Epidémiologie

Plan

1. Introduction
2. Enquête de cohorte
3. Enquête cas-témoins
4. Mesures de risques
5. Mesures d'association
6. Biais de sélection
7. Biais de classement
8. Biais de confusion
9. Stratégie d'analyse
10. Puissance
11. Modèles multivariés
12. Régression logistique

Philippe SAINT PIERRE

Université Paul Sabatier – Toulouse III

Institut de Mathématiques de Toulouse

philippe.saint-pierre@math.univ-toulouse.fr

Epidémiologie

11. Modèles multivariés

Philippe SAINT PIERRE

Université Paul Sabatier – Toulouse III

Institut de Mathématiques de Toulouse

philippe.saint-pierre@math.univ-toulouse.fr

11. Modèles multivariés

I. Principaux modèles multivariés

- Régression linéaire
- Régression logistique
- Modèle de Cox (analyse de survie)

II. Concepts de l'analyse multivariée

- Maximum de vraisemblance
- Intervalle de confiance
- Tests statistiques
- Interaction entre variables
- Codage des variables
- Sélection de modèle

I. Principaux modèles multivariés

- Dans les cours précédents
  - Relation univariée entre E et M ⟹ ne tient pas compte d'autres facteurs (covariables)
  - Méthode de Mantel-Haenszel ⟹ ajustement possible sur un nombre limités de facteurs F
    M et E doivent être binaire et F qualitatif
- Modèles multivariés
  - Prise en compte de plusieurs covariables simultanément avec leurs interactions
  - Effet d'une variable ajusté sur les autres variables
  - Covariables peuvent être qualitatives ou quantitatives
- Contexte
  - Soit Y une variable à expliquer
  - Soit $X_1, X_2, \dots, X_k$ des variables explicatives
  - On cherche à expliquer la variables Y par les variables $X_1, X_2, \dots, X_k$

Variables à expliquer et variables explicatives

- Les variables explicatives peuvent être de nature
  - Qualitative à 2 classes ou binaire (Ex: Non fumeur / Fumeur)
  - Qualitative ordonnée (Ex: Non fumeur / Fumeur passif / Fumeur)
  - Qualitative non ordonnée (Ex: Chômeur / Etudiants / Actif / Au foyer / Retraité)
  - Quantitative (Ex: Nombre de cigarettes par jour)

- La variable à expliquer peut être de nature
  - Quantitative (Ex : terme de naissance) ⟹ Régression linéaire
  - Binaire (Ex: terme<37sem / terme ≥ 37 sem) ⟹ Régression logistique
  - Qualitative (Ex : terme<32 / terme 32-37 / terme ≥ 37) ⟹ Régression logistique ordinale
  - Temps avant un évènement (Ex: temps avant le décès) ⟹ Modèle de Cox (analyse de survie)

Régression linéaire

- La variable à expliquer Y est quantitative
- La moyenne de la variable Y est exprimée comme une fonction linéaire des variables explicatives $X_1, X_2, \dots, X_k$

$$E(Y|X_1, X_2, \dots, X_k) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

- Ou encore

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon \quad \text{avec (en général)} \quad \varepsilon \sim N(0, \sigma^2)$$

- Estimation des paramètres de régression $\alpha, \beta_1, \beta_2, \dots, \beta_k$
  - Méthodes des moindres carrés

$$\hat{\beta} = \min_{\beta} \|Y - (\alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)\|^2$$

  - Maximum de vraisemblance (identique à l'estimateur des moindres carrés dans le cas gaussien)

Régression logistique

- La variable à expliquer Y est binaire (Ex: la maladie dans une enquête cas-témoins)

- La probabilité que $Y = 1$ est exprimée comme une fonction logistique d'une combinaison linéaire des variables explicatives $X_1, X_2, \dots, X_k$

$$P(Y = 1|X_1, X_2, \dots, X_k) = \frac{1}{1 + e^{-(\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k)}}$$

- Estimation des paramètres de régression $\alpha, \beta_1, \beta_2, \dots, \beta_k$
  - Maximum de vraisemblance

- Fonction logistique $y = \frac{1}{1+e^{-x}}$

Modèle de Cox (analyse de survie)

- La variable à expliquer Y est une durée avant un évènement (Ex : données de cohorte)

- L'évènement d'intérêt (Ex : décès) peut être observé ou non observé
  - Evènement non observé à cause de la censure (perdus de vue, exclus vivants)
  - On sait si l'évènement à eu lieu ou non
  - La durée d'intérêt n'est pas toujours observée ⟹ Analyse de survie

- Le taux d'incidence $\lambda(t)$ est exprimé comme un risque de base et d'une fonction log-linéaire des variables explicatives $X_1, X_2, \dots, X_k$

$$\lambda(t|X_1, X_2, \dots, X_k) = \lambda_0(t) \times e^{\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k}$$

- Estimation des paramètres de régression $\beta_1, \beta_2, \dots, \beta_k$ ⟹ Vraisemblance partielle de Cox

II. Concepts de l'analyse multivariée

Quelques notions essentielle pour l'analyse multivariée

- Maximum de vraisemblance
- Intervalles de confiance
- Tests statistiques
- Interaction entre variables
- Codage des variables
- Sélection de modèle

Maximum de vraisemblance

- Vraisemblance: consiste à calculer la probabilité d'observer un échantillon de données
- En général,
  - les observations X_i d'un échantillon sont supposées $i.i.d.$ (indépendantes et identiquement distribuées)
  - Les observations sont indépendantes et suivent la même la loi de probabilité connue de densité paramétrique $f_\theta(.)$ ⟹ la densité dépend du même paramètre θ
  - La vraisemblance dépend du paramètre d'intérêt θ

$$L(\theta) = f_\theta(x_1, \dots, x_n|\theta) = \prod_{i=1}^{n} f_\theta(x_i|\theta)$$

  - Le paramètre θ peut être estimé par l'estimateur du maximum de vraisemblance

$$\hat{\theta} = \max_{\theta \in \Theta} L(\theta)$$

Maximum de vraisemblance

- L'estimateur du maximum de vraisemblance $\hat{\theta}$ a de bonnes propriétés mathématiques
  - Estimateur converge presque sûrement : $\hat{\theta} \xrightarrow[n\to\infty]{} \theta$
  - En général, l'estimateur asymptotiquement sans biais : $E(\hat{\theta}) \xrightarrow[n\to\infty]{} \theta$
  - Estimateur de variance minimale parmi les estimateurs sans biais (borne de Cramer-Rao)
  - La variance de l'estimateur $\hat{\theta}$ peut être estimée par l'inverse de la matrice d'information de Fisher : $Var(\hat{\theta}) = \{I(\hat{\theta})\}^{-1}$ où $I(\theta) = -E\left(\frac{\partial^2 \ln L(\theta)}{\partial^2 \theta}\right)$
  - Estimateur asymptotiquement normal : $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow[n\to\infty]{} N(0, I(\theta)^{-1})$

⟹ permet de construire des tests statistiques: Wald, Likelihood Ratio Test, Score

Exemple

- Echantillon d'observations $X_1, \dots, X_n$ de loi de Bernoulli $B(p)$
- Ex : $X_i = 0$ si l'individu est non malade et $X_i = 1$ si l'individu est malade

  le paramètre p correspond à la probabilité d'être malade
- Sur un échantillon de taille n, on observe k individus malades

$$L(p) = {C_n}^k \prod_{i=1}^{k} P(X_i = 1) \prod_{i=1}^{n-k} P(X_i = 0) = {C_n}^k p^k (1-p)^{n-k}$$

- En remarquant que $k = \sum_{i=1}^{n} X_i$ et en annulant la dérivée de $L(p)$ par rapport à p

$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- Ex: $n = 20$ et $p = 5 \rightarrow \hat{p} = 0.25$

  $\rightarrow$ $L(0.1) = 0.03$; $L(0.5) = 0.015$; $L(0.25) = 0.2$

Intervalle de confiance

- Normalité asymptotique de l'estimateur du maximum de vraisemblance permet d'obtenir les intervalles de confiance des coefficients de régression

  - La matrice d'information de Fisher permet d'obtenir une estimation de la variance des coefficients de régression : soit $\hat{s}_{\hat{\theta}}$ une estimation de l'écart-type de $\hat{\theta}$

  - Intervalle de confiance de θ de niveau α

$$IC_{\alpha}(\theta) := [\theta_{inf}; \theta_{sup}] = \hat{\theta} \pm z_{\frac{\alpha}{2}} \sqrt{\widehat{Var}(\hat{\theta})} = \hat{\theta} \pm z_{\frac{\alpha}{2}} \times \hat{s}_{\hat{\theta}}$$

  - Soit $\theta = (\theta_1, \theta_2)$, un intervalle de confiance de $\theta_1 + \theta_2$ de niveau α

$$IC_{\alpha}(\theta_1 + \theta_2) := (\hat{\theta}_1 + \hat{\theta}_2) \pm z_{\frac{\alpha}{2}} \times \hat{s}_{(\hat{\theta}_1 + \hat{\theta}_2)} \quad \text{avec } \hat{s}_{(\hat{\theta}_1 + \hat{\theta}_2)} = \sqrt{\widehat{Var}(\hat{\theta}_1) + \widehat{Var}(\hat{\theta}_2) + 2\widehat{Cov}(\hat{\theta}_1, \hat{\theta}_2)}$$

Tests statistiques

- Normalité asymptotique de l'estimateur du maximum de vraisemblance permet de définir 3 statistiques de test utiles pour comparer des modèles emboîtés
- Soit $\theta = (\theta_1, \dots, \theta_k) \in \mathbb{R}^k$, on souhaite tester des hypothèses de la forme

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases}$$

  - Test de Wald (rejet de H_0 si χ_W> quantile à 95% d'une Chi-2 à k ddl)

$$\chi_W = \left(\hat{\theta} - \theta_0\right)' I\left(\hat{\theta}\right)\left(\hat{\theta} - \theta_0\right) \underset{H_0}{\rightarrow} \chi(k)$$

  - Test du rapport de vraisemblance (rejet de H_0 si χ_{LRT}> quantile à 95% d'une Chi-2 à k ddl)

$$\chi_{LRT} = 2\left(\ln L(\hat{\theta}) - \ln L(\theta_0)\right) \underset{H_0}{\rightarrow} \chi(k)$$

  - Test du score (rejet de H_0 si χ_S> quantile à 95% d'une Chi-2 à k ddl)

$$\chi_S = \left(\left.\frac{\partial \ln L(\theta)}{\partial \theta}\right|_{\theta_0}\right)' I(\theta_0)^{-1} \left(\left.\frac{\partial \ln L(\theta)}{\partial \theta}\right|_{\theta_0}\right) \underset{H_0}{\rightarrow} \chi(k)$$

Tests statistiques

- Ex : modèle de Cox $\rightarrow \lambda(t|X_1, X_2, \dots, X_k) = \lambda_0(t) \times e^{\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k}$
- En général : tester le lien entre une variable explicative et la variable à expliquer

$$\begin{cases} H_0 : \beta_i = 0 \\ H_1 : \beta_i \neq 0 \end{cases}$$

  - Rejet de $H_0 : \beta_i = 0 \rightarrow$ association statistiquement significative
  - Non rejet de $H_0 : \beta_i = 0 \rightarrow$ on ne rejette pas l'absence d'association

- On peut tester l'effet de chaque variable séparément avec les tests précédents

Soit $\hat{\beta}_0 = (\hat{\beta}_1, \dots, \hat{\beta}_{i-1}, 0, \hat{\beta}_{i+1}, \dots, \hat{\beta}_k)$

  - Test de Wald : $\chi_W = (\hat{\beta} - \hat{\beta}_0)' I(\hat{\beta})(\hat{\beta} - \hat{\beta}_0) = \frac{\widehat{\beta_i}^2}{\widehat{\sigma}^2(\widehat{\beta_i})} \underset{H_0}{\rightarrow} \chi(1)$ (ou encore $\frac{\widehat{\beta_i}}{\hat{s}_{\widehat{\beta_i}}} \underset{H_0}{\rightarrow} N(0,1)$)
  - Test du rapport de vraisemblance : $\chi_{LRT} = 2\left(\ln L(\hat{\beta}) - \ln L(\hat{\beta}_0)\right) \underset{H_0}{\rightarrow} \chi(1)$
  - Test du score : $\chi_S = \left(\frac{\partial \ln L(\beta)}{\partial \beta}\Big|_{\hat{\beta}_0}\right)' I(\hat{\beta}_0)^{-1} \left(\frac{\partial \ln L(\beta)}{\partial \beta}\Big|_{\hat{\beta}_0}\right) \underset{H_0}{\rightarrow} \chi(1)$

Tests statistiques

- Ces tests permettent de comparer les modèles emboîtés entre eux

↳ Le nombre de ddl des lois de Chi-2 est égal à la différence entre le nombre de paramètres de chaque modèle

- Exemple avec le test du rapport de vraisemblance
  - Modèle 1 (complet) : $\lambda(t|X_1, X_2, \dots, X_k) = \lambda_0(t) \times e^{\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k}$ ⟹ Vraisemblance L_1
  - Modèle 2 (X_2 n'a pas d'effet) : $\lambda(t|X_1, X_2, \dots, X_k) = \lambda_0{}'(t) \times e^{\beta_1{}' X_1 + \beta_3{}' X_3 + \dots + \beta_k{}' X_k}$ ⟹ Vraisemblance L_2
  - Modèle 3 (X_1 et X_2 n'ont pas d'effet) : $\lambda(t|X_1, X_2, \dots, X_k) = \lambda_0{}''(t) \times e^{\beta_3{}'' X_3 + \dots + \beta_k{}'' X_k}$ ⟹ Vraisemblance L_3
  - Test de l'association de la variable $X_2 \rightarrow H_0 : \beta_2 = 0 \quad (H_1 : \beta_i \neq 0)$

$$\chi_{LRT} = 2(\ln L_1 - \ln L_2) \underset{H_0}{\rightarrow} \chi(k - (k-1)) \equiv \chi(1)$$

  - Test simultanément les associations des variables X_1 et $X_2 \rightarrow H_0 : \beta_1 = \beta_2 = 0 \quad (H_1 : \exists \beta_i \neq 0)$

$$\chi_{LRT} = 2(\ln L_1 - \ln L_3) \underset{H_0}{\rightarrow} \chi(k - (k-2)) \equiv \chi(2)$$

  - Remarque : les modèles 2 et 3 sont emboîtés dans le modèle 1 (cas particuliers du modèle 1)

Interaction entre variables

- Exemple du modèle de Cox
  - Modèle sans interaction : $\lambda(t|X_1, X_2) = \lambda_0(t) \times e^{\beta_1 X_1 + \beta_2 X_2}$ ⟹ Vraisemblance L_1
  - Modèle avec interaction : $\lambda(t|X_1, X_2) = \lambda_0'(t) \times e^{\beta_1' X_1 + \beta_2' X_2 + \gamma X_1 X_2}$ ⟹ Vraisemblance L_2

- Test de l'interaction
  - $\begin{cases} H_0 : \gamma = 0 \\ H_1 : \gamma \neq 0 \end{cases}$
  - Test de Wald : $\chi_W = \frac{\hat{\gamma}}{\hat{s}_{\hat{\gamma}}} \underset{H_0}{\rightarrow} N(0,1)$
  - Test du rapport de vraisemblance : $\chi_{LRT} = 2(\ln L_2 - \ln L_1) \underset{H_0}{\rightarrow} \chi(1)$

- Effet de la variable X_1 (Exemple avec X_1 et X_2 binaires)
  - Modèle sans interaction : e^{β_1} (le risque est multiplié par e^{β_1} quand $X_1 = 1$)
  - Modèle avec interaction : e^{β_1} quand $X_2 = 0$
    $e^{\beta_1 + \gamma}$ quand $X_2 = 1$ (Rq: IC d'une somme de paramètres)

- Interprétation complexe → l'effet dépend de la valeur des autres variables
  → limiter le nombre d'interactions et les interactions d'ordre supérieur

Codage des variables

- Codage d'une variable qualitative à deux classes
  - Codage naturel $\rightarrow X = 0$ et $X = 1$
  - Ex : modèle de Cox $\rightarrow \lambda(t|X) = \lambda_0(t) \times e^{\beta X}$
    - Le risque est multiplié par e^{β} pour les patients codés 1 par rapport aux patients codés 0
  - Si on utilise un autre codage $\rightarrow X = 2$ et $X = 5$
    - L'estimation du coefficient de régression sera différente $\rightarrow \beta'$
    - Néanmoins, l'interprétation reste identique
    - Le risque est multiplié par $e^{3\beta'} = e^{\beta}$ pour les patients codé 5 par rapport aux patients codés 2
  - Le codage n'a pas d'importance

Codage des variables

- Codage d'une variable qualitative à k **classes ordonnées**
  - Ex : Niveau d'étude → primaire, secondaire, supérieur
  - **Codage en une variable binaire** ➡ facilité d'interprétation mais perte d'informations
  - **Codage "linéaire" en k classes**
    - $X = 0$ (primaire), $X = 1$ (secondaire) et $X = 2$ (supérieur)
    - Le codage implique une relation linéaire pour l'interprétation (dépend du codage)
      - Ex : modèle de Cox → $\lambda(t|X) = \lambda_0(t) \times e^{\beta X}$
      - Risque multiplié par e^{β} pour les patients codés 1 par rapport aux patients codés 0
      - Risque multiplié par $e^{2\beta}$ pour les patients codés 2 par rapport aux patients codés 0
      - Risque multiplié par e^{β} pour les patients codés 2 par rapport aux patients codés 1
    - Rien ne justifie cette relation linéaire → il faut tester cette hypothèse

Codage des variables

- Codage d'une variable qualitative à k **classes ordonnées**
  - Ex : Niveau d'étude → primaire, secondaire, supérieur
  - **Codage "dichotomique" en utilisant k-1 variables binaires**
    - Soit X_0 et X_1 deux variables binaires

Niveau d'étude	Codage linéaire	Codage dichotomique (1)	
	X	X_0	X_1
primaire	0	0	0
secondaire	1	1	0
supérieur	2	0	1

Niveau d'étude	Codage linéaire	Codage dichotomique (2)	
	X	X_0	X_1
primaire	0	0	0
secondaire	1	1	0
supérieur	2	1	1

Ex : modèle de Cox → $\lambda(t|X) = \lambda_0(t) \times e^{\beta_0 X_0 + \beta_1 X_1}$

- Risque multiplié par e^{β_0} pour les patients "secondaire" par rapport aux patients "primaire"
- Risque multiplié par e^{β_1} pour les patients "supérieur" par rapport aux patients "primaire"
- Risque multiplié par $e^{(\beta_1 - \beta_0)}$ pour les patients "supérieur" par rapport aux patients "secondaire"

- Risque multiplié par $e^{(\beta_1 + \beta_2)}$ pour les patients "supérieur" par rapport aux patients "primaire"
- Risque multiplié par e^{β_2} pour les patients "supérieur" par rapport aux patients "secondaire"

Codage des variables

- Codage d'une variable qualitative à k **classes ordonnées**
  - Ex : Niveau d'étude → primaire, secondaire, supérieur
  - **Codage "dichotomique" en utilisant k-1 variables binaires**
    - Il n'y a plus d'hypothèse de linéarité
    - Les effets sont spécifiques à chaque classe
    - L'interprétation est indépendante du codage
    - Ce codage doit toujours être étudié car il ne fait aucune hypothèse
  - On compare ensuite le codage "dichotomique" avec le codage "linéaire" avec un test du rapport de vraisemblance

Codage des variables

- **Codage d'une variable qualitative à k classes ordonnées**

  - Comparaison du codage "linéaire" avec le codage "dichotomique"

    ↳ L_1 (1 paramètre) ↳ L_2 ($k-1$ paramètres)

  - Le modèle avec codage "linéaire" est emboîté dans le modèle avec codage "dichotomique"

$$\chi_{LRT} = 2(\ln L_2 - \ln L_1) \underset{H_0}{\rightarrow} \chi((k-1)-1) \equiv \chi(k-2)$$

  - Résultat du test → H_0: codage "linéaire" contre H_1: codage "dichotomique"
    - Résultat significatif (Rejet de H_0) : on conserve le codage "dichotomique"

      ↳ le codage dichotomique s'ajuste mieux aux données

    - Résultat non significatif (Non rejet de H_0) : on conserve le codage "linéaire"

      ↳ les deux codages apportent la même information: garder modèle avec le moins de paramètres

Codage des variables

- Codage d'une variable qualitative à k **classes ordonnées**

  - Exemple : Lien entre le niveau d'étude et le risque de développer la maladie d'Alzheimer

Niveau d'étude	M1: Codage binaire	M2: Codage linéaire	M3: Codage dichotomique	
	Z	X	X_0	X_1
primaire	0	0	0	0
secondaire	1	1	1	0
supérieur	1	2	0	1

Ex : modèle de Cox

- Codage "binaire" $\rightarrow \lambda(t|X) = \lambda_0(t) \times e^{\gamma Z}$
- Codage "linéaire" $\rightarrow \lambda(t|X) = \lambda_0(t) \times e^{\beta X}$
- Codage "dichotomique" $\rightarrow \lambda(t|X) = \lambda_0(t) \times e^{\beta_0 X_0 + \beta_1 X_1}$

Test du rapport de vraisemblance

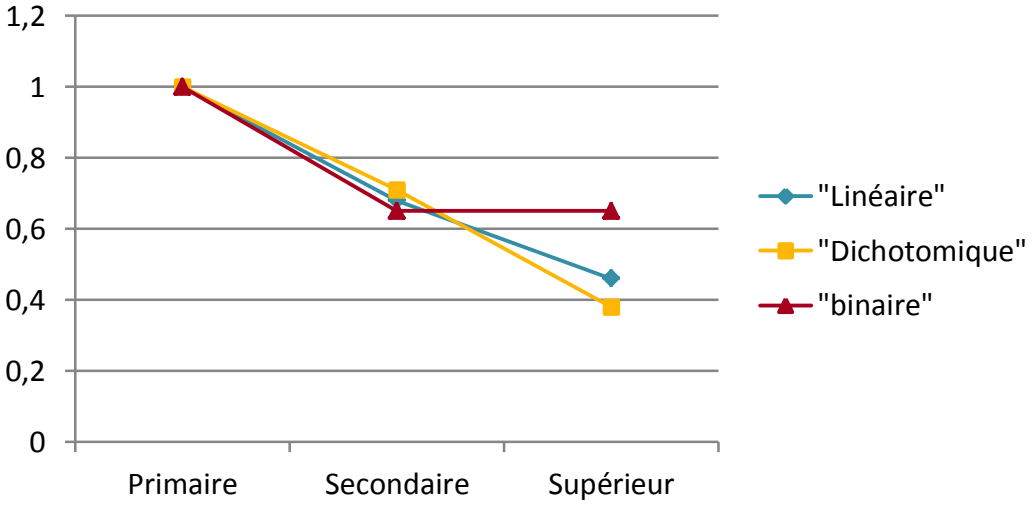
- M1 emboîté dans M3 $\rightarrow$ rejet du codage "binaire"

  $\chi_{LRT} = 2 \times (\ln L_{Dicho} - \ln L_{Bin}) = 5 > 3.84 = \chi_{95\%}(1)$

- M2 emboîté dans M3 $\rightarrow$ non rejet du codage "linéaire"

  $\chi_{LRT} = 2 \times (\ln L_{Lin} - \ln L_{Bin}) = 0.6 < 3.84 = \chi_{95\%}(1)$

- M1 n'est pas emboité dans M2



Codage des variables

- Codage d'une variable qualitative à k **classes non ordonnées**
  - Nécessite d'avoir un coefficient différent pour chaque classe car il n'y a pas d'ordre naturel

    ➡ Toujours utiliser en priorité un codage "dichotomique"

  - Exemple : Relation entre le cancer de la vessie et le type de tabac consommé

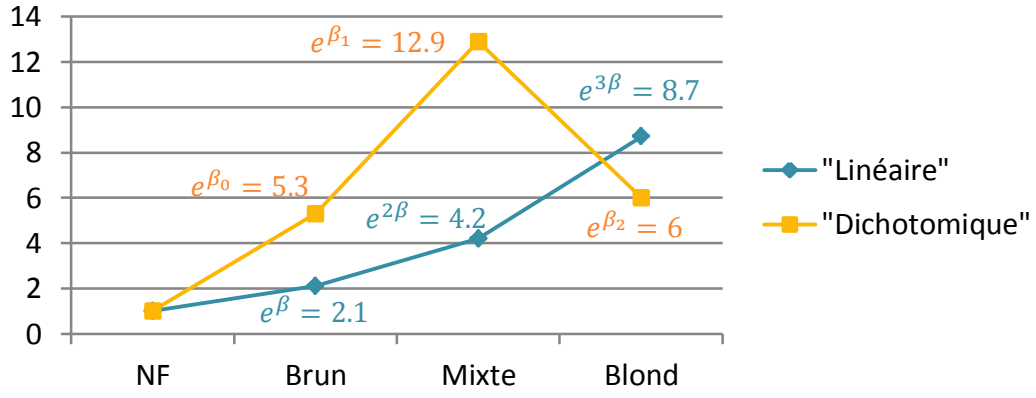
Type de tabac	Codage linéaire	Codage dichotomique		
	X	X_0	X_1	X_2
Non fumeur	0	0	0	0
Tabac brun	1	1	0	0
Tabac mixte	2	0	1	0
Tabac blond	3	0	0	1

Test du rapport de vraisemblance

- On rejette l'hypothèse de linéarité
- Le résultat peut dépendre du codage

Ex : modèle de Cox

- Codage "linéaire" $\rightarrow \lambda(t|X) = \lambda_0(t) \times e^{\beta X}$
- Codage "dichotomique" $\rightarrow \lambda(t|X) = \lambda_0(t) \times e^{\beta_0 X_0 + \beta_1 X_1 + \beta_2 X_2}$



Codage des variables

- Codage d'une variable quantitative
  - Recodage en une variable binaire → facilité d'interprétation mais perte d'informations
  - Recodage en une variable qualitative ordonnée
    - Facilité d'interprétation mais perte d'informations
    - Attention à l'hypothèse de linéarité → codage "linéaire" ou "dichotomique" à évaluer
  - Conserver une variable quantitative
    - Pas de perte d'informations (à privilégier si possible)
    - Permet d'avoir un seul coefficient → facilité d'interprétation
    - Attention à l'hypothèse de linéarité → à évaluer en utilisant des variables qualitatives
  - Attention: un modèle avec une variable quantitative ne sera jamais emboîté dans un modèle avec une variable qualitative (pas de comparaison possible avec les tests présentés)

Codage des variables

- Codage d'une variable quantitative

  - Exemple : Relation entre le cancer de la vessie et le nombre de cigarettes / jour

    Evaluer l'hypothèse de linéarité

Nombre de cigarettes	Codage linéaire	Codage dichotomique			
	X	X_0	X_1	X_2	X_3
0	0	0	0	0	0
[0;20[	1	1	0	0	0
[20;40[	2	0	1	0	0
[40;60[	3	0	0	1	0
≥ 60	4	0	0	0	1

Test du rapport de vraisemblance

$\ln L_{Lin} = -279$ et $\ln L_{Dicho} = -276$

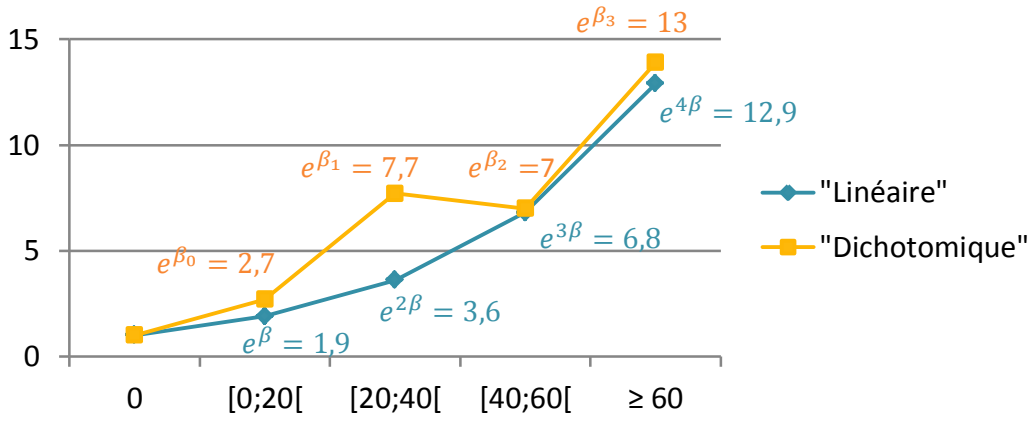
$\chi_{LRT} \underset{H_0}{\rightarrow} \chi(k-2) \equiv \chi(3)$

$\chi_{LRT} = 2 \times (-276 - (-279)) = 6 < 7.81 = \chi_{95\%}(3)$

⟶ On ne rejette pas l'hypothèse de linéarité

Ex : modèle de Cox

- Codage "linéaire" $\rightarrow \lambda(t|X) = \lambda_0(t) \times e^{\beta X}$
- Codage "dichotomique" $\rightarrow \lambda(t|X) = \lambda_0(t) \times e^{\beta_0 X_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3}$



Sélection de modèle

- Choix des variables à inclure dans le modèle initial
  - Facteurs de risque connus de la maladie (connaissances, bibliographie)
  - Variables d'appariement et/ou de stratification (car jamais parfait)
  - Facteurs de risques de la maladie identifiés avec l'analyse univariée

    ↳ Garder les variables tel que $p < 20\%$ ou 25%

- Attention : groupe de variable très corrélées entre elles
  - Faire des groupes de variables corrélées
  - Ex: variables socio-économiques → revenu, CSP, diplôme, scolarité
  - Dans chaque groupe, sélectionner une ou plusieurs variables à conserver

Sélection de modèle

- Codage des variables
  - Variables binaires
  - Variables qualitatives non ordonnées à k classes $\rightarrow$ codage "dichotomique"
  - Variables qualitatives ordonnées à k classes
    1. Modèle avec un codage utilisant la variable qualitative à k classes (codage "linéaire")
    2. Modèle avec un codage utilisant $k-1$ variables binaires (codage "dichotomique")
    3. Tester la linéarité en comparant les 2 codages (test du rapport de vraisemblance)
    4. Si le codage "linéaire" est rejeté (rejet de la linéarité) utiliser un codage dichotomique
  - Variable quantitative
    1. Modèle avec la variable quantitative
    2. Modèle avec un codage "linéaire"
    3. Modèle avec un codage "dichotomique"
    4. Si le codage "linéaire" est rejeté utiliser un codage dichotomique sinon utiliser la variable quantitative

Sélection de modèle

- Prise en compte des interactions
  - Rechercher les interactions éventuelles par une analyse stratifiée (Mantel-Haenszel)
  - En pratique, on recherche rarement les interactions d'ordre supérieur à 2 (Ex: $X_1 X_2 X_3$)
  - Retenir le minimum de termes d'interaction car interprétation devient difficile
  - En présence d'interaction entre X_1 et X_2
    - L'effet de X_1 varie en fonction de la valeur de X_2
    - Donner l'estimation de l'effet dans chaque strate de X_2
      - Ex : modèle de Cox avec X_1 et X_2 binaires $\rightarrow$ e^{β_1} pour $X_2 = 0$ et $e^{\beta_1+\gamma}$ pour $X_2 = 1$

Sélection de modèle

- Sélection des variables à inclure dans le modèle final

  - Tests de Wald, du LRT et du score pour comparer des modèles emboîtés

    - Ex: modèle de Cox M1 : $\lambda(t|X_1, X_2, X_3) = \lambda_0(t) \times e^{\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3}$

      M2 : $\lambda(t|X_1, X_2) = \lambda_0{}'(t) \times e^{\beta_1{}' X_1 + \beta_2{}' X_2}$

      - Si le test rejette $H_0: \beta_3 = 0$ ($p < 0.01$) on conserve le modèle M1
      - Si le test ne rejette pas $H_0: \beta_3 = 0$ ("M1≈M2") on sélectionne le modèle M2 (moins de paramètre)
        - La variable X_3 n'est pas associée au risque d'évènement
        - Retirer la variable X_3 ne modifie pas trop la vraisemblance
        - Il faut vérifier que retirer la variable X_3 ne modifie pas trop l'estimation des coefficients restants

  - On peut aussi choisir le modèle qui $\rightarrow$ maximise le coefficient R^2

    $\rightarrow$ minimise les critères AIC, BIC, C_p de Mallows

Sélection de modèle

- Procédures de sélection automatique

  - Si le modèle de départ à p variables, il y a 2^p sous-modèles à explorer

    ↳ Utilisation d'algorithme de sélection efficaces pour ne pas explorer tous les sous-modèles

  - Procédure forward (ascendante) : on part du modèle avec aucune variable et on ajoute les variables une a une. A chaque pas, on ajoute celle qui améliore le plus le modèle (la plus petite p-value avec le test du LRT, Wald, Score, qui engendre la plus forte augmentation ou diminution du R^2, AIC, BIC, C_p de Mallows). On s'arrête quand aucune variable n'améliore le modèle.

  - Procédure backward (descendante): on part du modèle avec toutes les variables et on retire les variables une a une. A chaque pas, on retire celle qui améliore le plus le modèle en étant supprimée.

  - Procédure stepwise : mixte des deux premières procédures (on ajoute ou on supprime à chaque étape)

  - En général on évite la procédure forward (l'effet d'une variable peut changer quand on ajoute d'autres variables)

Sélection de modèle

- Problème lié à la multiplicité des tests (comparaisons multiples)

  - Lorsque plusieurs tests sont réalisés successivement le risque d'erreur global de 1ère espèce α augmente avec le nombre de test

  - On montre que pour k tests de niveau α, le risque d'erreur global est

$$\alpha_{global} = 1 - (1 - \alpha)^k > \alpha$$

  - Il faut utiliser des méthodes pour corriger la multiplicités des tests

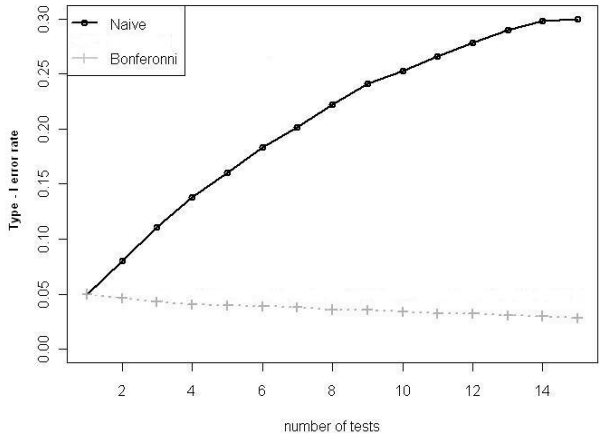
    - Méthode de Bonferroni : utiliser un risque $\frac{\alpha}{k}$ au lieu de α

      Le risque global est < α mais la puissance du test devient faible

    - Méthode de Holm-Bonferroni (plus puissante)

    - Correction de Šidák

    - False discovery rate



Sélection de modèle

- Règles hiérarchiques pour le retrait des variables
  - En présence d'interaction → conserver les variables qui interviennent dans une interaction
  - En cas de codage dichotomique
    - Conserver toutes les variables utilisés pour le codage dans le modèle
    - Sinon l'interprétation est modifiée !
    - Ex : Catégorie socio-professionnelle

Catégorie socio-professionnelles	Codage linéaire	Codage dichotomique			
	X	X_0	X_1	X_2	X_3
Ouvrier	0	0	0	0	0
Cadre	1	1	0	0	0
Agriculteur	2	0	1	0	0
Chômeur	3	0	0	1	0
Autres	4	0	0	0	1

- e^{β_1} représente l'augmentation du risque pour un "Agriculteur" par rapport à la catégorie "Ouvrier"
- Si on supprime la variable X_2

  ↳ e^{β_1} devient l'augmentation du risque pour les agriculteurs par rapport à la catégorie "Ouvrier" + "Chômeur"

Sélection de modèle

- Equilibre à trouver entre
  - Modèle saturé (toutes les variables et toutes les interactions)
    - Bonne adéquation
    - Risque de sur-ajustement (perte de puissance pour étudier le lien entre la maladie et l'exposition)
    - Interprétation difficile (beaucoup de coefficients)
  - Modèle non saturé (modèle avec peu de variables)
    - Moins bonne adéquation du modèle
    - Possibilité de confusion résiduelle
    - Interprétation plus facile (moins de coefficient)

Epidémiologie

12. Régression logistique

Philippe SAINT PIERRE

Université Paul Sabatier – Toulouse III

Institut de Mathématiques de Toulouse

philippe.saint-pierre@math.univ-toulouse.fr

12. Régression logistique

I. Exemples

II. Modèle logistique

III. Estimation du maximum de vraisemblance

IV. Fonction Logit

V. Odds ratio

VI. Intervalles de confiance

VII. Tests statistiques

VIII. Interaction entre variables

IX. Méthodes alternatives

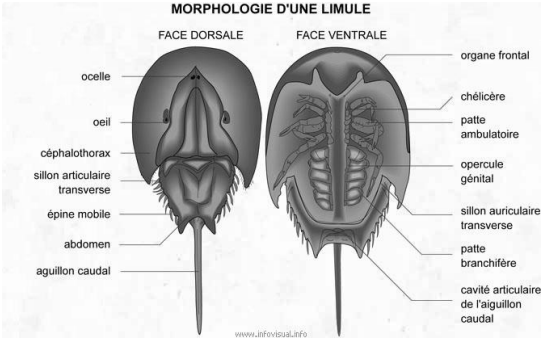
I. Exemples

- Exemple 1 : Enquête cas-témoins $\rightarrow$ étudier le lien entre une exposition et une maladie
  - Echantillon de m_1 malades et m_0 non malades
  - On observe le statut malade ($Y = 1$) ou non ($Y = 0$)
  - On observe l'exposition, fumeur ($X_1 = 1$) ou non ($X_1 = 0$)
  - On observe également d'autres covariables ($X_2, \dots, X_k$)

		Fumeur		
		E+	E-	
Cancer du poumon	M+	a	b	m_1
	M-	c	d	m_0

  - On cherche à étudier l'effet du tabac en tenant comte des covariables sur la probabilité d'être malade
  - Attention : on ne pourra pas estimer la probabilité d'être malade pour un individu (individus sélectionnés sur le statut malade non malades)

- Exemple 2 : Echantillon représentatif de 173 femelles limule
  - La présence ($Y = 1$) ou l'absence ($Y = 0$) de mâle dans l'entourage
  - La largeur de leur abdomen : X_1 en cm
  - La teinte de leur carapace : $X_2 = 1$ (foncée) ou $X_1 = 0$ (clair)



  - Objectif : expliquer la présence ou l'absence de partenaire en fonction des variables "largeur" et "teinte".

II. Modèle logistique

- Soit Y une variable binaire qu'on cherche à expliquer

- La probabilité que $Y = 1$ est exprimée comme une fonction logistique d'une combinaison linéaire des variables explicatives $X_1, X_2, \dots, X_k$

$$P(Y = 1|X_1, X_2, \dots, X_k) = \frac{1}{1 + e^{-(\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k)}}$$

- L'objectif est d'estimer les coefficients $\alpha, \beta_1, \dots, \beta_k$ à partir d'un échantillon

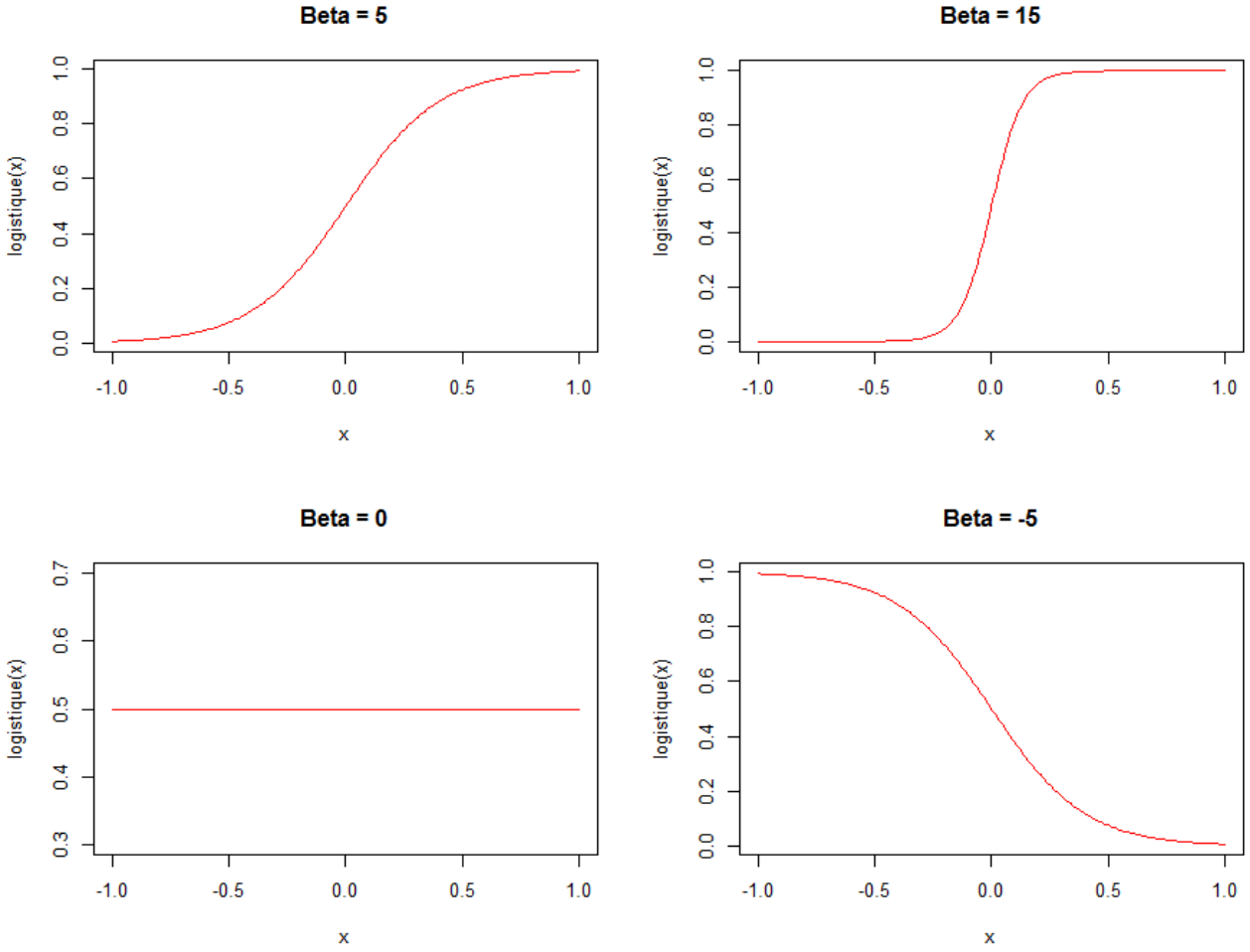
$$(Y_i, X_{i1}, \dots, X_{ik})_{i=1,\dots,n}$$

- Pour chaque individu, on pourra déduire la probabilité d'être malade (si indépendance des observations : pas le cas dans les enquêtes cas-témoins)

$$P(Y_i = 1|X_{i1}, X_{i2}, \dots, X_{ik}) = \frac{1}{1 + e^{-(\alpha+\beta_1 X_{i1}+\beta_2 X_{i2}+\cdots+\beta_k X_{ik})}}$$

Modèle logistique

- Fonction logistique : $y = \frac{1}{1+e^{-\beta x}} \rightarrow \beta \in \mathbb{R}, x \in \mathbb{R}, y \in [0,1]$



III. Estimation par maximum de vraisemblance

- Estimation des paramètres de régression $\alpha, \beta_1, \beta_2, \dots, \beta_k$ par maximum de vraisemblance à partir d'un échantillon $(Y_i, X_{i1}, \dots, X_{ik})_{i=1,\dots,n}$

$$L(\alpha, \beta_1, \dots, \beta_k) = \prod_{i=1}^{n} P(Y_i = y_i \mid X_{i1}, X_{i2}, \dots, X_{ik})$$

$$= \prod_{i=1}^{n} [P(Y_i = 1 \mid X_{i1}, X_{i2}, \dots, X_{ik})]^{y_i} [P(Y_i = 0 \mid X_{i1}, X_{i2}, \dots, X_{ik})]^{1-y_i}$$

$$= \prod_{i=1}^{n} \left[\frac{1}{1+e^{-(\alpha+\beta_1 X_{i1}+\beta_2 X_{i2}+\cdots+\beta_k X_{ik})}}\right]^{y_i} \left[\frac{e^{-(\alpha+\beta_1 X_{i1}+\beta_2 X_{i2}+\cdots+\beta_k X_{ik})}}{1+e^{-(\alpha+\beta_1 X_{i1}+\beta_2 X_{i2}+\cdots+\beta_k X_{ik})}}\right]^{1-y_i}$$

- Estimateur du maximum de vraisemblance

$$\hat{\beta} = (\hat{\alpha}, \widehat{\beta_1}, \dots, \widehat{\beta_k}) = \max_{\beta \in \mathbb{R}^{k+1}} \log(L(\alpha, \beta_1, \dots, \beta_k))$$

- Estimation par des méthodes itératives : algorithme de Nawton-Raphson

IV. Fonction Logit

- Modèle logistique

$$P(Y = 1|X_1, X_2, \dots, X_k) = \frac{1}{1 + e^{-(\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k)}}$$

- Considérons la fonction Logit(.)

$$\text{Logit}(p) = \ln\left(\frac{p}{1-p}\right)$$

- La fonction Logit(.) permet de "linéariser" $P(Y = 1|X_1, X_2, \dots, X_k)$

$$\text{Logit}(P(Y = 1|X_1, X_2, \dots, X_k)) = \ln\left(\frac{1}{e^{-(\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k)}}\right) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

V. Odds ratio

- Relation entre Logit(p) et l'odds ratio d'une enquête cas-témoins → M+ $(Y = 1)$ et M- $(Y = 0)$

  - Soit $P_1 = P(M+|E+)$ et $P_0 = P(M+|E-)$

$$OR = \frac{P_1/(1-P_1)}{P_0/(1-P_0)}$$

  - $\mathrm{Ln}(OR) = \mathrm{Ln}\left(\frac{P_1}{1-P_1}\right) - \mathrm{Ln}\left(\frac{P_0}{1-P_0}\right) = \mathrm{Logit}(P_1) - \mathrm{Logit}(P_0) = \mathrm{Ln}(OR)$

- Cas où il y a une seule variable explicative qui est l'exposition E

  - Modèle logistique : $P(M+|E) = \frac{1}{1+e^{-(\alpha+\beta E)}}$ ou $\mathrm{Logit}(P(M+|E)) = \alpha + \beta E$

  - $\mathrm{Ln}(OR) = \mathrm{Logit}(P_1) - \mathrm{Logit}(P_0) = \underbrace{\alpha + \beta \times 1}_{E=1 \text{ si E+}} - \underbrace{(\alpha + \beta \times 0)}_{E=0 \text{ si E-}} = \beta$

➡ $OR = \exp(\beta)$ → OR **brut** entre E et M

Odds ratio

- Cas où il y a d'autres variables explicatives en plus de l'exposition E

  - Modèle logistique : $P(M+|E,X_1,X_2,\dots,X_k) = \frac{1}{1+e^{-(\alpha+\beta E+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k)}}$

  - $\text{Logit}(P(M+|E,X_1,X_2,\dots,X_k)) = \alpha+\beta E+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k$

  - Soit $P_1 = P(M+|E+,X_1,X_2,\dots,X_k)$ et $P_0 = P(M+|E-,X_1,X_2,\dots,X_k)$

  - $\text{Ln}(OR) = \text{Logit}(P_1) - \text{Logit}(P_0)$

$$= \underbrace{\alpha+\beta\times 1+(\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k)}_{E=1 \text{ si E+}} - \underbrace{[\alpha+\beta\times 0+(\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k)]}_{E=0 \text{ si E-}} = \beta$$

➡ $OR = \exp(\beta)$ → OR entre E et M **ajusté** sur $X_1, X_2, \dots, X_k$

VI. Intervalles de confiance

- Modèle logistique $\rightarrow$ $\text{Logit}(P(M+|E)) = \alpha + \beta E$
  - Intervalle de confiance de β de niveau α

$$IC_\alpha(\beta) := [\beta_{inf}; \beta_{sup}] = \hat{\beta} \pm z_{\frac{\alpha}{2}} \sqrt{\widehat{Var}(\hat{\beta})}$$

  - Intervalle de confiance de $OR = \exp(\beta)$ de niveau α

$$IC_\alpha(OR) := \left[e^{\beta_{inf}}; e^{\beta_{sup}}\right]$$

- Modèle logistique $\rightarrow$ $\text{Logit}(P(M+|E, X_1)) = \alpha + \beta E + \beta_1 X_1 + \gamma E X_1$
  - Intervalle de confiance de $\beta + \gamma$ de niveau α

$$IC_\alpha(\beta+\gamma) := [\Gamma_{inf}; \Gamma_{sup}] = (\hat{\beta} + \hat{\gamma}) \pm z_{\frac{\alpha}{2}} \sqrt{\widehat{Var}(\hat{\beta}) + \widehat{Var}(\hat{\gamma}) + 2\widehat{Cov}(\hat{\beta}, \hat{\gamma})}$$

  - Intervalle de confiance de $OR = \exp(\beta + \gamma)$ de niveau α : $IC_\alpha := \left[e^{\Gamma_{inf}}; e^{\Gamma_{sup}}\right]$

    ↳ OR entre E et M en présence d'interaction dans la strate $X_1 = 1$

VII. Tests statistiques

- Modèle logistique $\rightarrow$ $\text{Logit}(P(M+|X_1, X_2, \dots, X_k)) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$

- Test de l'association entre une variable et la maladie

$$\begin{cases} H_0 : \beta_i = 0 \\ H_1 : \beta_i \neq 0 \end{cases} \Leftrightarrow \begin{cases} H_0 : OR_i = 1 \\ H_1 : OR_i \neq 1 \end{cases}$$

  - Rejet de $H_0 : \beta_i = 0$ $\rightarrow$ association statistiquement significative
  - Non rejet de $H_0 : \beta_i = 0$ $\rightarrow$ on ne rejette pas l'absence d'association

- Pour chaque variables on peut tester l'association entre la variable i et la maladie

Soit $\hat{\beta}_0 = (\hat{\beta}_1, \dots, \hat{\beta}_{i-1}, 0, \hat{\beta}_{i+1}, \dots, \hat{\beta}_k)$

  - Test de Wald : $\chi_W = (\hat{\beta} - \hat{\beta}_0)' I(\hat{\beta})(\hat{\beta} - \hat{\beta}_0) = \frac{\widehat{\beta_i}^2}{\widehat{Var}(\widehat{\beta_i})} \underset{H_0}{\rightarrow} \chi(1)$ (ou encore $\frac{\widehat{\beta_i}}{\hat{s}_{\widehat{\beta_i}}} \underset{H_0}{\rightarrow} N(0,1)$)

  - Test du rapport de vraisemblance : $\chi_{LRT} = 2\left(\ln L(\hat{\beta}) - \ln L(\hat{\beta}_0)\right) \underset{H_0}{\rightarrow} \chi(1)$

Tests statistiques

- Modèle logistique $\rightarrow$ $\text{Logit}(P(M+|X_1, X_2, \dots, X_k)) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$

- On peut comparer des modèles emboîtés entre eux, par exemple

  Modèle A : $\text{Logit}(P(M+|X_1, X_2, \dots, X_k)) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$

  Modèle B : $\text{Logit}(P(M+|X_1)) = \alpha + \beta_1 X_1$

$$\begin{cases} H_0 : \beta_2 = \cdots = \beta_k = 0 \\ H_1 : \exists\, \beta_i \neq 0, i = 2, \dots, k \end{cases} \Leftrightarrow \begin{cases} H_0 : OR_2 = \cdots = OR_k = 1 \\ H_1 : \exists\, OR_i \neq 1, i = 2, \dots, k \end{cases}$$

  - Rejet de H_0 $\rightarrow$ on rejette le modèle B
  - Non rejet de H_0 $\rightarrow$ on peut raisonnablement considérer le modèle B

- Test du rapport de vraisemblance : $\chi_{LRT} = 2(\ln L_A - \ln L_B) \underset{H_0}{\rightarrow} \chi(\underbrace{k+1-2}) \equiv \chi(k-1)$

  Nombre de paramètres du modèle A – Nombre de paramètres du modèle B

VIII. Interaction entre variables

- Modèle sans interaction : $\text{Logit}(P(M+|X_1,X_2)) = \alpha + \beta_1 X_1 + \beta_2 X_2$ ➡ Vraisemblance L_1
- Modèle avec interaction : $\text{Logit}(P(M+|X_1,X_2)) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \gamma X_1 X_2$ ➡ Vraisemblance L_2

- Test de l'interaction avec le test de Wald ou du rapport de vraisemblance

$$\begin{cases} H_0 : \gamma = 0 \\ H_1 : \gamma \neq 0 \end{cases}$$

  - Test de Wald : $\chi_W = \frac{\hat{\gamma}}{\hat{s}_{\hat{\gamma}}} \underset{H_0}{\rightarrow} N(0,1)$
  - Test du rapport de vraisemblance : $\chi_{LRT} = 2(\ln L_2 - \ln L_1) \underset{H_0}{\rightarrow} \chi(1)$

- OR entre variable X_1 et la maladie (Exemple avec X_1 et X_2 binaires)
  - Modèle sans interaction : $OR = e^{\beta_1}$
  - Modèle avec interaction : $OR = e^{\beta_1}$ quand $X_2 = 0$

    $OR = e^{\beta_1 + \gamma}$ quand $X_2 = 1$ (Rq: IC d'une somme de paramètres)

VIII. Méthodes alternatives

- Régression logistique est un modèle très populaire
- Régression logistique fait partie des Modèles Linéaires Généralisés (modèle binomial)
- Extension possible dans le cas où la variable à expliquer est qualitative ordonnée

  ↳ Régression logistique polytomique ou ordinale

- Méthodes alternatives $\rightarrow$ méthode d'apprentissage supervisé
  - Arbre de décision (algorithme CART)
  - Forêt aléatoire
  - Support Vector Machine
  - Méthode des k plus proches voisins