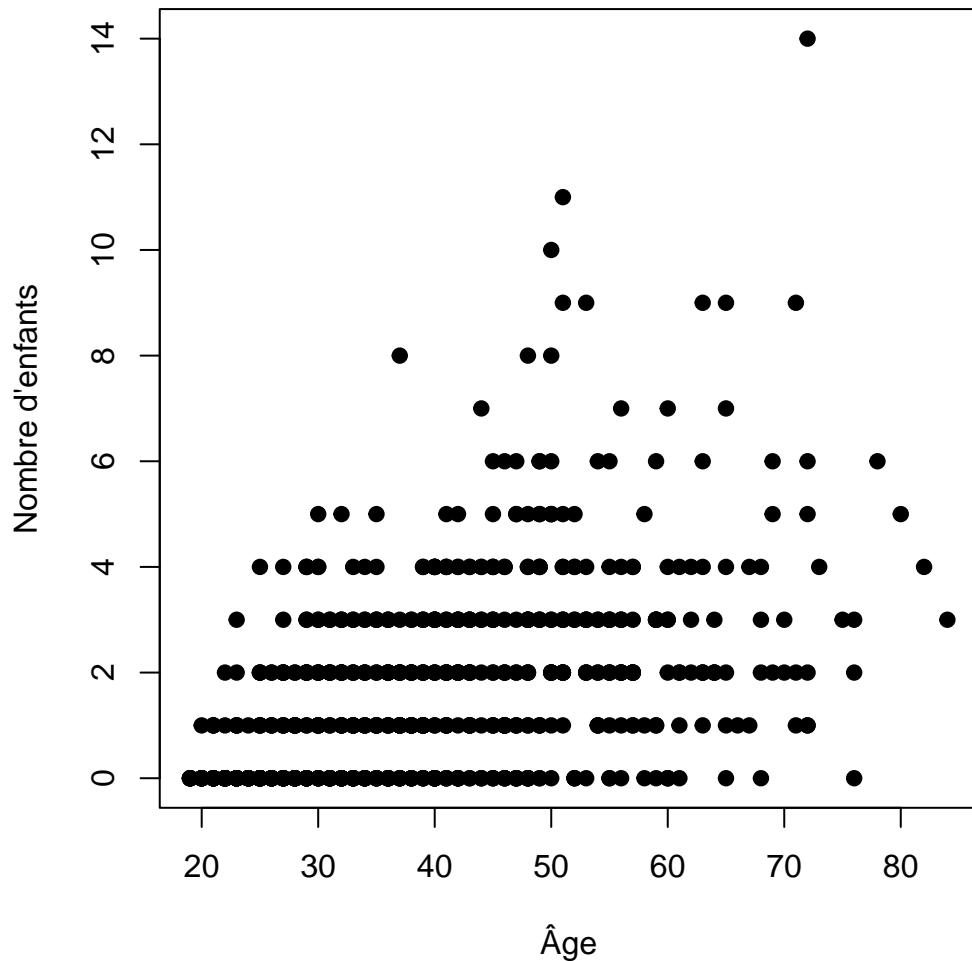
- mesure la **force** de la relation linéaire entre deux variables quantitatives avec une unité (en % de variance expliquée)
- varie entre 0 et 1 (avec 0 = pas de relation linéaire).
- représente la proportion de la variance de Y expliquée par la variance de X (et non pas le pourcentage de valeurs de Y expliquées par X).

1.B.1 Exemple sur R

1.B.1.1 SMP Représentation graphique de la relation entre l'âge et le nombre d'enfants dans le jeu de données **smp**.

```
plot(smp$age, smp$nb.enfants,
      xlab="Âge",
      ylab="Nombre d'enfants",
      main="Relation entre l'âge et le nombre d'enfants",
      pch=19) # pch = sert à choisir le type de point
```

Relation entre l'âge et le nombre d'enfants



Calcul du coefficient de corrélation de Pearson entre l'âge et le nombre d'enfants.

```
cor(  
  smp$age,  
  smp$nb.enfants,  
  use="complete.obs") # ignore les valeurs manquantes
```

```
[1] 0.498
```

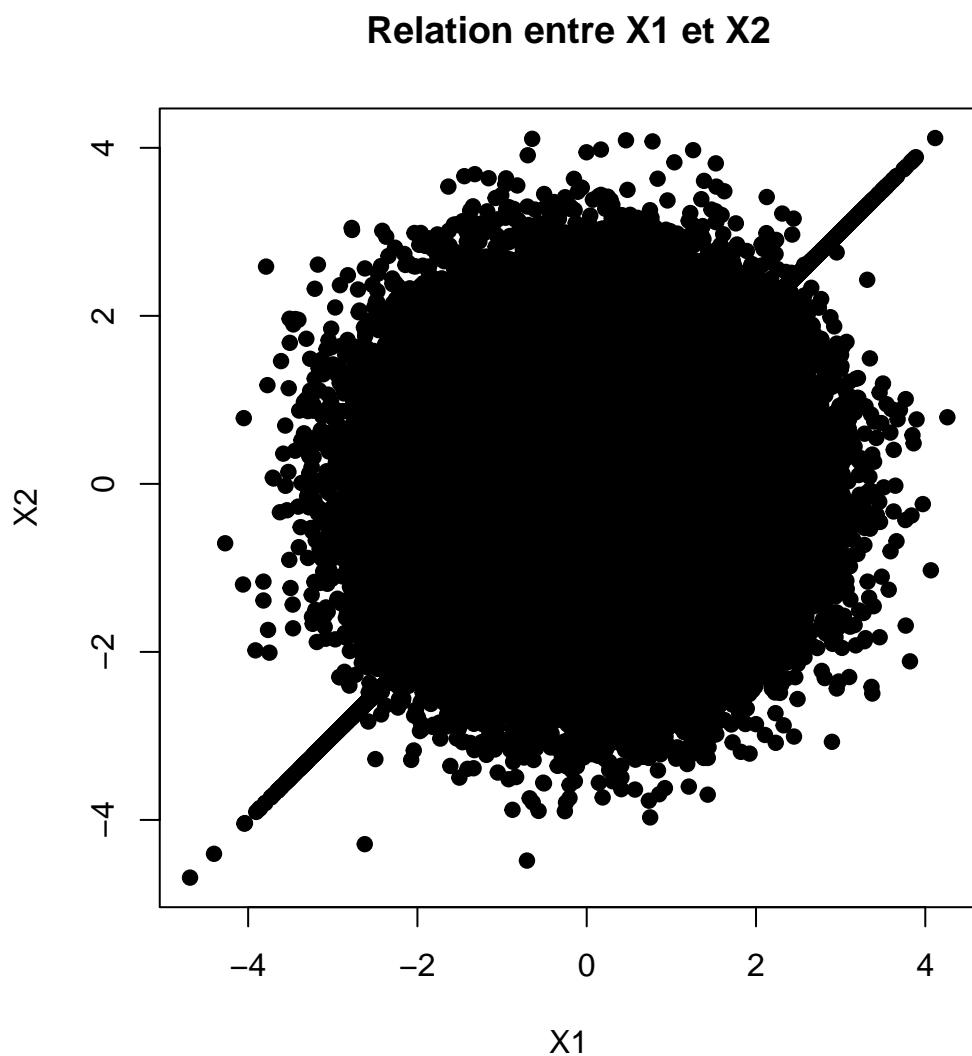
Le coefficient de corrélation est de 0.498, ce qui indique une dépendance linéaire positive entre l'âge et le nombre d'enfants.

```
set.seed(20230430)  
x <- rnorm(100000) # génère 100000 valeurs aléatoires suivant une loi normale  
y <- rnorm(100000)  
z <- rnorm(100000)
```

```
X1 <- c(x,y) # concatène les deux vecteurs x et y  
X2 <- c(x,z)
```

1.B.1.2 Données simulées Représentation graphique de la relation entre X1 et X2.

```
plot(X1, X2,  
      xlab="X1",  
      ylab="X2",  
      main="Relation entre X1 et X2",  
      pch=19)
```



Coorélation entre X1 et X2.

```
round(cor(X1, X2), 3)
```

```
[1] 0.498
```

Variance de X1 :

```
round(var(X1), 3)
```

```
[1] 1
```

logique ce soit = 1 car X1 est la concaténation de deux variables indépendantes de variance 1 (car générées par `rnorm` donc suivent une loi normale standard).

Pour calculer la variance partagée entre X1 et X2, on utilise la formule de la variance de la somme de deux variables aléatoires indépendantes :

$$Var(X1 + X2) = Var(X1) + Var(X2) + 2Cov(X1, X2)$$

Sur R :

```
# corrélation de Pearson mise au carré donne la part de variance partagée
rho2 <- (cor(X1, X2)^2)
rho2
```

```
[1] 0.248
```

Une autre manière d'obtenir ça :

1. Construire un modèle linéaire de Y en fonction de X
2. Extraire la part de variance résiduelle (non expliquée par X) du modèle
3. Variance expliquée par $X = 1 - \text{variance résiduelle}$

(mais `summary(lm())` donne directement le R^2 dans la partie "Multiple R-squared")

```
res <- lm(X1 ~ X2)
# summary donne l'info dans "Multiple R-squared"
summary(res)
```

Call:

```
lm(formula = X1 ~ X2)
```

Residuals:

Min	1Q	Median	3Q	Max
-5.079	-0.487	-0.001	0.486	4.867

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.00361	0.00194	1.86	0.063 .
X2	0.49770	0.00194	256.59	<2e-16 ***

```
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.868 on 199998 degrees of freedom  
Multiple R-squared: 0.248, Adjusted R-squared: 0.248  
F-statistic: 6.58e+04 on 1 and 2e+05 DF, p-value: <2e-16
```

```
# variance résiduelle  
round(var(residuals(res)),3)
```

```
[1] 0.754
```

```
# variance expliquée par X2  
1 - round(var(residuals(res)),3)
```

```
[1] 0.246
```

! Important

NB : la variance partagée n'est pas la même chose que la covariance !!

Covariance = mesure comment deux variables varient ensemble, positive si les deux variables augmentent ensemble, négative si l'une augmente quand l'autre diminue.

Paramètre	Symbole	Interprétation	Formule / calcul R
Coefficient de corrélation	r ou ρ	Direction et force de la relation linéaire entre deux variables quantitatives	<code>cor(X, Y)</code>
Variance partagée	r^2	Proportion de la variance de Y expliquée par X (force de la relation linéaire)	<code>rho2 <- cor(X, Y)^2</code>
Covariance	$\text{Cov}(X, Y)$	Mesure comment deux variables varient ensemble (positive : ensemble, négative : sens inverse)	<code>cov(X, Y)</code>

1.B.1.3 Matrice de corrélation Pour calculer la matrice de corrélation entre plusieurs variables quantitatives, on peut utiliser la fonction `cor()` en lui passant un data frame ou une matrice.

```
quanti <- c("age","nb.enfants","depression","schizophrenie",  
         "gravite","recherche.nouv","evit.danger","dep.recompense")  
round(cor(smp.d[,quanti],use="pairwise.complete.obs"),digits=3)
```

	age	nb.enfants	depression	schizophrenie	gravite
age	1.000	0.498	-0.093	-0.021	-0.127
nb.enfants	0.498	1.000	0.003	-0.003	-0.057

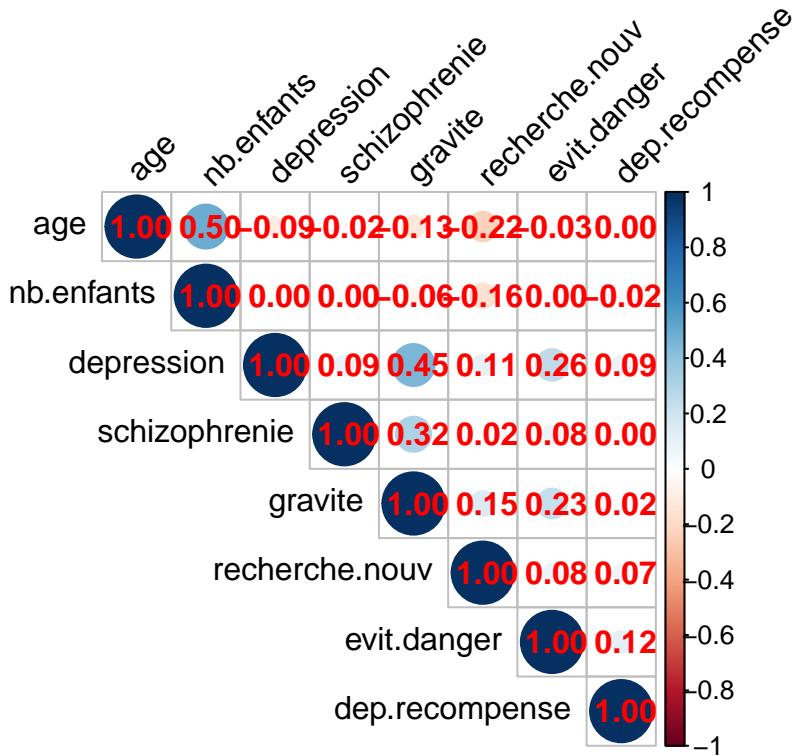
depression	-0.093	0.003	1.000	0.094	0.454
schizophrenie	-0.021	-0.003	0.094	1.000	0.318
gravite	-0.127	-0.057	0.454	0.318	1.000
recherche.nouv	-0.223	-0.159	0.109	0.022	0.154
evit.danger	-0.027	0.004	0.256	0.081	0.230
dep.recompense	-0.001	-0.023	0.089	-0.004	0.019
		recherche.nouv	evit.danger	dep.recompense	
age		-0.223	-0.027		-0.001
nb.enfants		-0.159	0.004		-0.023
depression		0.109	0.256		0.089
schizophrenie		0.022	0.081		-0.004
gravite		0.154	0.230		0.019
recherche.nouv		1.000	0.081		0.071
evit.danger		0.081	1.000		0.119
dep.recompense		0.071	0.119		1.000

On peut représenter ça graphiquement avec la fonction `corrplot()` du package `corrplot`.

```
library(corrplot)
```

```
corrplot 0.95 loaded
```

```
corrplot(
  round(cor(smp.d[,quanti],use="pairwise.complete.obs"),digits=3),
  method="circle",
  addCoef.col = "red",
  type="upper",
  tl.col="black",
  tl.srt=45)
```



Une matrice de corrélation est symétrique (mêmes valeurs de part et d'autre de la diagonale), car la corrélation entre X et Y est la même que celle entre Y et X .

i Note

A quoi ça peut servir dans une étude rétrospective ?

Dans une étude qui compare 2 techniques chirurgicales (A vs B), la matrice de corrélation sert surtout à explorer et comprendre les relations entre les nombreuses variables mesurées autour de l'intervention.

- Explorer les facteurs pré-opératoires entre eux (âge corrélé au score ASA, etc.)
- Repérer la colinéarité entre co-variables avant une régression ajustée
- Relier facteurs pré-op et outcomes post-op (âge, IMC, ASA vs durée d'hospitalisation, perte sanguine, etc.)
 - Notamment selon le type de chirurgie (groupe A vs groupe B) : est ce que les plus vieux ont plus de perte sanguine avec la technique A que B ?

1.C Concordance

1.C.1 Principle

Par exemple, “concordance” entre deux échographies identiques fait par deux médecins différents.

Variables quantitatives : concordance mesurée par le coefficient de corrélation intra-classe (ICC).

- ICC varie entre 0 (pas de concordance) et 1 (concordance parfaite).
- Dans l'exemple du score échographique, le coefficient de corrélation intraclasse =

$$\frac{\text{variance inter-patients}}{\text{variance inter-patients} + \text{variance inter-radiologues} + \text{variance résiduelle}}$$

Donc en gros : $\text{ICC} = \frac{\text{vrai signal}}{\text{vrai signal} + \text{bruit}}$

Vaut 1 quand il n'y a aucun bruit (variance inter-radiologues et résiduelle = 0).

Vaut 0 quand il n'y a que de bruit, donc les mesures sont totalement indépendantes entre elles.

1.C.2 Exemple sur R

Dans l'étude smp, 2 cliniciens sont présentés lors des entretiens : un junior et un senior.

A la fin de chaque entretien, ils remplissaient plusieurs questionnaires, dont un comportait l'échelle *CGI = Clinical Global Impression* (échelle de 1 à 7, 1 = pas malade, 7 = très malade).

Il est possible de quantifier le niveau de concordance entre les notes CGI données par le junior et le senior à l'aide du coefficient de corrélation intraclasse (ICC).

Dans ce cas : il s'agit d'un ICC de type "2-way random effects, absolute agreement, single rater/measurement" (notation ICC(2,1) de Shrout & Fleiss, 1979).

- 2-way random effects : les deux évaluateurs (junior et senior) sont considérés comme des échantillons aléatoires d'une population plus large d'évaluateurs possibles.
- Absolute agreement : on s'intéresse à l'accord absolu entre les évaluateurs, pas seulement à la corrélation.
- Single rater/measurement : on considère les notes individuelles de chaque évaluateur, pas une moyenne.

C'est le type d'ICC le plus couramment utilisé en pratique clinique.

```
psy::icc(
  smp.aij[,c("gravite.jun", "gravite.sen")]
)
```

```
$nb.subjects
[1] 796
```

```
$nb.raters
[1] 2
```

```
$subject.variance
[1] 2.4
```

```
$rater.variance
[1] -0.000228
```

```
$residual
```

```
[1] 0.257

$icc.consistency
[1] 0.903

$icc.agreement
[1] 0.903

• $subject.variance : variance sujets = variance signal
• $rater.variance : variance évaluateurs = variance bruit due aux différences entre évaluateurs
• $residual : variance résiduelle = variance bruit due aux autres sources d'erreur
• $icc.agreement : coefficient de corrélation intraclasse ICC vaut 0,9
```

Mais la librairie `irr` propose aussi une fonction `icc()` pour calculer le coefficient de corrélation intraclasse, il faut juste la paramétriser un peu plus mais l'output est plus clair.

```
library(irr)
```

```
Loading required package: lpSolve
```

```
Attaching package: 'irr'
```

```
The following object is masked from 'package:psy':
```

```
icc

irr::icc(
  smp.aij[,c("gravite.jun","gravite.sen")],
  model="twoway", # sinon oneway si un seul évaluateur par sujet
  type="agreement", # sinon consistency = uniformité des réponses
  unit="single" # single ou average
)
```

```
Single Score Intraclass Correlation
```

```
Model: twoway
Type : agreement
```

```
Subjects = 796
Raters = 2
ICC(A,1) = 0.903
```

```
F-Test, H0: r0 = 0 ; H1: r0 > 0
F(795,795) = 19.7 , p = 5.62e-295
```

```
95%-Confidence Interval for ICC Population Values:
0.89 < ICC < 0.915
```

1.D Résumé des paramètres de dépendance entre deux variables quantitatives

Résumé des principaux paramètres de dépendance entre deux variables quantitatives :

- Dépendance
 - Dépendance monotone ou linéaire
 - Coefficient de corrélation de Pearson
 - Coefficient de détermination = variance partagée (covariance)
 - Concordance = coefficient de corrélation intraclassé (ICC)
-

Paramètre	Symbole	Interprétation	Formule / calcul R
Coefficient de corrélation	r ou ρ	Direction et force de la relation linéaire entre deux variables quantitatives	<code>cor(X, Y)</code>
Variance partagée	r^2	Proportion de la variance de Y expliquée par X (force de la relation linéaire)	<code>rho2 <- cor(X, Y)^2</code>
Covariance	$\text{Cov}(X, Y)$	Mesure comment deux variables varient ensemble (positive : ensemble, négative : sens inverse)	<code>cov(X, Y)</code>
Coefficient de corrélation intraclassé	ICC	Mesure la concordance entre plusieurs mesures quantitatives	<code>psy::icc(dataframe)</code> ou <code>irr::icc(dataframe,</code> <code>model=...,</code> <code>type=...,</code> <code>unit=...)</code>

2 Variables catégorielles

2.A Dépendance

2.A.1 Chi2 et associés

- Existe-t-il une relation entre les deux variables catégorielles ?
- Si oui, quelle est la force de cette relation ?

Pour ces questions : on utilise

- **Le test du Chi2 d'indépendance** sert à évaluer l'existence d'une relation entre deux variables catégorielles.

Et des transformations normalisées du Chi2 pour évaluer la force de cette relation :

(pour normaliser le Chi2 : racine carré de $\chi^2/nombre\ d'observations$)

- **Le V de Cramer** pour la force de l'association entre deux variables catégorielles **non ordonnées** (type de chirurgie) (= on pourrait utiliser le coefficient de Pearson pour des variables binaires).
 - Varie entre 0 (pas d'association) et 1 (association parfaite)
- Le coefficient de Pearson : plutôt pour variables quantitatives, mais utilisable pour des variables binaires (0/1)
 - Varie entre -1 et +1
- **Le coefficient de Spearman** : pour variables **ordinaires** ou quantitatives non linéaires (rangées)
 - Coefficient de Spearman = corrélation de Pearson calculée sur les rangs des données
 - Varie entre -1 et +1

i Note

ANOVA ou V de Cramer pour variables catégorielles non ordonnées ?

- **ANOVA** :
 - 1 variable quantitative, 1 ou plusieurs variables catégorielles non ordonnées
 - compare les moyennes de la variable quantitative entre les différentes modalités de la variable catégorielle
 - test statistique (p-value)
 - outcome quantitatif
- **V de Cramer** :
 - Variables catégorielles non ordonnées
 - mesure la force de l'association entre les variables
 - valeur entre 0 et 1
 - outcome catégoriel

2.A.2 Odds-ratio et risque relatif

Pour des X et Y binaires (0/1) :

Exemple : association entre décès en USI et existence d'une infection ou non

	Décès (Y=1)	Pas de décès (Y=0)
Infection (X=1)	a	b
Pas d'infection(X=0)	c	d

- **Risque relatif (RR)** = risque de décès chez les patients infectés / risque de décès chez les patients non infectés = $[a/(a+b)] / [c/(c+d)]$
 - On a RR fois plus de risque de décès si on est infecté
 - $RR = \frac{\% \text{ de deces chez les infectes}}{\% \text{ de deces chez les non infectes}} = \frac{a}{a+b} / \frac{c}{c+d}$
- **Odds-ratio (OR)** = rapports des côtes = interprétation plus subtile
 - Odds de décès chez les patients infectés = a/b (côte)
 - Odds de décès chez les patients non infectés = c/d
 - $OR = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{a}{b} \times \frac{d}{c}$
 - Il y a OR fois plus de “morts par rapport aux vivants” si on est infecté que de “morts par rapport aux vivants” si on n'est pas infecté.

RR et OR sont positifs et varient de 0 à l'infini.

S'ils valent 1 : pas d'association entre X et Y, les variables sont indépendantes.

S'ils valent 0 ou sont très grands : forte association entre X et Y.

Rapport entre RR et OR :

- Si l'événement étudié est rare (<10%) : RR = OR
- Si l'événement est fréquent (>10%) : **OR surestime le RR** (OR > RR si RR > 1 ; OR < RR si RR < 1)

2.A.3 Exemple sur R

Sur fichier **smp.d** : force de l'association entre la variable “dépression” (**smp.d\$depression**) et le FDR “le prisonnier a un niveau élevé d'évitement du danger” (**smp.d\$evit.danger**).

La variable dépression est binaire (0 = non dépressif, 1 = dépressif).

La variable evit.danger n'est pas binaire, mais codée 1, 2 ou 3 pour “faible”, “moyen” ou “élevé”.

Il faut donc la recoder en binaire (0 = faible ou moyen, 1 = élevé).

```
smp.d$evit.danger.b <- smp.d$evit.danger > 2
smp.d$depression.b <- smp.d$depression == 1
tb <- table(
```

```

  smp.d$depression.b,
  smp.d$evit.danger.b,
  deparse.level=2
)
tb

```

```

          smp.d$evit.danger.b
smp.d$depression.b FALSE TRUE
    FALSE    338    96
    TRUE     131   126

```

Pour obtenir facilement le RR et l'OR, on peut utiliser la fonction `epi.2by2()` du package `epiR`.

```

epi.2by2(
  tb,
  method="cohort.count", # sinon case.control
  conf.level=0.95
)

```

	Outcome+	Outcome-	Total	Inc risk *
Exposure+	338	96	434	77.88 (73.68 to 81.70)
Exposure-	131	126	257	50.97 (44.69 to 57.24)
Total	469	222	691	67.87 (64.25 to 71.34)

Point estimates and 95% CIs:

Inc risk ratio	1.53 (1.34, 1.74)
Inc odds ratio	3.39 (2.43, 4.73)
Attrib risk in the exposed *	26.91 (19.65, 34.16)
Attrib fraction in the exposed (%)	34.55 (25.88, 42.85)
Attrib risk in the population *	16.90 (9.87, 23.93)
Attrib fraction in the population (%)	24.90 (19.77, 30.45)

Uncorrected chi2 test that OR = 1: $\text{chi2}(1) = 53.594$ $\text{Pr}>\text{chi2} = <0.001$

Fisher exact test that OR = 1: $\text{Pr}>\text{chi2} = <0.001$

Wald confidence limits

CI: confidence interval

* Outcomes per 100 population units

2.B Dépendance monotone

Une dépendance monotone ne peut s'envisager qu'entre des variables ordinaires (rangées) ou entre une variable ordinaire et une variable quantitative.

- **Coefficient de Spearman** : corrélation de Pearson calculée sur les rangs des données.
 - Varie entre -1 et +1

- Utilisable pour des variables ordinaires (rangées) ou quantitatives non linéaires

Problème : donner du sens à une corrélation basée sur des rangs !! dépend ++ du codage

2.B.1 Exemple sur R

En pratique : dans l'étude santé mentale en prison, les deux variables de tempérament : « recherche de nouveauté » et « évitemennt du danger » sont codées en 1, 2 et 3 pour, respectivement, « bas », « moyen » et « élevé ». Si l'on souhaite apprécier dans quelle mesure un niveau élevé de recherche de nouveauté est associé à un niveau bas d'évitement du danger, il est possible d'estimer un coefficient de corrélation de Spearman ou de Pearson :

Dans l'étude `smp` : variable “recherche de nouveauté” (`smp.d$recherche.nouv`) et variable “évitement du danger” (`smp.d$evit.danger`).

- les deux variables `recherche.nouv` et `evit.danger` sont codées 1, 2 ou 3 pour “faible”, “moyen” ou “élevé”.
- objectif : apprécier dans quelle mesure un niveau élevé de recherche de nouveauté est associé à un niveau bas d'évitement du danger.

```
table(smp$recherche.nouv)
```

	1	2	3
249	157	289	

```
table(smp$evit.danger)
```

	1	2	3
315	154	222	

Calcul du coefficient de corrélation de Spearman entre les deux variables.

```
cor(
  smp.d$recherche.nouv,
  smp.d$evit.danger,
  method="spearman",
  use="complete.obs") # ignore les valeurs manquantes
```

[1] 0.0785

Le coefficient de corrélation de Spearman est de 0.078, ce qui indique une très faible dépendance monotone positive entre la recherche de nouveauté et l'évitement du danger.

NB : si on avait utilisé le coefficient de Pearson :

```

cor(
  smp.d$recherche.nouv,
  smp.d$evit.danger,
  method="pearson",
  use="complete.obs") # ignore les valeurs manquantes

```

[1] 0.0807

Le coefficient de corrélation de Pearson est de 0.081 : les deux coefficients sont très proches (c'est assez fréquent quand les variables sont ordinaires avec peu de modalités).

2.C Concordance

2.C.1 Coefficient kappa de Cohen

Pour les variables catégorielles, on pourrait se dire que mesurer à quel point 2 variables s'accordent reviendrait à compter la proportion de fois où elles ont la même valeur.

Problème : cette proportion de concordance peut être due au hasard !!

On corrige ça avec le **kappa de Cohen**.

$$\text{kappa} = \frac{\text{concordance observée} - \text{concordance due au hasard}}{1 - \text{concordance due au hasard}}$$

- $\text{kappa} = 0$: concordance observée = concordance due au hasard
- $\text{kappa} = 1$: concordance parfaite

2.C.2 Sensibilité, spécificité, VPP, VPN

Le kappa de Cohen mesure une **concordance globale et symétrique** entre deux variables catégorielles.

Mais parfois, on s'intéresse à la capacité d'une variable à s'approcher d'une variable de référence = mesurer à quel point Y prédit correctement X .

Dans ce cas, il vaut mieux utiliser des paramètres asymétriques :

- Sensibilité = proportion de vrais positifs parmi les positifs réels = $P(Y = 1|X = 1)$
- Spécificité = proportion de vrais négatifs parmi les négatifs réels = $P(Y = 0|X = 0)$

Le problème avec la sensibilité et la spécificité : elles ne tiennent pas compte de la prévalence de la condition réelle X (c'est à dire la proportion de $X = 1$ dans la population).

- Valeur prédictive positive (VPP) = proportion de vrais positifs parmi les positifs prédits = $P(X = 1|Y = 1)$
- Valeur prédictive négative (VPN) = proportion de vrais négatifs parmi les négatifs prédits = $P(X = 0|Y = 0)$

Dans un tableau de contingence :

	Y=1 (test positif)	Y=0 (test négatif)
X=1 (condition réelle présente)	a (vrais positifs)	b (faux négatifs)
X=0 (condition réelle absente)	c (faux positifs)	d (vrais négatifs)

- Sensibilité = $a / (a + b)$
- Spécificité = $d / (c + d)$
- Valeur prédictive positive (VPP) = $a / (a + c)$
- Valeur prédictive négative (VPN) = $d / (b + d)$

On peut ainsi représenter une courbe ROC (Receiver Operating Characteristic) qui trace la sensibilité en fonction de 1 - spécificité pour différents seuils de décision.

2.C.2.1 Exemple 1 sur R

- Les deux cliniciens (junior et senior) posent un diagnostic de schizophrénie (1 = oui, 0 = non) pour chaque patient.
- Niveau d'accord inter-juges pour une variable catégorielle : kappa de Cohen.

```
psy::ckappa(
  smp.aij[,c("scz.jun","scz.sen")]
)
```

```
$table
  0  1
0 715 11
1 30 43
```

```
$kappa
[1] 0.65
```

Autre méthode avec le package **irr** :

```
irr::kappa2(
  smp.aij[,c("scz.jun","scz.sen")],
  weight="unweighted" # ou "equal" ou "squared" pour kappa pondéré
)
```

Cohen's Kappa for 2 Raters (Weights: unweighted)

```
Subjects = 799
Raters = 2
Kappa = 0.65
```

```
z = 18.6
p-value = 0
```

Le clinicien junior a posé le diagnostic de schizophrénie chez $30 + 43 = 73$ détenus alors que le clinicien senior l'a fait pour seulement $11 + 43 = 54$.

Au total, le coefficient kappa vaut 0,65.

2.C.2.2 Exemple 2 sur R

- Étude sur 244 patients déprimés hospitalisés : tâche de lecture de texte puis comptage de 1 à 10, enregistrement voix.
- Extraction de la fréquence fondamentale (hauteur de voix), connue pour être $\sim 75\text{--}140$ Hz chez les hommes et $\sim 170\text{--}250$ Hz chez les femmes.
- Objectif : voir dans quelle mesure la hauteur de voix prédit le sexe en testant un seuil de 155 Hz.
- Méthode : **calcul sensibilité et spécificité du seuil 155 Hz** pour discriminer hommes et femmes.

```
sexef <- vox$sex == 2
voix.aigue <- vox$moyf0>155
# moyenne des femmes avec voix aiguë = sensibilité
# = proportion de femmes avec test positif
mean(voix.aigue[sexef], na.rm=TRUE)
```

[1] 0.917

```
# moyenne des hommes sans voix aiguë = spécificité
# = proportion d'hommes avec test négatif
mean(!voix.aigue[!sexef], na.rm=TRUE)
```

[1] 0.933

Vérifier le seuil de 155 Hz avec une courbe ROC :

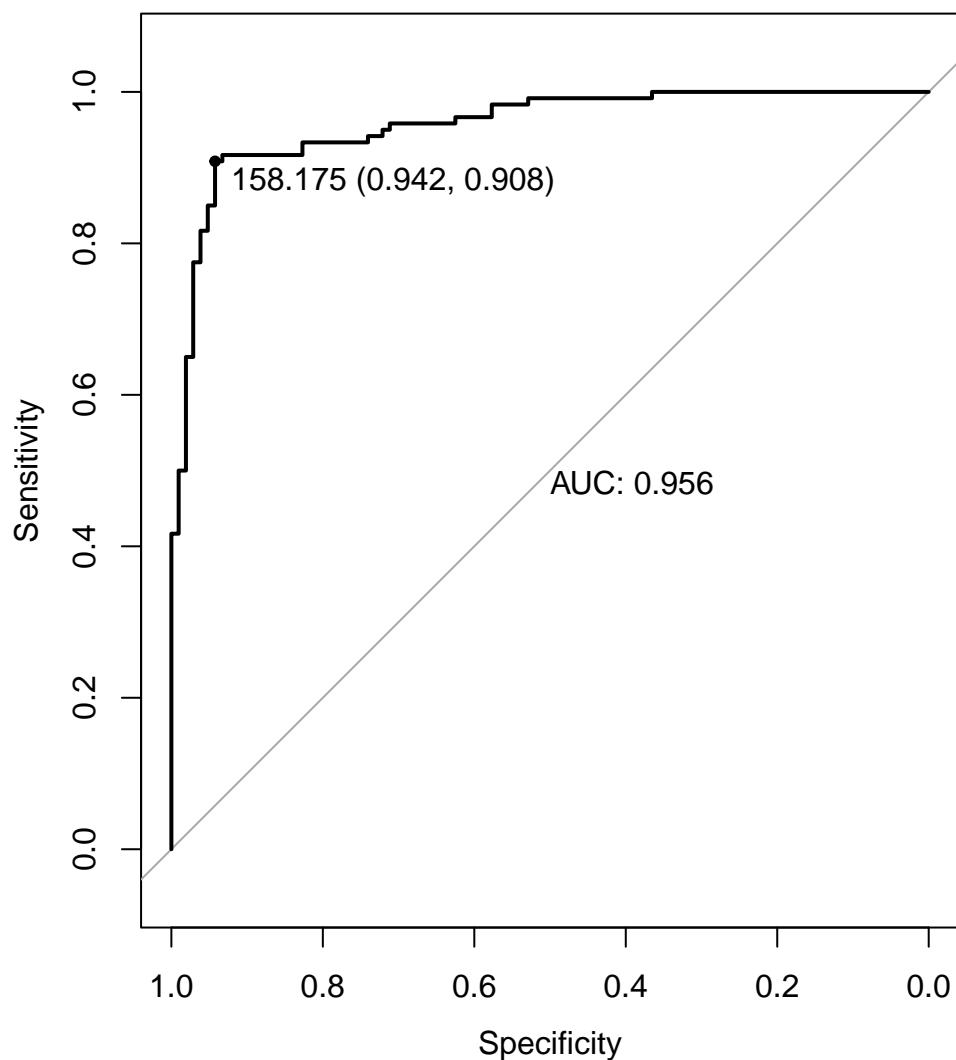
```
rocf0 <- roc(sexf~vox$moyf0)
```

Setting levels: control = FALSE, case = TRUE

Setting direction: controls < cases

```
plot(rocf0,
      main="Courbe ROC pour la hauteur de voix",
      print.thres="best",
      print.thres.best.method="youden",
      print.auc=TRUE)
```

Courbe ROC pour la hauteur de voix



Seuil optimal calculé par indice de Youden (maximise la somme de la sensibilité et de la spécificité)
= 158,17 Hz

AUC : calcule la qualité globale du test

- Aire comprise entre 0 et 1
- = probabilité que la hauteur de voix d'une femme soit plus élevée que celle d'un homme pris au hasard

76 Effet centre

8Introduction

Les mesures répétées désignent des situations où plusieurs observations sont recueillies sur une même unité statistique (souvent un patient) au cours du temps, ou lorsque les observations sont regroupées au sein de clusters (par exemple, des patients au sein d'un même hôpital ou des membres d'une même famille).

Le défi majeur posé par ces données réside dans la non-indépendance des observations.

En effet, les mesures effectuées sur un même patient à différents moments sont généralement corrélées entre elles.

Ignorer cette corrélation en utilisant des méthodes statistiques classiques (comme la régression linéaire standard ou l'ANOVA classique) viole l'hypothèse d'indépendance des résidus, conduisant à des estimations biaisées de la précision (intervalles de confiance trop étroits) et à une augmentation du risque d'erreur de type I.

Si l'analyse est relativement simple avec seulement deux points de mesure (analyse de l'évolution $Y_{aprs} - Y_{avant}$ ou ajustement sur la valeur basale), elle devient plus complexe avec trois mesures ou plus.

9Modèle linéaire

Imaginons qu'on corrèle la durée d'entretien (en minutes) Y avec le niveau de dépression X .

Y est supposé normalement distribué.

X est supposé quantitatif.

Corrélation : peut être intra- ou inter-centre

- Intra-centre : rechercher si un sujet avec un niveau de dépression X élevé a une durée d'entretien Y supérieure à celle d'un autre sujet du même centre présentant un niveau de dépression plus bas.
- Inter-centre : la durée moyenne des entretiens \hat{Y} réalisés dans les prisons avec un haut niveau de dépression est-elle supérieure ou inférieure à la durée moyenne des entretiens réalisés dans les autres prisons ?

1 Analyse naïve sans prise en compte du centre

Imaginons qu'on néglige l'effet centre et qu'on réalise une régression linéaire simple de Y en fonction de X .

On prend 3 centres comprenant chacun 9 détenus.

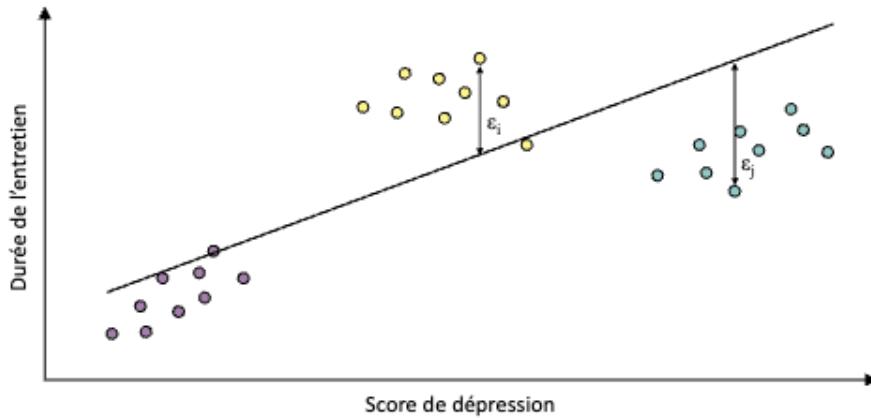


Figure 9.1: Estimation d'une corrélation entre « durée de l'entretien » et « score de dépression » sans tenir compte de l'effet centre (ici un centre par couleur). La droite de régression minimise la somme des ε_i^2 et néglige l'effet centre.

Ici :

- Effet centre particulièrement important
- Chaque centre semble associé à un niveau spécifique de dépression \hat{X}
- Et à un niveau moyen spécifique de durée d'entretien \hat{Y}

1.A Problèmes posés par cette analyse naïve

= biais de confusion par effet de groupe + non-indépendance des résidus

Problème 1 : Biais de confusion par effet de groupe.

- Si les centres pour lesquels le niveau de dépression est le plus élevé sont aussi ceux où la durée moyenne des entretiens est la plus longue, la pente de régression est **biaisée**.
- Elle mélange l'effet individuel (la relation durée-dépression pour un détenu) et l'effet du groupe. Le modèle ne peut distinguer si la dépression est liée à la durée de l'entretien ou à d'autres facteurs propres au centre (personnel, type de population, etc.).
- *Conséquence* : Le coefficient de régression estimé est faussé et ne représente pas la véritable association au niveau individuel.

Problème 2 : Non-indépendance des résidus.

- Les résidus ε_i ne sont pas indépendants. Les observations au sein d'un même groupe (centre) se ressemblent plus qu'avec celles d'autres groupes.
- C'est la **corrélation intra-classe**.
- *Conséquence* : Ce problème ne biaise pas l'estimation de la pente, mais il conduit à une **sous-estimation de son erreur-standard**.
- *Impact* : L'intervalle de confiance est artificiellement étroit et la p-value trop petite. Le risque est de conclure à tort à un effet significatif (Erreur de Type I).
- **Solutions** : **Bootstrap par grappe ou estimateur sandwich**

– Le Bootstrap par grappe (Cluster Bootstrap) :

* *Principe* :

- Puisque les individus d'un groupe ne sont pas indépendants, on ré-échantillonne les **groupes** (les centres) avec remise.
- Pour chaque groupe tiré, on inclut tous les individus qu'il contient.

* *Résultat* :

- On obtient une distribution de 1000 coefficients qui reflète l'incertitude liée à la variabilité *entre les groupes*. L'IC à 95% est construit à partir de cette distribution (ex: via les percentiles).
- **Cet intervalle sera presque toujours plus large** que l'IC naïf, reflétant une estimation plus honnête de l'incertitude.

– L'estimateur “Sandwich” (Estimateur Robuste de la Variance) :

* **Le problème** :

- Le modèle classique est “trop confiant”. Il pense que chaque ligne de données apporte une information unique.
- Or, si les patients d'un même centre se ressemblent, on a moins d'information réelle qu'on ne le croit.
- L'erreur-standard calculée classiquement est donc trop petite.

* **Solution** :

- Au lieu de se fier uniquement à la théorie (qui suppose l'indépendance), l'estimateur Sandwich regarde les résidus réels (les erreurs du modèle).
 - Si, dans un centre, tous les résidus vont dans le même sens (ex: le modèle se trompe toujours par excès pour ce centre), l'estimateur détecte cette corrélation. Il utilise cette “réalité du terrain” pour corriger mathématiquement la variance à la hausse.
- * **Principe :** La formule mathématique de la variance robuste se compose de trois blocs multipliés entre eux : $A \times B \times A$.
- * On l'appelle “Sandwich” uniquement parce que la correction (B) est coincée entre deux blocs identiques (A), comme une tranche de jambon entre deux tranches de pain.
1. **Le bloc A (La Théorie)** : C'est la variance calculée par le modèle classique. Elle suppose que tout est parfait (indépendance).
 2. **Le bloc B (La Réalité)** : C'est une correction calculée directement à partir des **données brutes** (les résidus). Si les erreurs sont corrélées dans les groupes, ce bloc B va “gonfler” la valeur.
 3. **Le calcul** : On multiplie Théorie \times Correction \times Théorie.
- * **Résultat** : Les estimations des coefficients (la pente) ne changent pas, mais les intervalles de confiance s'élargissent et les p-values augmentent, reflétant une incertitude plus honnête.

1.B Exemple R avec jeu de données fictif

1. Génération des données

```
set.seed(1)
# génération des Xi et Yi pour 5 temps de mesure dans 200 centres
x1 <- runif(200)*0.3 # Génération des xi pour le centre j
x2 <- runif(200)*0.3
x3 <- runif(200)*0.3
x4 <- runif(200)*0.3
x5 <- runif(200)*0.3
a <- rnorm(200)*100+300 # Les coefficients de la régression
b <- rnorm(200)*100+150 # sont fonction du centre
y1 <- a+b*x1+rnorm(200)*72.5
y2 <- a+b*x2+rnorm(200)*72.5
y3 <- a+b*x3+rnorm(200)*72.5
y4 <- a+b*x4+rnorm(200)*72.5
y5 <- a+b*x5+rnorm(200)*72.5

# dt est le fichier "large" (une ligne par centre)
dt <- data.frame(1:200,x1,y1,x2,y2,x3,y3,x4,y4,x5,y5)
names(dt)[1] <- "centre"
```

```
# conversion du fichier "large" en fichier "long"
dtl <- reshape(dt,idvar="centre",varying=2:11,v.names=c("x","y"),timevar =
  ~ "temps",direction="long")
dtl$centre <- as.factor(dtl$centre)
```

2. Régression linéaire simple sans prise en compte de l'effet centre

```
mod.lm <- lm(y ~ x, data=dtl)
summary(mod.lm)$coefficients
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	286.6649	8.232679	34.820363	1.591602e-174
x	153.6712	47.571204	3.230341	1.276828e-03

Ici, le calcul de l'erreur-standard (l'écart-type) et de la p-value repose sur l'hypothèse que les résidus sont indépendants, ce qui n'est pas le cas.

Impossible de le savoir si on n'a pas d'information sur les données !!

Attention parce que les résidus semblent "normaux" (cf. graphique ci-dessous), cela ne garantit pas leur indépendance.

```
par(mfrow=c(1,2))
hist(residuals(mod.lm),main="Histogramme des résidus",xlab="Résidus")
qqnorm(residuals(mod.lm),main="Q-Q plot des résidus")
qqline(residuals(mod.lm))
```

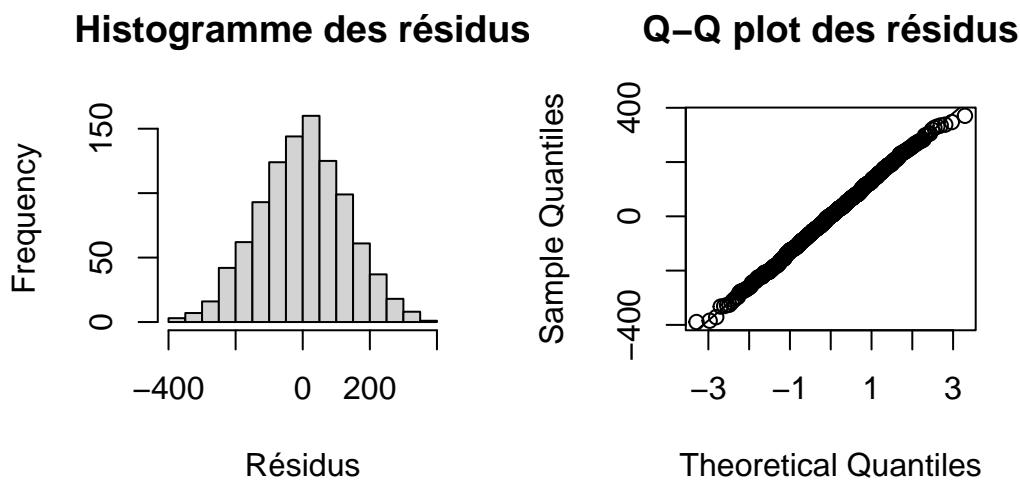


Figure 9.2: Graphique des résidus du modèle linéaire simple sans prise en compte de l'effet centre

Bootstrap

Il est possible de recourir à un bootstrap pour contourner cette limite, en utilisant la fonction `clusbootglm()` de la bibliothèque `ClusterBootstrap`.

```

set.seed(1)
library("ClusterBootstrap")
mod.clusboot <- clusbootglm(y~x,data=dtl,clusterid=centre)

```

```

$boot.coefs :
(Intercept)      x
286.6787       153.6561
$boot.sds
(Intercept)      x
11.03841        50.14656

```

L'écart type de $b = 50,14$ (par rapport à 47,57 naïf). Les coefficients ne changent pas.

Estimateur sandwich

```

library("sandwich")
sqrt(diag(vcovCL(lm(y~x,data=dtl),cluster=~centre))["x"])

```

```

x
50.73999

```

L'estimation de l'erreur type (50.74) est très proche de celle du bootstrap, prenant toujours en compte la dépendance des observations d'un même centre.

2 Analyse avec prise en compte de l'effet centre

il s'agit ici de calculer les moyennes de X et Y pour chaque centre puis d'en faire une régression linéaire

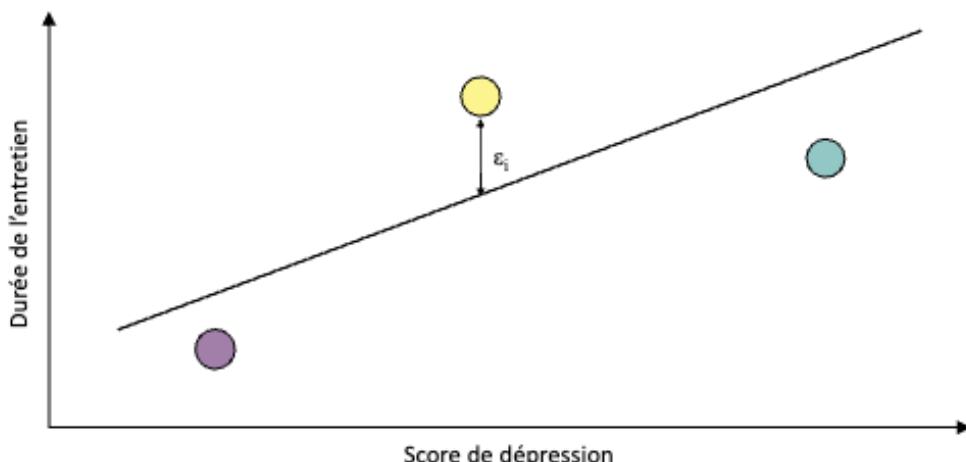


Figure 9.3: Estimation d'une pente inter-centres à partir des moyennes de X et Y pour chaque centre.

2.A Exemple R

```
#moyenne de Y et X par centre  
ymeans = tapply(dtl$y, dtl$centre, mean)  
xmeans = tapply(dtl$x, dtl$centre, mean)  
  
#régression linéaire des moyennes  
summary(lm(ymean ~ xmeans))
```

Call:

```
lm(formula = ymeans ~ xmeans)
```

Residuals:

Min	1Q	Median	3Q	Max
-309.22	-75.94	-4.72	73.62	308.89

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)							
(Intercept)	268.8	29.6	9.080	<2e-16 ***							
xmeans	272.9	190.4	1.434	0.153							

Signif. codes:	0	'***'	0.001	'**'	0.01	'*'	0.05	'..'	0.1	' '	1

Residual standard error: 111.4 on 198 degrees of freedom

Multiple R-squared: 0.01027, Adjusted R-squared: 0.005276

F-statistic: 2.055 on 1 and 198 DF, p-value: 0.1532

xmeans : fournit une estimation de la pente inter-centres.

Fournit une estimation de la pente inter-centres, avec une erreur type estimée en prenant en compte l'indépendance des centres.

Cela indique comment la moyenne de Y varie d'un centre à l'autre en fonction de la moyenne de X du centre. L'erreur type est ici estimée correctement car on travaille sur les centres (qui sont indépendants entre eux), et non sur les observations répétées.

3 Analyse intra-centres (modèles conditionnels, mixtes ou non)

3 méthodes :

1. Modèle linéaire avec effet fixe par centre (chaque centre)
2. Modèle linéaire avec pente commune (intercept variable par centre)
3. Modèle mixte avec effet aléatoire de centre (c'est à dire que les centres sont vus comme un échantillon aléatoire d'une population plus large de centres possibles)

3.A Modèle linéaire avec effet fixe par centre

Pour évaluer la corrélation intra-centre entre X et Y , on peut faire une régression linéaire dans chaque centre puis calculer la moyenne des pentes obtenues.

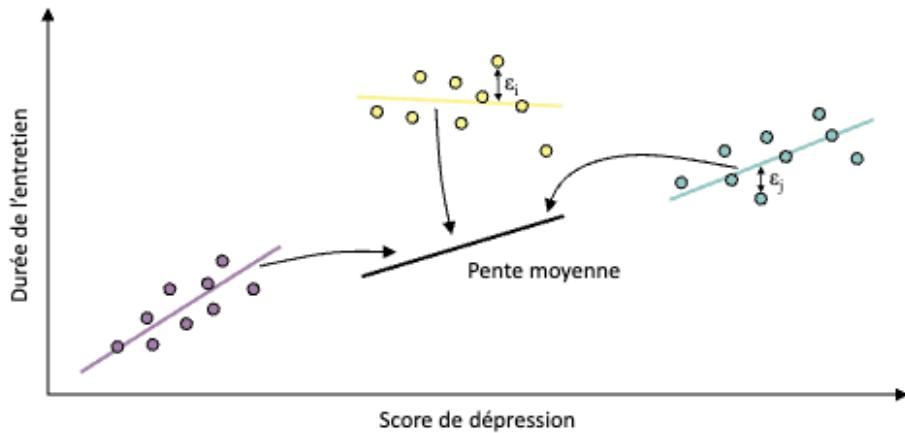


Figure 9.4: Approche conditionnelle artisanale pour laquelle une régression est réalisée dans chaque centre, une pente moyenne étant calculée dans un second temps.

C'est une approche “conditionnelle” car chaque pente est calculée “conditionnellement” à un centre donné.

L'équation serait :

$$Y_{ij} = a + [a.\text{centre}_j] + bx_{ij} + [b.\text{centre}_j]x_{ij} + \varepsilon_{ij}$$

- i : individu
- j : centre
- a : intercept global
- b : pente globale
- $[a.\text{centre}_j]$: déviation de l'intercept pour le centre j
- $[b.\text{centre}_j]$: déviation de la pente pour le centre j

Globalement : ça revient à faire une régression linéaire avec des interactions entre X et le centre.

Donc chaque centre a sa propre droite de régression (intercept et pente différents).

En R simplement : `lm(y ~ x * centre)`.

3.B Modèle linéaire avec pente commune (intercept variable par centre)

Variante plus simple : pente commune

Modèle proposé :

$$y_{ij} = a + [a.\text{centre}_j] + b x_{ij} + \varepsilon_{ij}$$

Ici :

- chaque centre j a son intercept propre : $a + [a.\text{centre}_j]$
- mais la pente b est identique dans tous les centres

Graphiquement : toutes les droites sont parallèles (même pente) mais décalées verticalement (intercepts différents).

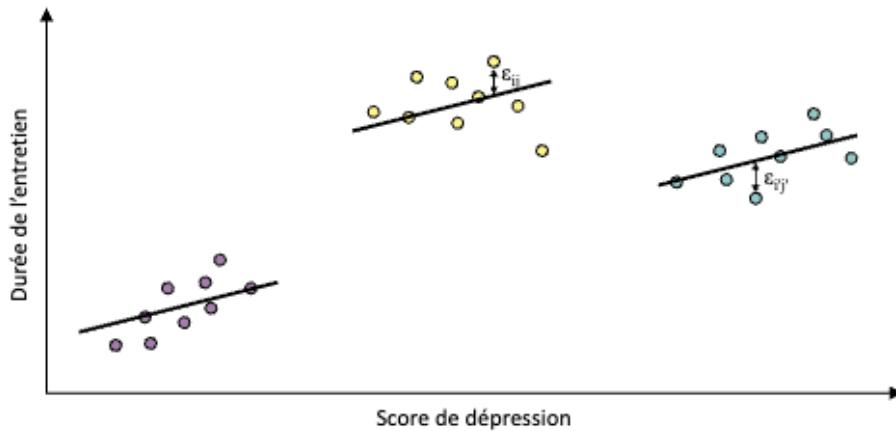


Figure 9.5: Approche conditionnelle avec pente commune, elle correspond au modèle : $y_{ij} = a + [a.\text{centre}_j] + b x_{ij} + \varepsilon_{ij}$

En R : `lm(y ~ x + centre)`

Avec 2 trucs importants à savoir sur l'interprétation de b :

- b : sorte de moyenne des pentes centre par centre, pondérée par la variance de x dans chaque centre.
 - Donc lié aux pentes que l'on obtiendrait si on faisait une régression séparée dans chaque centre, puis qu'on faisait une moyenne pondérée.
- erreur type de b : correcte uniquement si vraiment toutes les pentes sont égales entre centres.
 - Si en réalité les pentes diffèrent un peu entre centres, mais que le modèle force une pente commune, alors :
 - * b reste un estimateur moyen
 - * mais l'incertitude autour de b est mal évaluée → d'où la suggestion d'utiliser un estimateur sandwich (variance robuste) ou le bootstrap.

Résumé :

- Modèle plus simple, mais repose sur l'hypothèse forte : “même pente partout”
- Si cette hypothèse est fausse, le b affiché reste “une pente moyenne”, mais son écart-type est trop optimiste.

3.C Modèle mixte avec effet aléatoire de centre

Utile surtout si beaucoup de centres.

Passage aux modèles “mixtes” : centre comme effet aléatoire

Idée : plutôt que mettre une variable catégorielle [centre] avec 10 000 modalités, on introduit une variable aléatoire (centre).

Donc la variable catégorielle [centre] devient une variable aléatoire gaussienne (centre).

Nouveau modèle :

$$y_{ij} = a + (a_{\text{centre}} * j) + b, x * ij + \varepsilon_{ij}$$

Ou :

- $(a_{\text{centre}} * j)$: effet aléatoire pour le centre j, c'est à dire une variable aléatoire qui suit une distribution normale et qui modélise la variabilité des intercepts entre centres.
- b : pente commune à tous les centres.
- ε_{ij} : erreur résiduelle pour l'individu i dans le centre j.

Les effets sont “mixtes” car il y a à la fois des effets fixes (a, b) et des effets aléatoires ($a_{\{\text{centre}\}*j}$).

Ca aide car :

- Pas besoin de 9999 variables binaire (dummy) pour les centres.
- Processus aléatoire pour les centres.

Il faut le faire surtout s'il y a plus de 5 centres.

3.D Exemple R

Calcul des pentes de régressions de Y en fonction de X dans chaque centre

```
# fonction calculant la pente pour un centre donné
pente_intra=function(centre) {coef(lm(data=dtl[dtl$centre==centre,], y~x))["x"]}
# application de la fonction à tous les centres
pentes = sapply(levels(dtl$centre), pente_intra)

# affichage de la pente moyenne et de son erreur standard
cat("La pente moyenne est de", round(mean(pentes), 2),
    "avec une erreur standard de", round(sd(pentes) / sqrt(length(pentes)), 2),
    "\n")
```

La pente moyenne est de 131.66 avec une erreur standard de 41.6

- Modèle avec effet fixe par centre $Y_{ij} = a + [a.\text{centre}_j] + bx_{ij} + [b.\text{centre}_j]x_{ij} + \varepsilon_{ij}$

C'est surtout le paramètre b qui nous intéresse.

On utilise `contr.sum` pour que le coefficient x corresponde à la pente MOYENNE de tous les centres (sinon ce serait la pente du centre de référence). Le “point zéro” n'est plus le centre 1 mais la moyenne des centres.

```
dtlbis <- dtl
contrasts(dtlbis$centre) <- contr.sum
# estimation du modèle avec effet fixe par centre
mod.lmci <- lm(y~x*centre, data=dtlbis)
# affichage dans un cat()
cat("La pente intra-centres estimée est de",
    round(summary(mod.lmci)$coefficients["x", "Estimate"], 2),
    "avec une erreur standard de", round(summary(mod.lmci)$coefficients["x", "Std.
    Error"], 2), "\n")
```

La pente intra-centres estimée est de 131.66 avec une erreur standard de 39.47

- Modèle avec pente commune $Y_{ij} = a + [a.\text{centre}_j] + b x_{ij} + \varepsilon_{ij}$

```
mod.lmc <- lm(y~x+centre, data=dtl)
cat("La pente intra-centres estimée est de",
    round(summary(mod.lmc)$coefficients["x", "Estimate"], 2),
    "avec une erreur standard de", round(summary(mod.lmc)$coefficients["x", "Std.
    Error"], 2), "\n")
```

La pente intra-centres estimée est de 118.26 avec une erreur standard de 31.51

- Modèle mixte avec effet aléatoire de centre $Y_{ij} = a + (a_{\text{centre}*j}) + b, x * ij + \varepsilon_{ij}$

Il faut utiliser la fonction `lmer()` de la bibliothèque `lme4` ou `nlme`.

Syntaxe : `(1|centre)` signifie qu'on modélise un intercept aléatoire par centre.

```
library(nlme)
mod.lmer1 <- lmer(y~x+(1|centre), data=dtl)
summary(mod.lmer1)
```

Linear mixed model fit by REML. t-tests use Satterthwaite's method [
lmerModLmerTest]

Formula: $y \sim x + (1 | \text{centre})$

Data: dtl

REML criterion at convergence: 11949.7

Scaled residuals:

Min	1Q	Median	3Q	Max
-3.1811	-0.6001	0.0126	0.6147	2.7906

Random effects:

Groups	Name	Variance	Std.Dev.
centre	(Intercept)	11242	106.03
Residual		5723	75.65

Number of obs: 1000, groups: centre, 200

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	291.354	9.146	346.352	31.857	< 2e-16 ***
x	122.391	31.081	840.929	3.938	8.91e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:

(Intr)
x -0.509

La méthode ici utilisée est la **vraisemblance restreinte (REML)**, dont le principe est d'estimer les paramètres de variance en maximisant la vraisemblance des résidus (c'est à dire leur cohérence avec les données observées), ce qui est préférable pour estimer les paramètres de variance des effets aléatoires.

L'effet centre et le résidus sont donc caractérisés par des variances estimées (du fait qu'on les considère comme des variables aléatoires).

La variance de l'effet centre est donc estimée sensiblement plus importante que le bruit présent dans le modèle (résidus).

L'estimation de la pente vaut 122,39, légèrement différente des modèles précédents.

Pour interpréter la pente d'un modèle à effets mixtes :

- Si la variance intra-centre de X est très faible, la pente fixe du modèle mixte s'interprète comme une pente inter-centre.
 - Si la variance inter-centre de X est faible (donc faible effet centre), la pente fixe s'interprète comme une pente intra-centre.
4. Modèle mixte avec pente aléatoire par centre $Y_{ij} = a + (a_{\text{centre}*j}) + (b_{\text{centre}*j})x_{ij} + \varepsilon_{ij}$

3.E Conditions de validités des modèles mixtes

- La pente inter-centres est égale à la moyenne des pentes intra-centres (c'est à dire que les pentes des centres ne diffèrent pas trop entre elles) ;
- L'indépendance des résidus, ce qui peut poser des problèmes lorsque la structure de corrélation intra-centre est complexe, notamment lorsqu'elle est susceptible d'être négative, car les modèles mixtes classiques ne permettent pas de modéliser des corrélations négatives entre les observations d'un même centre ;
- Contribution des plus gros centres supérieure à celle des petits centres. Ca pose problème uniquement si la taille du centre est corrélée à la variable Y .
- Indépendance, normalité, homoscédasticité et nullité moyenne des effets aléatoires ; qu'il s'agisse des ordonnées à l'origine spécifiques à chacun des centres ou de leurs pentes ;

4 Modèle marginal = GEE (Generalized Estimating Equations)

Y : score de dépression, X durée de l'entretien, patients regroupés par centre et il existe une corrélation intra-centre.

Approche conditionnelle (modèles mixtes, etc.) : "Quel est l'effet de X dans un centre donné ?"

Approche marginale : "Quel est l'effet de X en moyenne dans la population, tous centres confondus ?"

→ on ne modélise pas les centres un par un ; on s'intéresse à la moyenne globale et on corrige seulement la corrélation.

4.A Définition

Dans l'approche conditionnelle : on réalise une série de modèles au sein même de chaque centre (on regarde ce qui se passe "à l'intérieur" des centres).

Dans l'approche marginale : on écrit un modèle très simple : en gros, comme $\text{lm}(y \sim x)$, mais on remplace l'estimateur classique par un autre estimateur, qui corrige pour la corrélation intra-centre.

Il faut faire l'hypothèse d'égalité des pentes inter-centres et intra-centre : la pente de la relation $X - Y$ est la même entre les centres (si on compare des centres entre eux) et à l'intérieur des centres (si on regarde les patients d'un même centre).

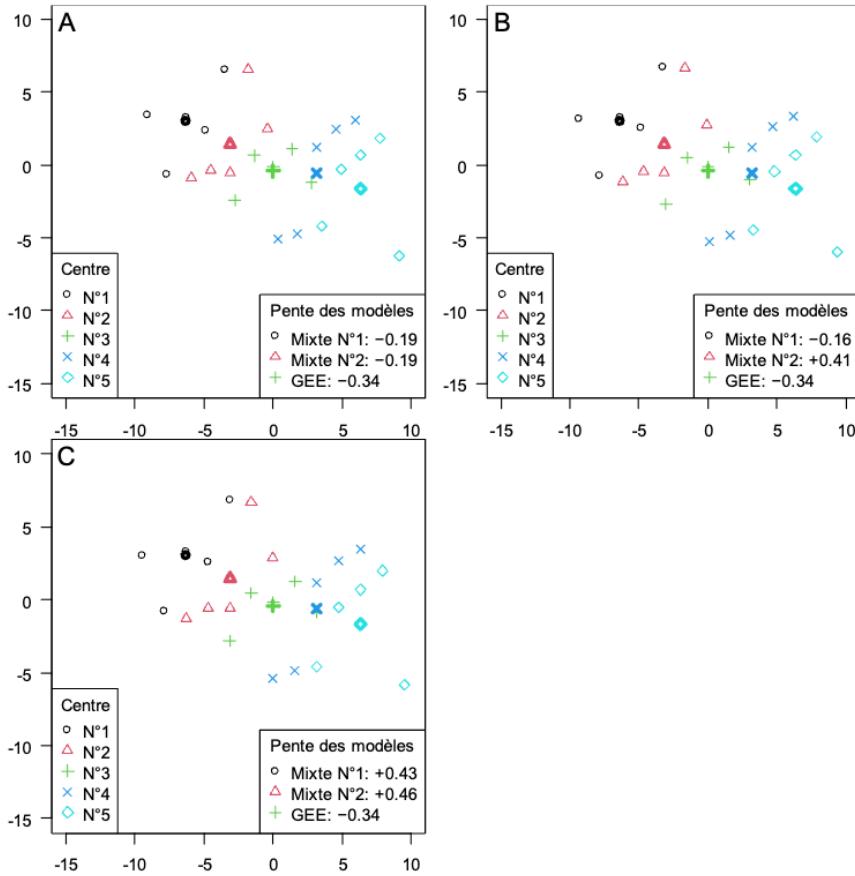
Sous cette hypothèse, l'estimation GEE de la pente est plus efficace (variance plus faible) que l'estimation naïve de $\text{lm}(y \sim x)$, c'est à dire même moyenne mais moins de variance.

4.B Principe des GEE

1. D'abord estimation d'à quel point les observations d'un même centre sont corrélées entre elles (corrélation intra-centre).
2. Ensuite, GEE utilise cette information pour ajuster la manière de calculer les coefficients du modèle linéaire

Si la corrélation intra-centre est faible, GEE donne des résultats très proches de $\text{lm}(y \sim x)$.

Si la corrélation intra-centre est forte, GEE fournit une estimation plus précise de la pente, mais l'interprétation de cette pente devient plus complexe.



Simulations de jeux de données multicentriques de taille modeste (4 observations par centre dans 5 centres).

Les panneaux A, B et C correspondent à des jeux légèrement différents.

Dans les encadrés sont données :

- N°1 : les pentes fixes d'un modèle à effets mixtes à intercept aléatoire $y_{ij} = a + [a.\text{centre}_j] + bx_{ij} + \varepsilon_{ij}$;
- N°2 : les pentes d'un modèle à effets mixtes à intercept et pente aléatoires ($y_{ij} = a + [a.\text{centre}_j] + bx_{ij} + [b.\text{centre}_j]x_{ij} + \varepsilon_{ij}$).
- Les pentes d'un modèle linéaire estimé par GEE sont également estimées.

Le symbole en gras représente la moyenne de chaque centre (barycentre des points du centre).

Une modification minime des données conduit les modèles à effets mixtes à estimer tantôt la pente inter-centres (décroissante donc négative), et tantôt la pente intra-centre (positive)

```
##is# Exemple R
```

Il faut utiliser la library **gee**.

Dans la syntaxe, l'utilisation de **order** est indispensable pour que les observations soient regroupées par centre.

L'option **corstr="exchangeable"** indique que la corrélation intra-centre est supposée identique entre toutes les paires d'observations d'un même centre (structure de corrélation dite "échangeable"), c'est à dire que la corrélation entre les observations 1 et 2 d'un centre est la même que celle entre les observations 1 et 3, etc.

```
dtlgee <- dtl[order(dtl$centre,dtl$temp),]  
mod.gee <- gee(  
  y~x,  
  data=dtlgee,  
  id=centre,  
  corstr="exchangeable"  
)
```

```
Beginning Ggee S-function, @(#) geeformula.q 4.13 98/01/27
```

```
running glm to get initial regression estimate
```

```
(Intercept)          x  
 286.6649     153.6712
```

```
summary(mod.gee)
```

```
GEE: GENERALIZED LINEAR MODELS FOR DEPENDENT DATA  
gee S-function, version 4.13 modified 98/01/27 (1998)
```

Model:

```
Link:           Identity  
Variance to Mean Relation: Gaussian  
Correlation Structure: Exchangeable
```

Call:

```
gee(formula = y ~ x, id = centre, data = dtlgee, corstr = "exchangeable")
```

Summary of Residuals:

Min	1Q	Median	3Q	Max
-391.96041	-86.83861	2.71576	83.68505	368.27139

Coefficients:

	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
(Intercept)	291.3500	9.135526	31.891982	9.410923	30.958709
x	122.4175	31.121353	3.933554	34.743959	3.523419

Estimated Scale Parameter: 16929.16

Number of Iterations: 2

Working Correlation

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	1.0000000	0.6610401	0.6610401	0.6610401	0.6610401
[2,]	0.6610401	1.0000000	0.6610401	0.6610401	0.6610401
[3,]	0.6610401	0.6610401	1.0000000	0.6610401	0.6610401
[4,]	0.6610401	0.6610401	0.6610401	1.0000000	0.6610401
[5,]	0.6610401	0.6610401	0.6610401	0.6610401	1.0000000

10 Modèle linéaire généralisé

1 Principles généraux

Si Y n'est plus une variable quantitative normalement distribuée, mais une variable binaire (ex: succès/échec), on peut utiliser un modèle linéaire généralisé (GLM) avec une fonction de lien logit (régression logistique).

NB : condition de validité de l'estimateur du maximum de vraisemblance de la régression logistique = au moins 5 à 10 évènements par variable explicative incluse dans le modèle.

On peut faire une "régression logistique conditionnelle" :

- vraisemblance partielle au sein de chaque centre puis agréger ces vraisemblances en une seule valeur
 - permet d'obtenir des OR intra-centres, mais empêche l'étude de l'effet centre lui-même.
- "modèle linéaire généralisé à effets mixtes = GLMM" (Generalized Linear Mixed Model) :
 - la variable (centre) est vue comme un effet aléatoire.

On peut aussi faire une approche "marginale" :

- Ignorer l'effet centre en premier lieu (modèle logistique simple), avec si possible un bootstrap par grappe ou un estimateur sandwich pour corriger l'erreur standard pour prendre en compte la corrélation intra-centre.
- Utiliser un modèle GEE avec une fonction de lien logit pour obtenir des OR "marginales" corrigées de la corrélation intra-centre.

2 Exemple R avec données simulées

On reprend le jeu de données simulées mais transforme Y en variable binaire selon un seuil de 300.

```
dtl$y.b <- ifelse(dtl$y>300,1,0)
dtlbis$y.b <- ifelse(dtlbis$y>300,1,0)
dtlgee$y.b <- ifelse(dtlgee$y>300,1,0)
```

1. Modèle non ajusté : régression logistique simple sans prise en compte de l'effet centre

```
modglm <- glm(y.b~x,data=dtl,family="binomial")
library(gtsummary)
# tableau gt summary avec modify footnote pour dire comment les OR ont été obtenus
# (en anglais)
tbl_glm <-tbl_regression(
  modglm,
  exponentiate = TRUE,
  label = list(x ~ "Duration of interview (X)")
) %>%
```

```

modify_footnote(
  estimate ~ "Odds Ratios (OR) calculated from logistic regression model
  ↵ without accounting for center effect."
)
tbl_glm

```

Characteristic	OR ¹	95% CI	p-value
Duration of interview (X)	4.33	1.03, 18.4	0.047

¹Odds Ratios (OR) calculated from logistic regression model without accounting for center effect.

Abbreviations: CI = Confidence Interval, OR = Odds Ratio

OR calculé caractérise la relation entre X et Y en ignorant l'effet centre et en supposant que toutes les observations sont indépendantes.

2. Estimation par bootstrap par grappe

Pour corriger l'erreur standard résultant de la corrélation intra-centre, on peut utiliser un bootstrap par grappe ou un estimateur sandwich.

```

# Bootstrap par grappe
modboot <- clusbootglm(
  y.b~x,
  data=dtl,
  clusterid=centre,
  family="binomial")
modboot$boot.coefs

```

```
(Intercept)           x
-0.07908927   1.47229199
```

```
modboot$boot.sds
```

```
(Intercept)           x
  0.1520077   0.7915578
```

```
exp(modboot$boot.coefs)[2]
```

```
           x
 4.359215
```

L'OR estimé après bootstrap est légèrement différent, et l'erreur standard est plus grande, reflétant l'incertitude accrue due à la corrélation intra-centre.

Le rapport entre le coefficient non exponentié et l'erreur standard permet de tester la présence d'une corrélation :

```
cat("Le rapport coefficient/erreur standard est de",
    round((modboot$boot.coefs[2]) / (modboot$boot.sds[2]), 2),
    "\n")
```

Le rapport coefficient/erreur standard est de 1.86

Le rapport entre le coefficient (non exponentié) et son erreur standard correspond au z-score du test de Wald.

Il sert à tester l'hypothèse nulle H_0 : le coefficient de X est nul ($OR = 1$), c'est-à-dire « pas d'association entre X et Y » après correction de la corrélation intra-centre.

Si ce z-score est inférieur en valeur absolue à 1,96, on ne met pas en évidence d'association significative au seuil de 5 %.

Ici, le coefficient vaut 1,87 donc $< 1,96 \rightarrow$ pas de preuve d'une association significative entre X et Y après correction de la corrélation intra-centre.

3. Modèle ajusté à effet fixe : régression logistique avec ajustement sur la variable catégorielle [centre]

```
modglmaj <- glm(y.b~x+centre,data=dtl,family="binomial")
summary(modglmaj)$coefficients["x", ]
```

```
Estimate Std. Error      z value   Pr(>|z|)
2.19692245 1.24982624 1.75778230 0.07878456
```

```
exp(summary(modglmaj)$coefficients["x", "Estimate"])
```

[1] 8.997281

Le problème est que le modèle compte 199 variables indicatrices (dummy) pour les centres, ce qui est lourd.

Il faudrait au mieux 2000 observations pour respecter la règle des 10 événements par variable, or il y en a 535.

4. Modèle logistique conditionnel :

La fonction `clogit()` permet de faire une régression logistique conditionnelle en utilisant la vraisemblance partielle.

Vraisemblance partielle = méthode d'estimation qui permet d'éliminer les paramètres de nuisance (ici, les intercepts spécifiques à chaque centre) en conditionnant sur le nombre d'événements observés dans chaque groupe.

Au lieu d'estimer la probabilité absolue de l'événement, on estime la probabilité qu'un individu ait l'événement sachant le nombre total d'événements observés dans son centre.

Cela permet d'estimer l'association intra-centre sans avoir à estimer les coefficients de chaque centre.

C'est mathématiquement équivalent à un modèle de Cox stratifié.

```
library(survival)
summary(clogit(y.b~x+strata(centre),dtl))
```

Call:

```
coxph(formula = Surv(rep(1, 1000L), y.b) ~ x + strata(centre),
      data = dtl, method = "exact")
```

n= 1000, number of events= 535

	coef	exp(coef)	se(coef)	z	Pr(> z)
x	1.754	5.779	1.115	1.573	0.116

	exp(coef)	exp(-coef)	lower .95	upper .95
x	5.779	0.173	0.6494	51.43

Concordance= 0.551 (se = 0.037)
Likelihood ratio test= 2.48 on 1 df, p=0.1
Wald test = 2.47 on 1 df, p=0.1
Score (logrank) test = 2.49 on 1 df, p=0.1

Le coefficient obtenu correspond au Log-Odds Ratio intra-centre.

5. Modèle mixte à pente commune

Ici, la pente commune correspond à une pente commune au sein des différents centres.

```
moglmer1 <- glmer(y.b~x+(1|centre),data=dtl,family="binomial")
summary(moglmer1)$coefficients["x", ]
```

	Estimate	Std. Error	z value	Pr(> z)
x	1.92515054	1.01661177	1.89369294	0.05826578

```
exp(summary(moglmer1)$coefficients["x", "Estimate"])
```

[1] 6.856181

6. Modèle mixte autorisant des pentes propres à chaque centre

```
modglmer2 <- glmer(y.b~x+(1+x|centre),data=dtl,family="binomial")
```

boundary (singular) fit: see help('isSingular')

```
summary(modglmer2)$coefficients["x", ]
```

Estimate	Std. Error	z value	Pr(> z)
2.3676541	1.0784455	2.1954323	0.0281326

```
exp(summary(modglmer2)$coefficients["x", "Estimate"])
```

[1] 10.67233

boundary (singular) fit: see help('isSingular') nous alerte sur un possible problème numérique : cela signifie que le modèle a du mal à estimer certains paramètres en raison d'une matrice de variance-covariance singulière, souvent causée par un manque de variabilité dans les données ou une sur-paramétrisation du modèle.

7. Modèle marginal GEE avec fonction de lien logit

```
modgee <- gee(  
  y.b~x,  
  data=dtlgee,  
  id=centre,  
  corstr="exchangeable",  
  family="binomial"  
)
```

Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27

running glm to get initial regression estimate

(Intercept)	x
-0.07879536	1.46485165

```
summary(modgee)$coefficients["x", ]
```

Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
1.1122609	0.6119977	1.8174265	0.6426709	1.7306851

```
exp(summary(modgee)$coefficients["x", "Estimate"])
```

[1] 3.041227

Au final, on a 7 OR différents avec des IC et des pvalue différentes :

OR selon la méthode d'estimation

Estimation Method	Odds Ratio (OR)	Standard Error	p-value
Non ajuste (GLM)	4.33	0.74	0.0466
Bootstrap ajuste (clusbootglm)	4.36	0.79	0.0629
Effet fixe ajuste (GLM + centre)	9.00	1.25	0.0788
Conditionnel (clogit)	5.78	1.12	0.1157
Pente commune mixte (glmer)	6.86	1.02	0.0583
Pente aléatoire mixte (glmer)	10.67	1.08	0.0281
Marginal (GEE)	3.04	0.64	0.0835

Quel modèle choisir en pratique ?

Le choix du modèle dépend avant tout de la question scientifique et de la manière dont l'échantillon de centres est considéré.

1. Objectif : mesurer une association globale dans la population (OR marginal)

Si l'échantillon a vocation à représenter une population cible, et que la question est :

« Quelle est la force de l'association entre Y.b et X globalement, dans cette population, en tenant compte du fait que les sujets sont regroupés par centre ? »,

alors l'approche logistique simple avec correction de l'erreur standard par bootstrap de grappes est souvent la plus naturelle.

Concrètement, on ajuste un modèle logistique standard `glm(y.b ~ x, family = binomial)` puis on corrige l'incertitude (erreur standard, IC, p-value) par un bootstrap par centre.

- La pente estimée (et l'OR correspondant) reste très transparente à interpréter : c'est un OR marginal moyen, sur l'ensemble des sujets, exposés vs non exposés.
- Le bootstrap corrige l'optimisme de l'erreur standard dû à la corrélation intra-centre, sans introduire de structure de modèle supplémentaire.
- L'approche reste techniquement simple et robuste, au prix d'un peu de calcul.

Dans la même philosophie, un modèle GEE logistique (avec `gee()` et `family = binomial`) fournit aussi un OR marginal, mais cette fois-ci via une construction plus sophistiquée, qui impose de choisir une structure de corrélation (ex. « exchangeable »). Dans les situations standard, la logistique simple + bootstrap par centre suffit souvent, et a l'avantage d'être plus transparente.

2. Objectif : contrôler l'effet centre comme facteur de confusion (OR conditionnel)

Si la question est :

« Quelle est la relation entre Y.b et X à l'intérieur des centres, en traitant le centre comme un facteur de confusion ou de nuisance ? »,

alors il s'agit d'estimer un OR conditionnel au centre.

Plusieurs modèles répondent à cette logique :

- la régression logistique conditionnelle (`clogit(y.b ~ x + strata(centre))`) ;

- le modèle mixte à pente commune ($\text{glmer}(y.b \sim x + (1|\text{centre}))$) ;
- le modèle mixte avec pentes aléatoires par centre ($\text{glmer}(y.b \sim x + (1 + x|\text{centre}))$).

Dans tous les cas, l'OR est conditionnel au centre : il répond à une question du type « à centre donné, quelle est l'association entre X et Y.b ? ».

Cette interprétation est plus délicate, car l'OR dépend d'une information (le centre) qui est rarement observable ou utilisable en pratique au moment de la prise de décision.

Ces modèles sont théoriquement valides, mais :

- reposent sur des hypothèses fortes (homogénéité des effets, distribution normale des effets aléatoires, structure de corrélation, etc.) ;
- peuvent conduire à des problèmes numériques (convergence, singularité) dès que la structure devient un peu complexe (pentes aléatoires, peu d'événements par centre, centres très hétérogènes).

Il s'agit de modèles puissants, mais qui nécessitent un usage prudent et une expertise spécifique, surtout pour les modèles mixtes avec pentes aléatoires.

3. Modèle à effets fixes de centre : cas très limité

La régression logistique avec effet fixe de centre ($\text{glm}(y.b \sim x + \text{centre, family = binomial})$) introduit une variable indicatrice pour chaque centre.

En pratique, ce modèle est à éviter dès que le nombre de centres est un peu important :

- il consomme énormément de degrés de liberté (une vingtaine de centres = une vingtaine de paramètres supplémentaires) ;
- il viole rapidement la règle « 10 événements par variable » ;
- il n'apporte pas d'information synthétique sur la variabilité entre centres.

Il n'est raisonnable que si le nombre de centres est très faible (par exemple < 5) et que chaque centre dispose de beaucoup de sujets.

4. Marge vs conditionnel : conséquences sur l'OR

Les modèles GEE donnent un OR marginal, c'est-à-dire une comparaison des cotes de prévalence de la maladie chez tous les exposés vs tous les non exposés, après « gommage » de l'effet centre.

À l'inverse, la logistique conditionnelle et les modèles mixtes (pente commune ou pentes aléatoires) produisent des OR conditionnels au centre. En présence d'une forte variabilité du risque de base entre centres, ces OR conditionnels peuvent être beaucoup plus grands que l'OR marginal issu d'un GEE ou d'un modèle simple corrigé par bootstrap.

En résumé :

- OR marginal (GLM + bootstrap, GEE) : mesure l'effet « moyen » dans la population globale ;
- OR conditionnel (clogit, GLMM) : mesure l'effet « à centre donné », avec une interprétation plus abstraite.

5. Synthèse pratique

- Si l'objectif est une mesure simple, robuste et interprétable de la force d'association globale entre X et Y.b dans une population représentée par l'échantillon, la solution la plus raisonnable

est :

- modèle logistique simple $\text{glm}(y.b \sim x, \text{family} = \text{binomial})$
- avec correction de l'erreur standard par bootstrap de grappes sur le centre.
- Si l'objectif est de contrôler strictement l'effet du centre comme facteur de confusion et d'obtenir un OR au sein des centres, la logistique conditionnelle ou un modèle mixte à pente commune sont des candidats possibles, mais leur usage doit rester réservé aux situations où la question scientifique l'exige vraiment et avec un contrôle soigneux des hypothèses et de la convergence.
- Les modèles mixtes avec pentes aléatoires n'apportent un gain réel que si l'on souhaite modéliser explicitement l'hétérogénéité de l'effet de X d'un centre à l'autre et si l'échantillon contient suffisamment d'information pour les estimer correctement.
- Les GEE binaires se justifient surtout lorsque l'on souhaite un OR marginal avec une estimation plus « théorique » de la variance, en acceptant une certaine complexité de mise en œuvre et d'interprétation.

En pratique, sauf question très spécifique sur la structure centre par centre, l'association globale X-Y.b dans une étude multicentrique sera souvent décrite de façon honnête et pragmatique par un modèle logistique simple combiné à un bootstrap par grappe sur le centre, en explicitant clairement que l'OR rapporté est un OR marginal ajusté pour la corrélation intra-centre.