Sommaire

Introduction

Un modèle multivarié permet d'exprimer une variable Y en fonction de plusieurs variables X_i

En épidémiologie,
- Y caractérise la maladie (ou sa distribution dans la population)
- les X_i caractérisent les facteurs de risque de la maladie ou des variables d'ajustement. Ils peuvent être qualitatifs ou quantitatifs.

• Les méthodes multivariées supposent une certaine **modélisation** de la réalité.

Par exemple, on peut modéliser (représenter) la relation entre Y et les X_i

- sous forme linéaire
- et/ou en supposant l'absence d'interaction entre les X_i

=> Les conclusions qu'on tire des analyses multivariées sont en partie conditionnées par le bien-fondé des hypothèses faites (par exemple linéarité).

Principaux modèles multivariés utilisés en épidémiologie

- **Régression linéaire multiple** (Y est quantitative)

$$\bar{Y}=E(Y|X_1,...,X_p)=\alpha+\beta_1X_1+...+\beta_pX_p=\alpha+\sum_{i=1}^{p}\beta_iX_i$$

$\bar{Y}$: moyenne de Y connaissant $X_1,\ldots,X_p$

- **Régression logistique** (Y = 0/1)

$$P(M^+|X_1,...,X_p)\frac{1}{1+\exp\left\{-\left(\alpha+\sum_{i=1}^{p}\beta_iX_i\right)\right\}}$$

$P(M^+|X_1,..., X_p)$: probabilité de maladie connaissant $X_1,\ldots,X_p$

Y avec plus de 2 classes : régression logistique multinomiale

- **Modèle de Cox** (Y = incidence instantanée λ)

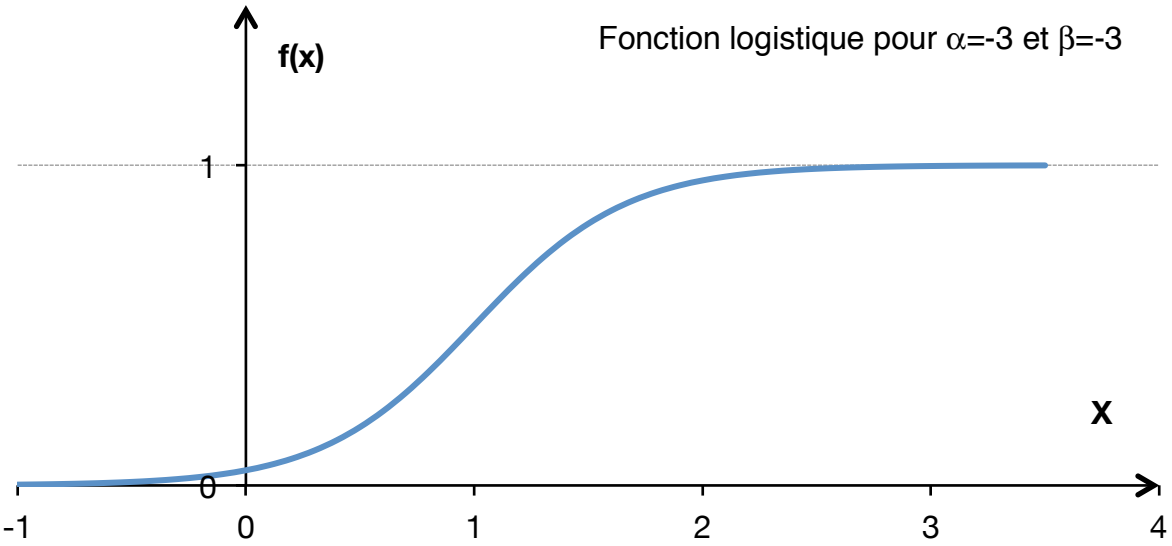
$$\lambda(t,X_1,...,X_p)=\lambda_0(t)\exp\left\{\sum_{i=1}^{p}\beta_iX_i\right\}$$

Définition du modèle logistique

1. Une seule variable X

$$P(M^{+}|X) = \frac{1}{1+e^{-(\alpha+\beta x)}}$$

- Fonction logistique : $f(x) = \frac{1}{1+e^{-(\alpha+\beta x)}}$



- L'association entre la maladie M et X est mesurée par β.

Si X est dichotomique (X=1 : exposé et X=0 : non exposé), β permet de retrouver l'odds ratio.

... β permet de retrouver l'odds ratio

$$P(M^+ \mid X) = \frac{1}{1+e^{-(\alpha+\beta x)}}$$

Par définition : $OR = \frac{P_1/1-P_1}{P_o/1-P_o}$ avec $P_o = P(M+|X=0)$, $P_1 = P(M+|X=1)$

$\boxed{X=1}$

$$P(M^+) = P_1 = \frac{1}{1+e^{-(\alpha+\beta)}} \qquad 1- P_1 = 1- \frac{1}{1+e^{-(\alpha+\beta)}} = \frac{e^{-(\alpha+\beta)}}{1+e^{-(\alpha+\beta)}}$$

$\boxed{X=0}$

$$P(M^+) = P_o = \frac{1}{1+e^{-\alpha}} \qquad 1- P_o = \frac{e^{-\alpha}}{1+e^{-\alpha}}$$

d'où : $$OR = \frac{P_1 / 1-P_1}{P_0 / 1-P_0} = \exp(\beta)$$

Autre écriture du modèle logistique

$$P(M^+|X) = \frac{1}{1+e^{-(\alpha+\beta x)}}$$

Définition : Logit P = $\text{Ln}\ \frac{P}{1-P}$ = Ln P - Ln(1-P)

P = P(M^+|X) —> $\boxed{\text{Logit}\, P = \alpha + \beta X}$

Si X est dichotomique, β permet de retrouver l'odds ratio :

$$\text{Ln OR} = \text{Ln}\left(\frac{P_1/1-P_1}{P_o/1-P_o}\right) = \text{Ln}(P_1/1\text{-}P_1) - \text{Ln}(P_o/1\text{-}P_o)$$

$$= \text{Logit}\ P_1 - \text{Logit}\ P_o$$

$$= (\alpha+\beta) - \alpha$$

$$= \beta$$

$$\boxed{\text{Ln OR} = \text{Logit}\ P_1 - \text{Logit}\ P_o}$$

Définition du modèle logistique

2. Deux variables

Logit P = $\alpha + \beta E + \beta' X$

exp(β) = OR lié à l'exposition E ***ajusté*** sur X

OR ajusté car Logit P varie de β quand E passe de 0 à 1 (non exposé à exposé) à condition que X reste fixe.

> Dans la réalité, X peut être différent chez les exposés et les non exposés (et donc ne pas être fixe). L'interprétation de β suppose qu'on "force" (par le calcul) X à être identique chez les exposés et les non exposés. C'est par définition, ce qu'on appelle ajuster sur X.

3. Plusieurs variables X_i

$$P(M^+|X_1,\ldots, X_k) = \frac{1}{1+\exp-\left(\alpha + \sum_{i=1}^{k}\beta_i X_i\right)} \quad \text{ou} \quad \text{Logit P} = \alpha + \sum_{i=1}^{k}\beta_i X_i$$

Si X_i prend les valeurs 0 et 1,

exp(β_i) = OR lié à l'exposition X_i ***ajusté*** sur les autres variables

Utilisation du modèle logistique dans les enquêtes cas-témoins

◊ On peut utiliser le modèle logistique dans les enquêtes cas-témoins.

Cela signifie que les coefficients β obtenus donnent toujours les odds-ratios par la relation $OR = e^{\beta}$.

On peut noter que cela est cohérent avec le fait que OR peut être estimé dans tous les types d'enquête.

◊ Démonstration

Soit f_0 et f_1 les fractions de sondage des témoins (Y = 0) et des malades (Y = 1).

On a : $P(Y=1|X) = f_1 \dfrac{1}{1+\exp\left\{-\left(\alpha+\sum\beta_i X_i\right)\right\}}$

$P(Y=0|X) = 1-P(Y=1|X) = f_0 \dfrac{\exp\left\{-\left(\alpha+\sum\beta_i X_i\right)\right\}}{1+\exp\left\{-\left(\alpha+\sum\beta_i X_i\right)\right\}}$

D'où : $\dfrac{P(Y=1|X)}{1-P(Y=1|X)} = \dfrac{f_1}{f_0}\exp\left(\alpha+\sum\beta_i X_i\right)$

et donc :

$$\text{Logit } P = \text{Ln}\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right) = \text{Ln}\left(\frac{f_1}{f_0}\right) + \left(\alpha+\sum\beta_i X_i\right) = \alpha' + \sum\beta_i X_i$$

Les coefficients β_i (et donc les OR_i) ne changent pas, mais la constante α n'est pas interprétable.

Estimation des paramètres

Modèle : $P(M^+ \mid X) = \frac{1}{1+e^{-(\alpha+\beta x)}}$

Observations : échantillon de N sujets -> (x_i, y_i) i=1,N

α et β sont estimés par la méthode du maximum de vraisemblance : les estimations obtenues sont notées $\widehat{\alpha}$ et $\widehat{\beta}$.

Vraisemblance d'un échantillon

Définition : la vraisemblance d'un échantillon est la probabilité d'observer l'échantillon. Elle est notée V (en anglais L pour likelihood).

Exemple : variable dichotomique (M^+/ M^-)

P = pourcentage de malades dans la population

• Echantillon de 1 sujet	• Echantillon de 2 sujets
- sujet malade : V = P	- M^+M^+ : $V = P^2$
- sujet non malade : V = 1-P	- M^-M^- : $V = (1-P)^2$
	- M^+M^-: $V = P(1-P) + (1-P)P = 2P(1-P)$

• De façon générale, soit un échantillon de n sujets avec k malades

Notons p_i la probabilité d'observer le sujet i

La vraisemblance de l'échantillon est $V = c \prod_{i=1}^{n} p_i$

où c = nombre d'échantillons possibles avec k malades

$c = C_n^k = \frac{n!}{k!\,(n-k)!}$ où $n! = n(n-1)\ldots 1 \quad 0! = 1$

Comme on a :

$p_i = P$ si le sujet i est malade

$p_i = (1-P) = Q$ si le sujet i est non malade

On obtient : $\boxed{V = C_n^k\, P^k\, (1-P)^{n-k}}$

• La valeur de la vraisemblance dépend du nombre de malades dans l'échantillon, mais aussi de P.
Considérons un échantillon de n=20 sujets où on observe k=5 malades

- P=10%

$$V = C_{20}^{5}\, 0{,}10^{5}\, 0{,}90^{15} = \frac{20!}{5!15!} 0{,}10^{5}\, 0{,}90^{15} = 0{,}03$$

- P = 25%, $V' = C_{20}^{5}\, 0{,}25^{5}\, 0{,}75^{15} = 0{,}20$

→ *fonction de vraisemblance* : $V = C_{20}^{5} P^{5} (1\text{-}P)^{15}$

• Lorsqu'on ne connaît pas P, mais qu'on observe n et k sur un échantillon, on se sert de la vraisemblance pour estimer P :

Estimation du maximum de vraisemblance = valeur p_O qui maximise V (connaissant n et k)

Ici, l'estimation du maximum de vraisemblance est $\frac{k}{n}$.

• Propriétés :

- Les estimateurs du maximum de vraisemblance ont une distribution asymptotiquement normale

- La méthode du maximum de vraisemblance permet de calculer la variance des paramètres estimés

Estimation des paramètres du modèle logistique

Dans un échantillon de n sujet, les observations sont les couples (x_i , y_i) où y_i = 1 ou 0 selon qu'il s'agit d'un malade ou d'un non malade et x_i est la valeur de x pour le sujet i.

La vraisemblance s'écrit toujours : $V = c \prod_{i=1}^{n} p_i$

avec $c = \frac{n!}{k!\,(n-k)!}$ et p_i = probabilité d'observer le sujet i

Le modèle logistique permet d'exprimer la probabilité P d'être malade en fonction de la valeur de X : $P = P(M^+|X) = \frac{1}{1+e^{-(\alpha+\beta x)}}$

Pour le sujet i, on a donc :

$p_i = P(M^+|X=x_i) = \frac{1}{1+e^{-(\alpha+\beta x_i)}}$ si le sujet est malade

$p_i = P(M^-|x_i) = 1 - P(M^+|x_i) = \frac{e^{-(\alpha+\beta x_i)}}{1+e^{-(\alpha+\beta x_i)}}$ s'il est non malade

On obtient finalement : $\boxed{V = c \frac{\prod_{\text{non malades}} e^{-(\alpha+\beta x_i)}}{\prod_{\text{tous}} \left(1+e^{-(\alpha+\beta x_i)}\right)}}$

Vraisemblance de l'échantillon : $V = c \dfrac{\prod_{\text{non malades}} e^{-(\alpha+\beta x_i)}}{\prod_{\text{tous}} \left(1+e^{-(\alpha+\beta x_i)}\right)}$

Les estimations du maximum de vraisemblance de α et β sont les valeurs $\hat{\alpha}$ et $\hat{\beta}$ qui rendent V maximum.

Il n'y a de solutions explicites pour $\hat{\alpha}$ et $\hat{\beta}$ que dans le cas d'un seul coefficient β et d'une variable X dichotomique. Sinon (quand il y a plusieurs variables X), il faut procéder par itérations, et les valeurs obtenues sont des approximations numériques

NB : en fait les logiciels donnent LnV et non V, et ne font pas figurer "c" (ce qui ne change rien pour la recherche du maximum)

Exemple[1]

Grossesse extra-utérine et antécédent de salpingite

- La variable maladie est notée "ct" (0 = accouchement; 1= GEU)
- La variable "exposition" est notée "salp" (salpingite clinique prouvée ou suspectée : 0 = non ; 1 = oui)

Le modèle est donc : $\boxed{\text{Logit}P = \alpha + \beta \text{salp}}$

ou : $P\left(M^{+} \middle| \text{salp}\right) = \dfrac{1}{1+e^{-(\alpha+\beta\ \text{salp})}}$

```
. tab ct salp

           | salp clinique prouv
    0:acc  |       ou susp
    1:GEU  |         0          1 |     Total
-----------+----------------------+----------
         0 |     1,122         26 |     1,148
         1 |       466         91 |       557
-----------+----------------------+----------
     Total |     1,588        117 |     1,705
```

[1] Sauf mention contraire, les exemples s'appuient sur une enquête cas-témoins sur les facteurs de risque de la grossesse extra-utérine (GEU). Les données sont disponibles à l'adresse https://sites.google.com/site/master2eq/ sous les formats Stata et Excel (fichier geu)

$$\text{Logit}\,P = \alpha + \beta \text{salp}$$

```
. tab ct salp

           |  salp clinique prouv
    0:acc  |       ou susp
    1:GEU  |         0          1 |     Total
-----------+----------------------+----------
         0 |     1,122         26 |     1,148
         1 |       466         91 |       557
-----------+----------------------+----------
     Total |     1,588        117 |     1,705
```

Contribution à la vraisemblance :

91 sujets $E^+ M^+$: $p_1 = \dfrac{1}{1+e^{-(\alpha+\beta)}}$

466 sujets $E^- M^+$: $p_2 = \dfrac{1}{1+e^{-\alpha}}$

26 sujets $E^+ M^-$: $p_3 = \dfrac{e^{-(\alpha+\beta)}}{1+e^{-(\alpha+\beta)}}$

1122 sujets $E^- M^-$: $p_4 = \dfrac{e^{-\alpha}}{1+e^{-\alpha}}$

$$V = C\, p_1^{91}\, p_2^{466}\, p_3^{26}\, p_4^{1122} = \frac{\left[e^{-(\alpha+\beta)}\right]^{26}\left[e^{-\alpha}\right]^{1122}}{\left[1+e^{-(\alpha+\beta)}\right]^{117}\left[1+e^{-\alpha}\right]^{1588}}$$

Exemple (suite)

Grossesse extra-utérine et antécédent de salpingite

$$\text{Logit}\, P = \alpha + \beta \text{salp}$$

```
. logit ct salp

Iteration 0:   log likelihood = -1077.2309
Iteration 1:   log likelihood = -1023.1123
Iteration 2:   log likelihood = -1023.0534
Iteration 3:   log likelihood = -1023.0534

Logistic regression                               Number of obs   =       1705
                                                  LR chi2(1)      =     108.36
                                                  Prob > chi2     =     0.0000
Log likelihood = -1023.0534                       Pseudo R2       =     0.0503

------------------------------------------------------------------------------
          ct |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        salp |   2.131445    .229102     9.30   0.000     1.682413    2.580476
       _cons |  -.8786825   .0551107   -15.94   0.000    -.9866974   -.7706675
------------------------------------------------------------------------------
```

$\widehat{\alpha}$ = -0,879 $\quad s_{\widehat{\alpha}}$ = 0,0551

$\widehat{\beta}$ = 2,131 $\quad s_{\widehat{\beta}}$ = 0,229

$$\widehat{OR} = e^{\widehat{\beta}} = 8{,}42$$

Intervalle de confiance

I.C. de β : $\widehat{\beta} \pm z_{\alpha/2}\, s_{\widehat{\beta}} = 2{,}131 \pm 1{,}96 \times 0{,}229 = [1{,}682\ ;\ 2{,}580]$

D'où : I.C. de OR : $[e^{1{,}682}\ ;\ e^{2{,}580}] = [5{,}38\ ;\ 13{,}20]$

• On peut obtenir directement l'odds ratio et son intervalle de confiance avec la commande "logistic" :

```
. logistic ct salp

Logistic regression                               Number of obs   =       1705
                                                  LR chi2(1)      =     108.36
                                                  Prob > chi2     =     0.0000
Log likelihood = -1023.0534                       Pseudo R2       =     0.0503

------------------------------------------------------------------------------
          ct | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        salp |   8.427032   1.93065     9.30   0.000     5.378519    13.20343
       _cons |   .4153298   .0228891   -15.94   0.000     .3728059    .4627041
------------------------------------------------------------------------------
```

• On retrouve les mêmes résultats avec le tableau initial :

$$OR = \frac{ad}{bc} = \frac{91 \times 1122}{26 \times 466} = 8{,}43$$

```
. cc ct salp,w

                |  salp clinique prouv ou |
                |  susp                   |                 Proportion
                |    Exposed    Unexposed |      Total        Exposed
----------------+-------------------------+-------------------------
          Cases |          91          466 |          557      0.1634
       Controls |          26         1122 |         1148      0.0226
----------------+-------------------------+-------------------------
          Total |         117         1588 |         1705      0.0686
                |                         |
                |       Point estimate    |    [95% Conf. Interval]
                |-------------------------+-------------------------
     Odds ratio |           8.427039      |     5.378522    13.20344 (Woolf)
 Attr. frac. ex.|           .8813344      |     .8140753    .9242621 (Woolf)
 Attr. frac. pop|           .1439882      |
                +-------------------------------------------------
                                  chi2(1) =   116.21  Pr>chi2 = 0.0000
```

Tests des paramètres

Hypothèse testée : $H_0 : OR = 1 \Leftrightarrow \beta=0 \Leftrightarrow P_0=P_1$

$H_1 : OR \neq 1$

- Test de χ^2 “habituel”

	E^+	E^-	
$M+$	a	b	m_1
$M-$	c	d	m_0
	n_1	n_0	n

$$\chi^2_c = \frac{\left(|ad-bc| - \frac{n}{2}\right)^2 n}{n_1 n_2 m_1 m_2}$$

- Test de Wald

si H_0 est vraie, $z = \frac{\hat{\beta} - 0}{s_{\hat{\beta}}}$

=> on calcule $z_0 = \frac{\hat{\beta}}{s_{\hat{\beta}}}$ et on compare au seuil de la loi normale centrée réduite $N(0,1)$

Souvent exprimé sous forme de χ^2 :

si H_0 est vrai, $\chi^2_0 = \frac{\hat{\beta}^2}{s^2_{\hat{\beta}}} = \frac{\hat{\beta}^2}{var(\hat{\beta})}$ suit une loi de χ^2 à 1 ddl

• Test du rapport des vraisemblances

modèle 1 : Logit P = α X n'est pas associé à P
vraisemblance : V_1

modèle 2 : Logit P = $\alpha' + \beta X$ vraisemblance : V2

$V_1 \leq V_2$ car le modèle 2 est plus complet

β est différent de 0 si V_1 est inférieure à V_2

Test : $2 \ln \frac{V_2}{V_1} = 2 \ln(V_2) - 2 \ln(V_1)$ suit une loi de χ^2 à 1ddl

Exemple

Association GEU - Antécédent de salpingite

- Test du χ^2 "habituel"

```
. tab ct salp,chi

           |  salp clinique prouv
    0:acc  |        ou susp
    1:GEU  |         0          1 |     Total
-----------+----------------------+----------
         0 |     1,122         26 |     1,148
         1 |       466         91 |       557
-----------+----------------------+----------
     Total |     1,588        117 |     1,705

      Pearson chi2(1) = 116.2093   Pr = 0.000
```

- Test de Wald

```
. logistic ct salp

Logistic regression                               Number of obs   =       1705
                                                  LR chi2(1)      =     108.36
                                                  Prob > chi2     =     0.0000
Log likelihood = -1023.0534                       Pseudo R2       =     0.0503

------------------------------------------------------------------------------
          ct | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        salp |   8.427032    1.93065     9.30   0.000     5.378519    13.20343
       _cons |   .4153298   .0228891   -15.94   0.000     .3728059    .4627041
------------------------------------------------------------------------------
```

- Test du rapport des vraisemblances

```
. logistic ct salp

Logistic regression                               Number of obs   =       1705
                                                  LR chi2(1)      =     108.36
                                                  Prob > chi2     =     0.0000
Log likelihood = -1023.0534                       Pseudo R2       =     0.0503

------------------------------------------------------------------------------
          ct | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        salp |   8.427032    1.93065     9.30   0.000     5.378519    13.20343
       _cons |   .4153298   .0228891   -15.94   0.000     .3728059    .4627041
------------------------------------------------------------------------------

. est store a

. logistic ct if salp!=.

Logistic regression                               Number of obs   =       1705
                                                  LR chi2(0)      =      -0.00
                                                  Prob > chi2     =          .
Log likelihood = -1077.2309                       Pseudo R2       =    -0.0000

------------------------------------------------------------------------------
          ct | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   .4851916    .025054   -14.01   0.000     .4384899    .5368674
------------------------------------------------------------------------------

. est store b


. lrtest a b

Likelihood-ratio test                             LR chi2(1)  =     108.36
(Assumption: b nested in a)                       Prob > chi2 =     0.0000
```

χ^2 "habituel"	Wald	Rapport des vraisemblances
116,2	86,49 (9,30^2)	108,36

• Le test du rapport de vraisemblance ne s'applique qu'à 2 modèles emboîtés

M_2 est emboîté dans M_1 (ou est un sous-modèle de M_1) si, en partant de M_1 et en imposant certaines relations entre ses coefficients, on retrouve le modèle M_2. C'est le cas, en particulier, quand on passe de M_1 à M_2 en enlevant une variable.

exemple : les 2 modèles suivants ne sont pas comparables (pas emboîtés)

modèle 1 : Logit $P = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

modèle 2 : Logit $P = \alpha' + \beta'_1 X_1 + \beta'_2 X_2 + \beta'_4 X_4$

• De plus, les deux vraisemblances doivent être calculées sur les mêmes sujets

```
. logistic ct salp

Logistic regression                               Number of obs   =       1705
                                                  LR chi2(1)      =     108.36
                                                  Prob > chi2     =     0.0000
Log likelihood = -1023.0534                       Pseudo R2       =     0.0503

------------------------------------------------------------------------------
          ct | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        salp |   8.427032    1.93065     9.30   0.000     5.378519    13.20343
       _cons |   .4153298   .0228891   -15.94   0.000     .3728059    .4627041
------------------------------------------------------------------------------

. est store a

. logistic ct

Logistic regression                               Number of obs   =       1725
                                                  LR chi2(0)      =       0.00
                                                  Prob > chi2     =          .
Log likelihood = -1097.2925                       Pseudo R2       =     0.0000

------------------------------------------------------------------------------
          ct | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   .4986968   .0254822   -13.62   0.000      .451172    .5512276
------------------------------------------------------------------------------

. est store b
.
. lrtest a b

observations differ: 1705 vs. 1725
```

• Les tests de Wald et du rapport de vraisemblance s'étendent au test de plusieurs coefficients à la fois

Exemple avec 2 paramètres

modèle 1 : Logit P = $\alpha + \beta_1 X_1$
modèle 2 : Logit P = $\alpha' + \beta'_1 X_1 + \beta'_2 X_2 + \beta'_3 X_3$

H_o : $\beta'_2 = 0$ et $\beta'_3 = 0$ H_1 : $\beta'_2 \neq 0$ ***ou*** $\beta'_3 \neq 0$

• **Test du rapport des vraisemblances**

Test : $2 \text{ Ln}\frac{V_2}{V_1} = 2 \text{ Ln } (V_2) - 2 \text{ Ln } (V_1)$ suit une loi de χ^2 à **2** ddl

• **Test de Wald**

$\chi_0^2 = \frac{\hat{\beta}^2}{\text{var}(\hat{\beta})}$ à 1 ddl pour un paramètre s'étend en un χ^2 à 2 ddl

En effet : $z_0 = \frac{\hat{\beta}}{s_{\hat{\beta}}} <=> \chi_0^2 = \frac{\hat{\beta}^2}{\text{var}(\hat{\beta})} = \left(\hat{\beta}\right)' \left[\text{var}(\hat{\beta})\right]^{-1} \left(\hat{\beta}\right)$

extension : $\theta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$ $\text{var}(\theta) = \begin{pmatrix} \text{var}(\beta_1) & \text{cov}(\beta_1,\beta_2) \\ \text{cov}(\beta_1,\beta_2) & \text{var}(\beta_2) \end{pmatrix}$

Test : $\chi_0^2 = \left(\hat{\theta}\right)' \left[\text{var}(\hat{\theta})\right]^{-1} \left(\hat{\theta}\right)$ à 2 ddl

Exemple

univf : niveau d'études universitaires; fprof : activité prof. femme

```
. logistic ct salp univf fprof

Logistic regression                               Number of obs   =       1230
                                                  LR chi2(3)      =      71.66
                                                  Prob > chi2     =     0.0000
Log likelihood = -730.93713                       Pseudo R2       =     0.0467

------------------------------------------------------------------------------
          ct | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        salp |   7.392144   1.944662     7.60   0.000     4.414114    12.37933
       univf |   .8476072   .1323978    -1.06   0.290      .624073    1.151208
       fprof |   1.369012    .198271     2.17   0.030     1.030694    1.818381
       _cons |   .3314003   .0405161    -9.03   0.000     .2607878    .4211322
------------------------------------------------------------------------------
```

Le modèle logistique s'écrit : $\text{Logit}\, P = \alpha + \beta_1 \text{salp} + \beta_2 \text{univf} + \beta_3 \text{fprof}$

L'hypothèse testée est :

$H_o : \beta_2 = 0$ et $\beta_3 = 0$

$H_1 : \beta_2 \neq 0$ ou $\beta_3 \neq 0$ nb : dans le cas présent, le test simultané de β_2 et β_3 a un intérêt limité. Il n'est donné qu'à titre d'exemple. Voir plus loin des situations où c'est plus utile (variables indicatrices)

```
. logistic ct salp univf fprof

Logistic regression                             Number of obs   =       1230
                                                LR chi2(3)      =      71.66
                                                Prob > chi2     =     0.0000
Log likelihood = -730.93713                     Pseudo R2       =     0.0467

------------------------------------------------------------------------------
          ct | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        salp |   7.392144   1.944662     7.60   0.000     4.414114    12.37933
       univf |   .8476072   .1323978    -1.06   0.290      .624073    1.151208
       fprof |   1.369012    .198271     2.17   0.030     1.030694    1.818381
       _cons |   .3314003   .0405161    -9.03   0.000     .2607878    .4211322
------------------------------------------------------------------------------
```

$H_0 : \beta_2 = 0$ et $\beta_3 = 0$
$H_1 : \beta_2 \neq 0$ ou $\beta_3 \neq 0$

• Test de Wald

```
. testparm univf fprof

 ( 1)  [ct]univf = 0
 ( 2)  [ct]fprof = 0

           chi2(  2) =    5.06
         Prob > chi2 =    0.0797
```

• Test du rapport des vraisemblances

```
. est store a

. logistic ct salp if univf!=. & fprof!=.

Logistic regression                             Number of obs   =       1230
                                                LR chi2(1)      =      66.53
                                                Prob > chi2     =     0.0000
Log likelihood = -733.50452                     Pseudo R2       =     0.0434

------------------------------------------------------------------------------
          ct | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        salp |   7.151568   1.872462     7.51   0.000       4.2809    11.94724
       _cons |   .3995128   .0260965   -14.05   0.000     .3515034    .4540795
------------------------------------------------------------------------------

. est store b

.
. lrtest a b

Likelihood-ratio test                           LR chi2(2)  =      5.13
(Assumption: b nested in a)                     Prob > chi2 =    0.0767
```

Exercice 1

Cet exercice, ainsi que les suivants (exercice 2 et 3), font l'objet de la 1ère séance de TP. Ils reposent sur le fichier de données geu.dta (format Stata) ou geu.xls (format Excel) disponible sur le site web. Ils seront réalisés et corrigés en TP avec le logiciel Stata. Le format Excel vous permet cependant d'avoir accès aux données sans le logiciel Stata.

1. Calculez l'odds ratio de GEU associé à un antécédent d'accouchement (variable aacc) et son intervalle de confiance. Est-il significativement différent de 1 ?

2. a) Calculez l'odds ratio de GEU associé à un antécédent d'accouchement et son intervalle de confiance après ajustement sur l'âge (prendre la variable age30 (0 : < 30 ans 1 : ≥ 30 ans)).
b) Testez si cet odds ratio est différent de 1 par les différentes méthodes que vous connaissez.
c) L'âge est-il facteur de confusion pour la relation entre le risque de GEU et l'antécédent d'accouchement ?

3. Les antécédents obstétricaux (accouchement, fcs, ivg et geu) sont-ils globalement liés au risque de GEU ?

(les variables correspondantes dans le fichier de données sont aacc, afcs, aivg, et ageu)

Interprétation des coefficients et codage des variables

De façon générale, si Logit P = $\alpha + \beta X$, l'odds ratio entre les catégories $X = x_1$ et $X = x_o$ est donné par :

Ln OR = Logit P_1 - Logit P_o = $(\alpha + \beta x_1) - (\alpha + \beta x_o) = \beta(x_1 - x_o)$

d'où : $OR = e^{\beta(x_1 - x_o)}$

Exemple : X = age

```
. logit ct agea

Iteration 0:   log likelihood = -1095.6718
Iteration 1:   log likelihood = -1068.2162
Iteration 2:   log likelihood = -1068.1198
Iteration 3:   log likelihood = -1068.1198

Logistic regression                             Number of obs   =       1721
                                                LR chi2(1)      =      55.10
                                                Prob > chi2     =     0.0000
Log likelihood = -1068.1198                     Pseudo R2       =     0.0251

------------------------------------------------------------------------------
          ct |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        agea |   .0772543   .0105806     7.30   0.000     .0565166    .0979919
       _cons |  -2.965807   .3185008    -9.31   0.000    -3.590057   -2.341556
------------------------------------------------------------------------------
```

Logit P = -3,00 + 0,0772 age

OR entre $X_o = 29$ et $X_1 = 30$: OR = exp(0,0772) = 1,08

OR entre $X_o = 25$ et $X_1 = 40$: OR = exp(15×0,0772) = $1{,}08^{15}$
= 3,18

1. X qualitative à 2 classes (dichotomique)

X=1 : exposé
X=0 : non exposé

$e^{\beta} = OR$

Autre codage de X

X= +1 : exposé (x_1)
X= -1 : non exposé (x_o)

$OR = e^{\beta(x_1 - x_0)} = e^{2\beta}$

Coefficient $\beta \leftrightarrow$ odds ratio

```
. gen z=2*salp-1
(20 missing values generated)

. tab salp z,miss
     salp |
  clinique |
  prouv ou |                z
     susp |        -1         1         . |     Total
----------+---------------------------------+----------
        0 |     1,588         0         0 |     1,588
        1 |         0       117         0 |       117
        . |         0         0        20 |        20
----------+---------------------------------+----------
    Total |     1,588       117        20 |     1,725

. logistic ct salp

Logistic regression                               Number of obs   =       1705
                                                  LR chi2(1)      =     108.36
                                                  Prob > chi2     =     0.0000
Log likelihood = -1023.0534                       Pseudo R2       =     0.0503

------------------------------------------------------------------------------
          ct | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        salp |   8.427032    1.93065     9.30   0.000     5.378519    13.20343
       _cons |   .4153298   .0228891   -15.94   0.000     .3728059    .4627041
------------------------------------------------------------------------------

. logistic ct z

Logistic regression                               Number of obs   =       1705
                                                  LR chi2(1)      =     108.36
                                                  Prob > chi2     =     0.0000
Log likelihood = -1023.0534                       Pseudo R2       =     0.0503

------------------------------------------------------------------------------
          ct | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           z |   2.902935   .3325341     9.30   0.000     2.319163    3.633652
       _cons |   1.205675   .1381113     1.63   0.103     .9632176    1.509164
------------------------------------------------------------------------------
```

$8{,}427 = 2{,}903^2$

Interprétation des coefficients et codage des variables

2. X qualitative nominale à plus de 2 classes

Y = GEU	X = induction de la grossesse	X=0 -> Non induite X=1 -> Hcg X=2 -> Clomid X=3 -> Autre
Logit P = $\alpha+\beta X$		

Entre les catégories X=0 et X=1 : $\boxed{OR_{0,1} = e^{\beta}}$

$Ln(OR_{o,1})$ = Logit P_1 - Logit P_o = $(\alpha+\beta) - \alpha = \beta$

Entre les catégories X=0 et X=2 : $\boxed{OR_{0,2} = e^{2\beta} = \left(OR_{0,1}\right)^2}$

X=2 -> E^+, X=0 -> E^-

D'où : $Ln(OR_{0,2})$ = Logit P_1 - Logit P_o = $(\alpha+2\beta) - \alpha = 2\beta$

On obtient donc : OR associé à Clomid est le carré de celui associé à Hcg

Y = GEU	X = induction de la grossesse	X=0 -> Non induite X=1 -> Hcg X=2 -> Clomid X=3 -> Autre
Logit P = $\alpha+\beta X$		

En prenant la catégorie "Non induite" comme référence :

- $OR_{"Hcg"} = e^{\beta}$
- $OR_{"Clomid"} = e^{2\beta} = (OR_{"Hcg"})^2$
- $OR_{"Autre"} = e^{3\beta} = (OR_{"Hcg"})^3$

Le modèle contient ces relations entre les odds-ratios dans son écriture (c'est-à-dire indépendamment des données observées).

Si le codage avait été :

X'=0 -> Non induite
X'=1 -> Clomid
X'=2 -> Hcg
X'=3 -> Autre

Logit P = $\alpha'+\beta'X'$

On aurait obtenu (toujours avec la catégorie "Non induite" comme référence) :

- $OR_{"Clomid"} = e^{\beta'}$
- $OR_{"Hcg"} = e^{2\beta'} = (OR_{"Clomid"})^2$
- $OR_{"Autre"} = e^{3\beta'} = (OR_{"Clomid"})^3$

Une variable X qualitative nominale à plus de 2 classes ne doit jamais être incluse dans un modèle logistique sous sa forme initiale.

Exemple

```
. tab ct gind

    0:acc |       0:non ind 1:hcg 2:clomid 3:autr
    1:GEU |         0          1          2          3 |     Total
----------+--------------------------------------------+----------
        0 |     1,099         13         31          4 |     1,147
        1 |       528         10         29          5 |       572
----------+--------------------------------------------+----------
    Total |     1,627         23         60          9 |     1,719


. logit ct gind

Logistic regression                               Number of obs   =       1719
                                                  LR chi2(1)      =       9.09
                                                  Prob > chi2     =     0.0026
Log likelihood = -1088.9259                       Pseudo R2       =     0.0042

------------------------------------------------------------------------------
          ct |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        gind |   .3382189   .1113441     3.04   0.002     .1199885    .5564494
       _cons |  -.7318449   .0527975   -13.86   0.000    -.8353261   -.6283637
------------------------------------------------------------------------------
```

Logit P= $\alpha + \beta X = -0,73 + 0,34\ X$

$OR_{Hcg}=e^{0,34}=1,40$ $OR_{Clomid}=e^{0,68}=1,97$ $OR_{Autre}=e^{0,102}=2,77$

```
. tab ct gind2

    0:acc |       0:non ind 1:clomid 2:hcg 3:autr
    1:GEU |         0          1          2          3 |     Total
----------+--------------------------------------------+----------
        0 |     1,099         31         13          4 |     1,147
        1 |       528         29         10          5 |       572
----------+--------------------------------------------+----------
    Total |     1,627         60         23          9 |     1,719


. logit ct gind2

Logistic regression                               Number of obs   =       1719
                                                  LR chi2(1)      =       7.45
                                                  Prob > chi2     =     0.0064
Log likelihood = -1089.7504                       Pseudo R2       =     0.0034

------------------------------------------------------------------------------
          ct |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       gind2 |   .3714249   .1357612     2.74   0.006     .1053378    .6375119
       _cons |  -.7265713   .0526058   -13.81   0.000    -.8296766   -.6234659
------------------------------------------------------------------------------
```

Logit P = $\alpha' + \beta' X = -0,73 + 0,37\ X'$

$OR_{Clomid}=e^{0,37}=1,45$ $OR_{Hcg}=e^{0,74}= 2,10$ $OR_{Autre}=e^{1,11}=3,03$

Décomposition d'une variable qualitative à plus de 2 classes

		X_1	X_2	X_3
	0	0	0	0
	1	1	0	0
X =	2	0	1	0
	3	0	0	1

$$\text{Logit } P = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

X_1, X_2 et X_3 sont appelées des variables indicatrices.

Lorsque X a k classes, (k-1) variables indicatrices suffisent.

Si on en prenait k, il n'y aurait plus unicité des coeffcients du modèle :

		X_o	X_1	X_2	X_3
	0	1	0	0	0
	1	0	1	0	0
X =	2	0	0	1	0
	3	0	0	0	1

$$\text{Logit } P = \alpha + \beta_o X_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

On peut constater que les modèles

Logit $P = \alpha + \beta_o X_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

et

Logit $P = (\alpha+a) + (\beta_o-a)X_o + (\beta_1-a)X_1 + (\beta_2-a)X_2 + (\beta_3-a)X_3$

donnent les mêmes valeurs de Logit P quelle que soit la valeur de X.

		X_1	X_2	X_3
	0	0	0	0
	1	1	0	0
X =	2	0	1	0
	3	0	0	1

Logit $P = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

X_1, X_2 et X_3 sont appelées des variables indicatrices.

• $\beta_1 = \text{Ln}(OR_{o,1})$

$\text{Ln}(OR_{o,1}) = \text{Logit } P_1 - \text{Logit } P_o$

P_1 correspond à $X_1 = 1, X_2 = 0, X_3 = 0$
P_o correspond à $X_1 = 0, X_2 = 0, X_3 = 0$

D'où : $\text{Ln}(OR_{o,1}) = (\alpha+\beta_1) - \alpha = \beta_1$

De même :

• $\beta_2 = \text{Ln}(OR_{o,2})$

• $\beta_3 = \text{Ln}(OR_{o,3})$

• Ln(OR) entre les catégories 1 et 2 = $\beta_2 - \beta_1$

Dans Stata, des variables indicatrices sont construites en ajoutant i. devant le nom de la variable.

- voir "factor variables" dans la documentation ou le help de Stata

- l'ancienne méthode nécessitait d'ajouter "xi" devant la commande (voir la syntaxe dans le help). Elle est toujours utile pour les commandes non compatibles avec "factor variables".

Exemple
(suite)

```
. logit ct i.gind

Logistic regression                             Number of obs   =       1719
                                                LR chi2(3)      =       9.20
                                                Prob > chi2     =     0.0267
Log likelihood = -1088.8724                     Pseudo R2       =     0.0042

------------------------------------------------------------------------------
          ct |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        gind |
           1 |   .4706954   .4239421     1.11   0.267    -.3602159    1.301607
           2 |   .6663683   .2637133     2.53   0.012     .1494998    1.183237
           3 |   .9562032    .672907     1.42   0.155    -.3626703    2.275077
             |
       _cons |  -.7330597   .0529515   -13.84   0.000    -.8368426   -.6292767
------------------------------------------------------------------------------
```

4 variables indicatrices sont créées correspondant aux 4 classes 0, 1, 2 et 3 de gind. Elles n'apparaissent pas dans la liste des variables, mais sont accessibles avec les noms 0.gind, 1.gind, 2.gind et 3.gind.

Par défaut, la catégorie gind = 0 (valeur la plus petite de gind) est prise comme référence. Cela se traduit par l'omission de la variable 0.gind dans le modèle logistique.

On peut imposer une autre catégorie de référence (ou de base) en remplaçant i. par ib(#) où # est la valeur de la catégorie de référence voulue ou par ib(##) où # est le numéro d'ordre de la catégorie de référence voulue.
Par exemple, pour prendre comme référence la catégorie codée 2, il faut écrire ib(2).gind ou ib(#3).gind.

```
. logit ct ib(2).gind

Logistic regression                             Number of obs   =       1719
                                                LR chi2(3)      =       9.20
                                                Prob > chi2     =     0.0267
Log likelihood = -1088.8724                     Pseudo R2       =     0.0042

------------------------------------------------------------------------------
          ct |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       gind |
          0 |  -.6663683   .2637133    -2.53   0.012    -1.183237   -.1494998
          1 |  -.1956729   .4936232    -0.40   0.692    -1.163157    .7718109
          3 |   .2898349   .7188469     0.40   0.687    -1.119079    1.698749
            |
       _cons |  -.0666914   .2583425    -0.26   0.796    -.5730333    .4396505
------------------------------------------------------------------------------

. logit ct ib(#3).gind

Logistic regression                             Number of obs   =       1719
                                                LR chi2(3)      =       9.20
                                                Prob > chi2     =     0.0267
Log likelihood = -1088.8724                     Pseudo R2       =     0.0042

------------------------------------------------------------------------------
          ct |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       gind |
          0 |  -.6663683   .2637133    -2.53   0.012    -1.183237   -.1494998
          1 |  -.1956729   .4936232    -0.40   0.692    -1.163157    .7718109
          3 |   .2898349   .7188469     0.40   0.687    -1.119079    1.698749
            |
       _cons |  -.0666914   .2583425    -0.26   0.796    -.5730333    .4396505
------------------------------------------------------------------------------
```

Exemple
(suite)

```
. tab ct gind

    0:acc |       0:non ind 1:hcg 2:clomid 3:autr
    1:GEU |         0          1          2          3 |     Total
----------+--------------------------------------------+----------
        0 |     1,099         13         31          4 |     1,147
        1 |       528         10         29          5 |       572
----------+--------------------------------------------+----------
    Total |     1,627         23         60          9 |     1,719
```

$$OR_{Hcg} = \frac{10 \times 1099}{13 \times 528} = 1{,}60$$

```
. logit ct i.gind

Logistic regression                               Number of obs   =       1719
                                                  LR chi2(3)      =       9.20
                                                  Prob > chi2     =     0.0267
Log likelihood = -1088.8724                       Pseudo R2       =     0.0042

------------------------------------------------------------------------------
          ct |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        gind |
           1 |   .4706954   .4239421     1.11   0.267    -.3602159    1.301607
           2 |   .6663683   .2637133     2.53   0.012     .1494998    1.183237
           3 |   .9562032    .672907     1.42   0.155    -.3626703    2.275077
             |
       _cons |  -.7330597   .0529515   -13.84   0.000    -.8368426   -.6292767
------------------------------------------------------------------------------
```

$$\text{Logit } P = -0{,}73 + 0{,}47\ X_1 + 0{,}67\ X_2 + 0{,}96\ X_3$$

$$OR_{Hcg} = e^{0{,}47} = 1{,}60 \qquad OR_{Clomid} = e^{0{,}67} = 1{,}95$$

$$OR_{Autre} = e^{0{,}96} = 2{,}61$$

Lorsqu'on décompose X en variables indicatrices, celles-ci doivent être toutes présentes (ou absentes) du modèle *en même temps*.

Supposons qu'on enlève X_3

		X_1	X_2	X_3
	0	0	0	0
	1	1	0	0
X =	2	0	1	0
	3	0	0	1

$$\text{Logit } P = \alpha' + \beta'_1 X_1 + \beta'_2 X_2$$

Les catégories 0 et 3 de X ne sont plus distinguables avec les variables restantes X_1 et X_2.

Le coefficient β'_1 de la variable X_1 n'a plus le sens précédent.

Il devient le logarithme de l'odds ratio entre la catégorie X=1 et les catégories X = 0 et X = 3 réunies.

=> retirer X_3 revient à regrouper les catégories X = 0 et X = 3.

$$\text{Logit } P = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

Le test de l'association entre X et la maladie doit être un test global

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \qquad H_1 : \beta_1 \neq 0 \text{ ou } \beta_2 \neq 0 \text{ ou } \beta_3 \neq 0$$

```
. logistic ct i.gind

Logistic regression                               Number of obs   =       1719
                                                  LR chi2(3)      =       9.20
                                                  Prob > chi2     =     0.0267
Log likelihood = -1088.8724                       Pseudo R2       =     0.0042

------------------------------------------------------------------------------
          ct | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       gind |
          1 |   1.601107   .6787768     1.11   0.267     .6975257    3.675197
          2 |   1.947153     .51349     2.53   0.012     1.161253    3.264925
          3 |   2.601799   1.750769     1.42   0.155     .6958158    9.728665
            |
      _cons |   .4804368   .0254398   -13.84   0.000     .4330757    .5329772
------------------------------------------------------------------------------
```

Test de Wald

```
. testparm i.gind  // équivalent à testparm 1.gind 2.gind 3.gind

 ( 1)  [ct]1.gind = 0
 ( 2)  [ct]2.gind = 0
 ( 3)  [ct]3.gind = 0

           chi2(  3) =    9.36
         Prob > chi2 =    0.0248
```

Test du rapport de vraisemblance

```
. est store a

. logistic ct if gin!=.

Logistic regression                               Number of obs   =       1719
                                                  LR chi2(0)      =       0.00
                                                  Prob > chi2     =          .
Log likelihood = -1093.4734                       Pseudo R2       =     0.0000

------------------------------------------------------------------------------
          ct | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   .4986922   .0255265   -13.59   0.000     .4510891    .5513189
------------------------------------------------------------------------------

. est store b

. lrtest a b

Likelihood-ratio test                             LR chi2(3)  =       9.20
(Assumption: b nested in a)                       Prob > chi2 =     0.0267
```

On présente les résultats de la façon suivante :

Induction de la grossesse	Total	Témoins %	Cas %	OR et IC	p
non	1627	96	92	1	< 0,03
Hcg	23	1,1	1,8	1,6 [0,7 ; 3,7]	
Clomid	60	2,7	5,1	1,9 [1,2 ; 3,3]	
Autre	9	0,35	0,87	2,6 [0,7 ; 9,7]	

```
. logistic ct i.gind

Logistic regression                               Number of obs   =       1719
                                                  LR chi2(3)      =       9.20
                                                  Prob > chi2     =     0.0267
Log likelihood = -1088.8724                       Pseudo R2       =     0.0042

------------------------------------------------------------------------------
          ct | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        gind |
           1 |   1.601107   .6787768     1.11   0.267     .6975257    3.675197
           2 |   1.947153     .51349     2.53   0.012     1.161253    3.264925
           3 |   2.601799   1.750769     1.42   0.155     .6958158    9.728665
             |
       _cons |   .4804368   .0254398   -13.84   0.000     .4330757    .5329772
------------------------------------------------------------------------------

. testparm i.gind  // équivalent à testparm 1.gind 2.gind 3.gind

 ( 1)  [ct]1.gind = 0
 ( 2)  [ct]2.gind = 0
 ( 3)  [ct]3.gind = 0

           chi2(  3) =    9.36
         Prob > chi2 =    0.0248
```

Interprétation des coefficients et codage des variables

3. X qualitative ordinale

Exemple

Consommation de tabac comptée en cigarettes par jour

```
. tab ct tabfc

    0:acc |          0:nf;1:1-9;2:10-19;3:>=20
    1:GEU |         0          1          2          3 |     Total
----------+--------------------------------------------+----------
        0 |       809        158        106         71 |     1,144
        1 |       264         78        113         94 |       549
----------+--------------------------------------------+----------
    Total |     1,073        236        219        165 |     1,693
```

En prenant les non fumeuses comme référence, il faut 3 odds ratios pour décrire l'ensemble de l'association GEU - tabac :

Nombre de cigarettes par jour	0	1-9	10-19	≥ 20
OR	1	1,51	3,27	4,06

Pour inclure la variable tabac dans un modèle logistique, il y a deux possibilités :

• "conserver" les 3 odds ratios en décomposant la variable tabac en variables indicatrices (comme précédemment).

• "modéliser" l'association GEU - tabac pour représenter la relation dose-effet (de façon similaire à une régression linéaire)

A. Décomposition en variables indicatrices

Le modèle correspondant est : Logit $P = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

```
logistic ct i.tabfc

Logistic regression                             Number of obs   =       1693
                                                LR chi2(3)      =     107.61
                                                Prob > chi2     =     0.0000
Log likelihood = -1012.8708                     Pseudo R2       =     0.0504

------------------------------------------------------------------------------
          ct | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       tabfc |
          1  |   1.512802   .2352084     2.66   0.008      1.11542    2.051756
          2  |    3.26676   .4987294     7.75   0.000     2.421955    4.406242
          3  |   4.057085   .6997357     8.12   0.000     2.893378     5.68883
             |
       _cons |   .3263288   .0231302   -15.80   0.000     .2840026    .3749631
------------------------------------------------------------------------------
```

On retrouve effectivement les 3 odds ratios précédents.

Choix de la catégorie de référence pour la décomposition :

- Il s'impose en pratique quand il y a une catégorie "non exposé".
- Sinon, plusieurs choix sont possibles

Exemple : niveau d'études

```
. tab ct et4

    0:acc |      0:aucune;1:prim;2:sec,tech;3:univ
    1:GEU |         0          1          2          3 |     Total
----------+--------------------------------------------+----------
        0 |         4         93        781        258 |     1,136
        1 |         6         36        377        127 |       546
----------+--------------------------------------------+----------
    Total |        10        129      1,158        385 |     1,682
```

Référence : X=0 (pas d'études)

```
. logistic ct i.et4

Logistic regression                             Number of obs     =      1,682
                                                LR chi2(3)        =       4.49
                                                Prob > chi2       =     0.2134
Log likelihood = -1057.9179                     Pseudo R2         =     0.0021

------------------------------------------------------------------------------
          ct | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        et4 |
          1 |   .2580646   .1741118    -2.01   0.045      .068776    .9683216
          2 |   .3218097   .2087054    -1.75   0.080     .0902743    1.147188
          3 |   .3281655   .2147959    -1.70   0.089     .0909811     1.18368
            |
       _cons |        1.5   .9682455     0.63   0.530     .4232948    5.315442
------------------------------------------------------------------------------
```

```
. testparm i.et4

 ( 1)  [ct]1.et4 = 0
 ( 2)  [ct]2.et4 = 0
 ( 3)  [ct]3.et4 = 0

           chi2(  3) =    4.38
         Prob > chi2 =    0.2235
```

Référence : X=1 (niveau d'études primaire)

```
. logistic ct ib1.et4

Logistic regression                             Number of obs     =      1,682
                                                LR chi2(3)        =       4.49
                                                Prob > chi2       =     0.2134
Log likelihood = -1057.9179                     Pseudo R2         =     0.0021

------------------------------------------------------------------------------
          ct | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        et4 |
          0 |   3.874999   2.614396     2.01   0.045     1.032715    14.53995
          2 |   1.247012   .2569675     1.07   0.284      .832658    1.867561
          3 |   1.271641   .2851439     1.07   0.284     .8194019    1.973477
            |
       _cons |   .3870968   .0759839    -4.84   0.000     .2634727    .5687264
------------------------------------------------------------------------------
```

```
. testparm i.et4

 ( 1)  [ct]0.et4 = 0
 ( 2)  [ct]2.et4 = 0
 ( 3)  [ct]3.et4 = 0

           chi2(  3) =    4.38
         Prob > chi2 =    0.2235
```

Finalement, le choix de la catégorie de référence :

- n'a pas de conséquence pour retrouver les différents odds ratios : $1{,}247 = \dfrac{0{,}322}{0{,}258}$

- peut changer "l'impression" donnée par les odds ratios présentés ou leur précision si une catégorie a un effectif petit

B. Modélisation de l'association GEU - tabac

La décomposition en variables indicatrices ne tient pas compte de l'ordre des catégories de la variable tabac.
=> perte d'information et perte de puissance possible s'il y a une relation dose-effet.

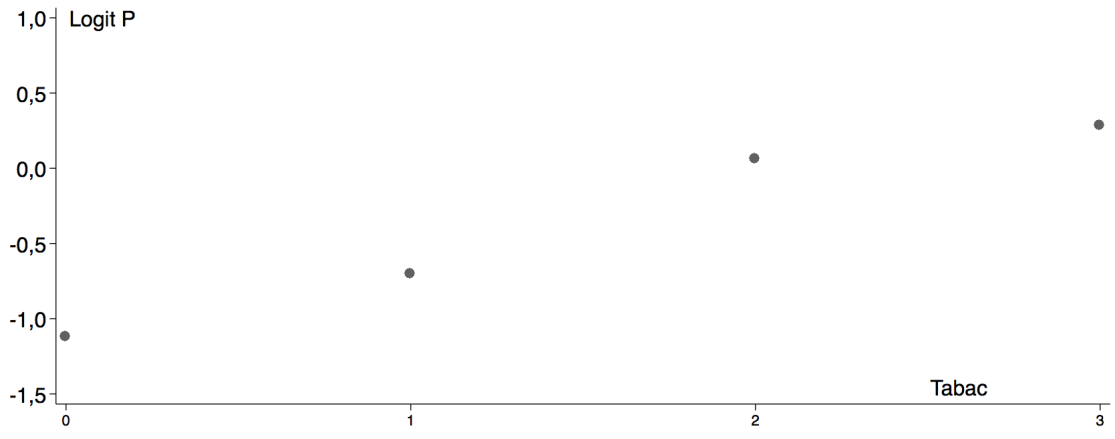
0:acc / 1:GEU	0:nf;1:1-9;2:10-19;3:>=20 0	1	2	3	Total
0	809	158	106	71	1,144
1	264	78	113	94	549
Total	1,073	236	219	165	1,693
OR(observé)	1	1,51	3,27	4,06	

Le calcul des Logit P pour les 4 catégories de tabac donne :

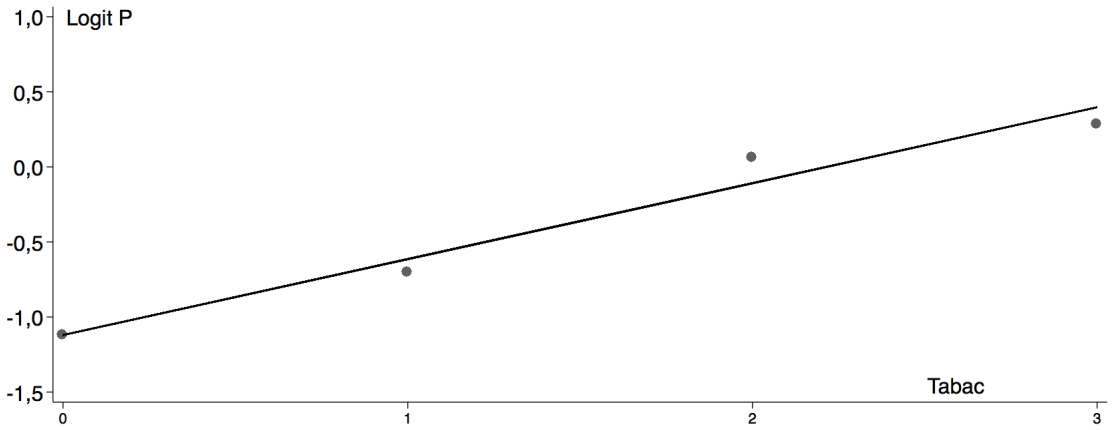
$$\text{Logit } P_0 = \text{Ln}\left(\frac{P_o}{1-P_o}\right) = \text{Ln}\left(\frac{264/1073}{809/1073}\right) = \text{Ln}\left(\frac{264}{809}\right) = -1{,}12$$

de même : Logit P_1 = - 0,71 Logit P_2 = 0,064 Logit P_3 = 0,28

D'où la représentation graphique :



=> on peut raisonnablement représenter la variation des logits en fonction de la quantité de cigarettes par une droite.



Le modèle correspondant est : Logit P = $\alpha + \beta X$

```
. logit ct tabfc

Logistic regression                               Number of obs   =       1693
                                                  LR chi2(1)      =     104.99
                                                  Prob > chi2     =     0.0000
Log likelihood = -1014.1795                       Pseudo R2       =     0.0492

------------------------------------------------------------------------------
          ct |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       tabfc |    .505858   .0499975    10.12   0.000     .4078648    .6038512
       _cons |  -1.120282   .0673178   -16.64   0.000    -1.252222   -.9883411
------------------------------------------------------------------------------
```

exp(β) = exp(0,506) = 1,66

Nombre de cigarettes par jour	0	1-9	10-19	≥ 20
X	0	1	2	3
OR observé	1	1,51	3,27	4,07
OR donné par le modèle Logit P = $\alpha + \beta X$	1	1,66	2,75	4,56

Les odds ratios donnés par le modèle sont différents des odds ratios observés, mais on a résumé l'association GEU - tabac par un seul paramètre : β = 0,506 (ou OR = 1,66).

```
. logit ct tabfc

Logistic regression                               Number of obs   =       1693
                                                  LR chi2(1)      =     104.99
                                                  Prob > chi2     =     0.0000
Log likelihood = -1014.1795                       Pseudo R2       =     0.0492

------------------------------------------------------------------------------
          ct |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       tabfc |    .505858   .0499975    10.12   0.000     .4078648    .6038512
       _cons |  -1.120282   .0673178   -16.64   0.000    -1.252222   -.9883411
------------------------------------------------------------------------------
```

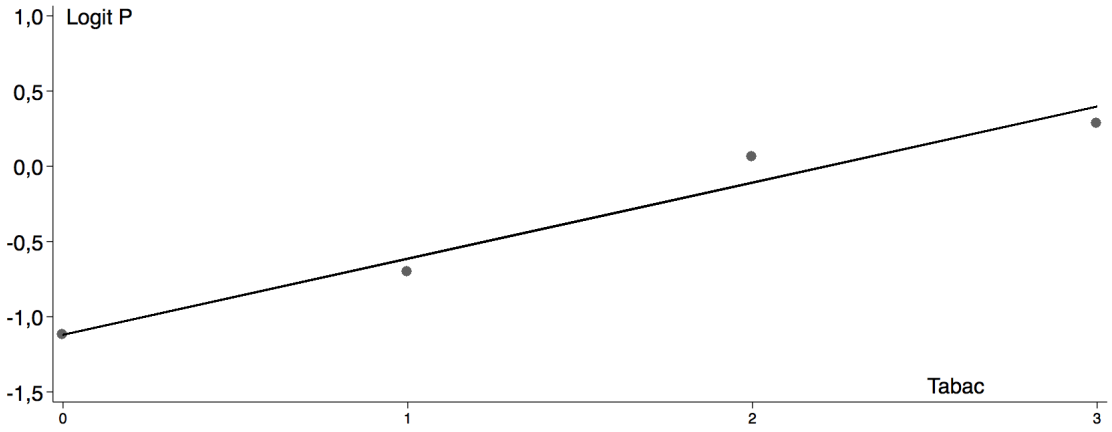
Nombre de cigarettes par jour	0	1-9	10-19	≥ 20
X	0	1	2	3
OR observé	1	1,51	3,27	4,07
OR donné par le modèle Logit P = $\alpha + \beta$ X	1	1,66	2,75	4,56

Pour trouver avec Stata l'odds ratio correspondant à X=2 et son intervalle de confiance, la commande est lincom (linear combination).

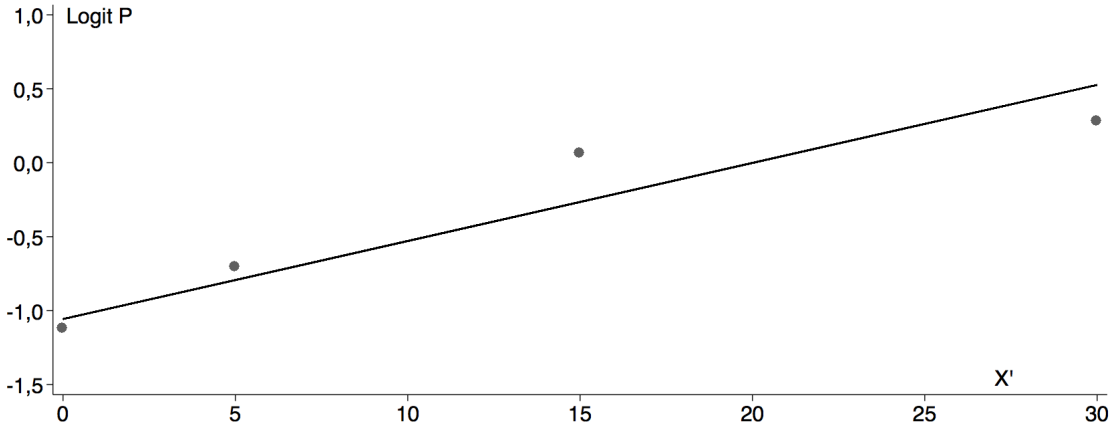
```
. lincom 2*tabfc,or

 ( 1)  2*[ct]tabfc = 0

------------------------------------------------------------------------------
          ct | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   2.750316   .2750177    10.12   0.000     2.260824    3.345789
------------------------------------------------------------------------------
```

Choix des valeurs de X

Nombre de cigarettes par jour	0	1-9	10-19	≥ 20
X	0	1	2	3
OR observé	1	1,51	3,27	4,07
OR donné par le modèle Logit P = $\alpha + \beta$ X	1	1,66	2,75	4,56



Nombre de cigarettes par jour	0	1-9	10-19	≥ 20
X' (centre de la classe)	0	5	15	30
OR observé	1	1,51	3,27	4,07
OR donné par le modèle Logit P = $\alpha + \beta$ X'	1	1,30	2,21	4,87

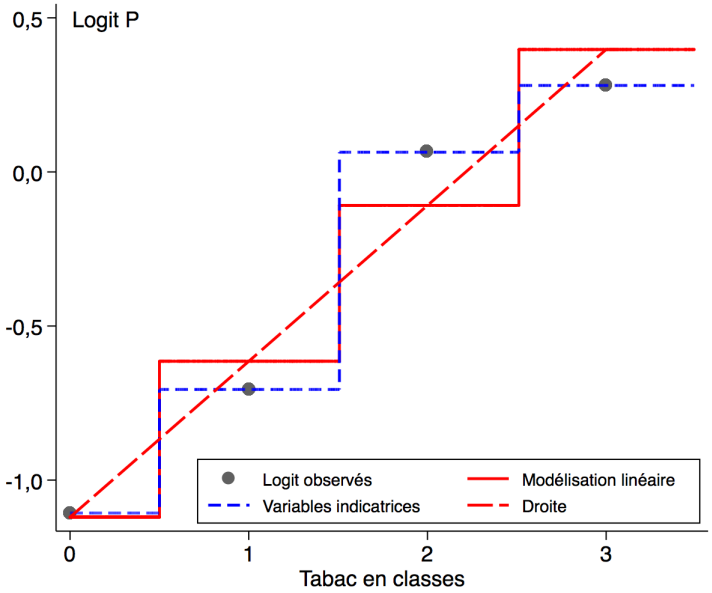


NB : il n'y a pas de test pour comparer les deux codages, mais on peut, pour chacun d'eux, tester s'il s'écarte de la linéarité.

C. Choix entre variables indicatrices et modélisation linéaire

En fait, il est abusif de représenter la relation tabac/GEU avec la modélisation linéaire par une droite car la variable tabac est en classes avec 4 valeurs.

Il vaut mieux la représenter par un escalier (qui a des "marches régulières")



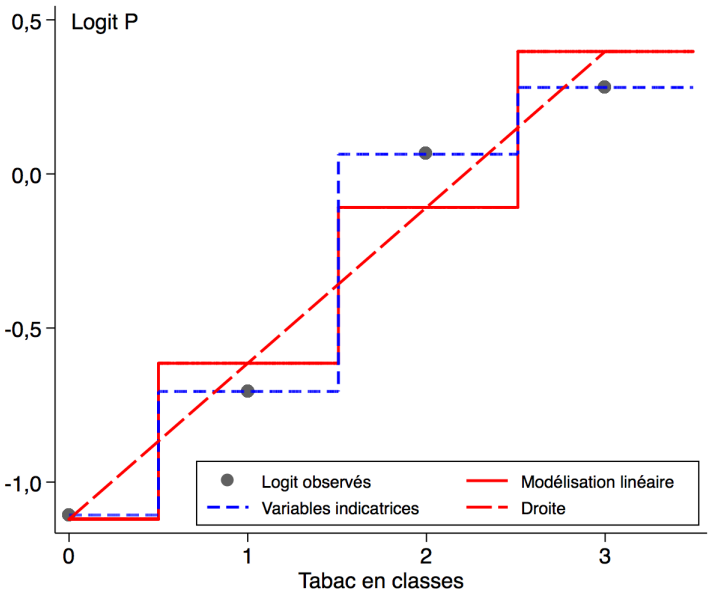
Modèle avec variables indicatrices :

Logit P = $\alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

Modèle "linéaire" : Logit P = $\alpha + \beta$ X

(avec X = 0, 1, 2 ou 3)

Droite : pas de modèle correspondant



Modèle avec variables indicatrices :

Logit P = $\alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

Modèle "linéaire" : Logit P = $\alpha + \beta$ X (avec X = 0, 1, 2 ou 3)

Droite : pas de modèle correspondant

Modèle avec variables indicatrices :

- représente la relation par un escalier (ce qui a peu de chances d'être la réalité, mais le fait de mettre la variable tabac en classes n'est pas non plus très réaliste ...)

- est le plus proche possible des observations puisqu'il passe par tous les points observés

- est sensible aux fluctuations d'échantillonnage (les hauteurs des "marches" peuvent être très variables selon l'échantillon, surtout si une classe a un effectif petit)

Modélisation linéaire :

- utilise l'information concernant l'ordre des catégories de X

- représente la relation par une droite (ou du moins par un escalier régulier) (ce qui ne peut être qu'une approximation de la réalité)

- résume et synthétise la relation par un coefficient (la pente de la droite), ce qui donne un test de tendance

On peut comparer les 2 modèles car ils sont emboîtés (voir poly spécifique). Le test est un test de linéarité

Comparaison entre :

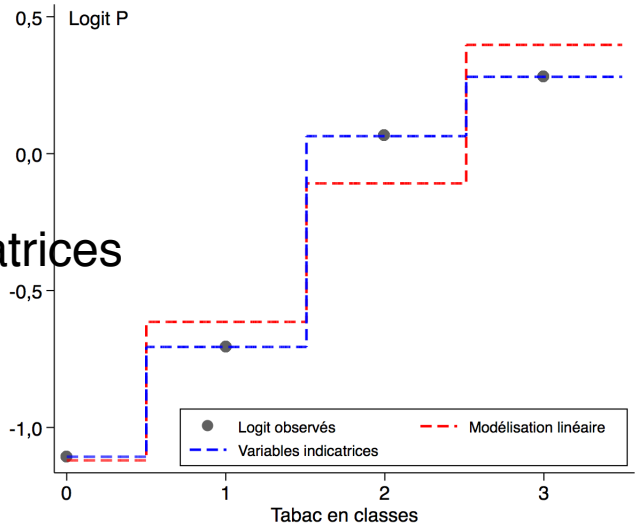
• la modélisation linéaire

Logit P = $\alpha + \beta X$

• la décomposition en variables indicatrices

Logit P = $\alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

Test du rapport de vraisemblance (test de linéarité)




```
. logit ct tabfc

Logistic regression                             Number of obs   =       1693
                                                LR chi2(1)      =     104.99
                                                Prob > chi2     =     0.0000
Log likelihood = -1014.1795                     Pseudo R2       =     0.0492

------------------------------------------------------------------------------
          ct |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       tabfc |    .505858   .0499975    10.12   0.000     .4078648    .6038512
       _cons |  -1.120282   .0673178   -16.64   0.000    -1.252222   -.9883411
------------------------------------------------------------------------------

. est store lin

. logit ct i.tabfc

Logistic regression                             Number of obs   =       1693
                                                LR chi2(3)      =     107.61
                                                Prob > chi2     =     0.0000
Log likelihood = -1012.8708                     Pseudo R2       =     0.0504

------------------------------------------------------------------------------
          ct |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       tabfc |
          1  |   .4139636   .1554786     2.66   0.008     .1092311    .7186961
          2  |   1.183799   .1526679     7.75   0.000     .8845749    1.483022
          3  |   1.400465   .1724725     8.12   0.000     1.062425    1.738505
             |
       _cons |   -1.11985     .07088   -15.80   0.000    -1.258772   -.9809276
------------------------------------------------------------------------------

. est store indic
. lrtest indic lin

Likelihood-ratio test                           LR chi2(2)  =      2.62
(Assumption: lin nested in indic)               Prob > chi2 =    0.2702
```

Différence non significative :

- linéarité non rejetée
- on retient la modélisation linéaire

Interprétation des coefficients et codage des variables

4. X quantitative

• On peut inclure X dans le modèle avec son codage initial sous la forme : Logit P = $\alpha + \beta$ X

=> la liaison entre Logit P et X est alors supposée linéaire

D'autres modélisations sont possibles (voir le chapitre consacré aux variables quantitatives).

Exercice 2

1. Pour tester si la liaison entre la consommation de tabac et Logit P est linéaire, on procède comme cela vient d'être indiqué en prenant la variable tabfc des pages précédentes.
On considère donc les deux modèles logistiques :

M_1 : Logit $P = \alpha_0 + \beta_1 tabfc_1 + \beta_2 tabfc_2 + \beta_3 tabfc_3$ où $tabfc_1$, $tabfc_2$ et $tabfc_3$ sont les variables indicatrices des catégories de tabfc.

M_2 : Logit $P = \alpha + \beta$ tabfc

Ecrivez les hypothèses testées lorsqu'on compare ces deux modèles. Faites le test et donnez votre conclusion.

2. Etudiez le lien entre le nombre de fausses couches spontanées antérieures (variable nafcs) et le risque de GEU.

Interaction

E : exposition (en 0/1)
X : variable en 0/1

Il y a interaction entre E et X si l'odds ratio associé à E est différent pour X=0 et X=1.

Le modèle Logit P = $\alpha+\beta E+\gamma X$ "tient compte" des deux variables E et X, mais suppose l'absence d'interaction :

X = 0 -> Ln OR_o = Logit P_1 - Logit P_o = $\alpha + \beta - \alpha = \beta$

X = 1 -> Ln OR_1 = Logit P_1 - Logit P_o = $\alpha + \beta + \gamma - (\alpha + \gamma) = \beta$

En présence d'une interaction, on a $OR_1 \neq OR_o$
C'est-à-dire :

X = 0 -> Ln $OR_o = \beta$
X = 1 -> Ln $OR_1 = \beta + \delta$

d'où le modèle : Logit P = $\alpha + \beta E + \gamma X + \delta EX$

avec EX = 0 si E = 0 ou X = 0
EX = 1 si E = X = 1

Les variables d'interaction sont construites par Stata avec le système de "factor variables".
La syntaxe est i.var1##i.var2

Exemple

Induction de la grossesse : 0 = non induite
1 = induction par Clomid

Age30 : 0 = âge < 30 ans
1 = âge ≥ 30 ans

```
. logit ct i.clomid##i.age30

Logistic regression                             Number of obs   =       1683
                                                LR chi2(3)      =      44.67
                                                Prob > chi2     =     0.0000
Log likelihood =  -1046.124                     Pseudo R2       =     0.0209

------------------------------------------------------------------------------
          ct |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.clomid |    1.234339   .3807449     3.24   0.001     .4880927    1.980585
     1.age30 |    .6511233   .1072584     6.07   0.000     .4409006    .8613459
             |
clomid#age30 |
       1 1   |   -1.127027   .5354195    -2.10   0.035    -2.17643    -.0776236
             |
       _cons |     -1.0267   .0744555   -13.79   0.000    -1.17263    -.8807696
------------------------------------------------------------------------------
```

La commande i.clomid##i.age30 créée :

- des variables indicatrices pour clomid et age30 et exclut du calcul celles correspondant aux catégories de référence.

- des termes d'interaction combinant les variables indicatrices :
0.clomid#0.age30, 0.clomid#1.age30, 1.clomid#0.age30 et 1.clomid#1.age30.
Ceux correspondant aux catégories de référence sont exclus, il ne reste donc ici que le dernier.

Exemple (suite)

Induction de la grossesse : 0 = non induite
1 = induction par Clomid

Age30 : 0 = âge < 30 ans
1 = âge ≥ 30 ans

```
. logit ct i.clomid##i.age30

Logistic regression                             Number of obs   =       1683
                                                LR chi2(3)      =      44.67
                                                Prob > chi2     =     0.0000
Log likelihood =  -1046.124                     Pseudo R2       =     0.0209

------------------------------------------------------------------------------
          ct |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.clomid |    1.234339   .3807449     3.24   0.001      .4880927    1.980585
     1.age30 |    .6511233   .1072584     6.07   0.000      .4409006    .8613459
             |
clomid#age30 |
       1 1   |   -1.127027   .5354195    -2.10   0.035     -2.17643   -.0776236
             |
       _cons |     -1.0267   .0744555   -13.79   0.000     -1.17263   -.8807696
------------------------------------------------------------------------------
```

Logit P = -1,027 + 1,234 CLOMID+ 0,651 AGE30
- 1,127 CLOMID×AGE30

Il y a **deux** odds-ratios associés au Clomid :

- chez les femmes de moins de 30 ans : OR = $e^{1,234}$ = 3,43
- chez les femmes de plus de 30 ans: OR = $e^{1,234-1,127}$ = 1,11

Le test du coefficient de la variable d'interaction est un test de comparaison de ces deux odds-ratios. Il est ici à la limite de la signification (p=0,035).
On peut le reproduire avec la commande testparm :

```
. testparm 1.clomid#1.age30
( 1)  [ct]1.clomid#1.age30 = 0
chi2(  1) =    4.43
      Prob > chi2 =    0.0353
```

Exemple (suite)

Attention à bien mettre ## et pas #.

```
. logit ct i.clomid##i.age30

Logistic regression                             Number of obs     =       1683
                                                LR chi2(3)        =      44.67
                                                Prob > chi2       =     0.0000
Log likelihood =  -1046.124                     Pseudo R2         =     0.0209

------------------------------------------------------------------------------
          ct |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.clomid |    1.234339   .3807449     3.24   0.001     .4880927    1.980585
     1.age30 |    .6511233   .1072584     6.07   0.000     .4409006    .8613459
             |
clomid#age30 |
       1 1   |   -1.127027   .5354195    -2.10   0.035    -2.17643    -.0776236
             |
       _cons |     -1.0267   .0744555   -13.79   0.000    -1.17263    -.8807696
------------------------------------------------------------------------------
```

```
. logit ct i.clomid#i.age30

Logistic regression                             Number of obs     =       1683
                                                LR chi2(3)        =      44.67
                                                Prob > chi2       =     0.0000
Log likelihood =  -1046.124                     Pseudo R2         =     0.0209

------------------------------------------------------------------------------
          ct |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
clomid#age30 |
       0 1   |    .6511233   .1072584     6.07   0.000     .4409006    .8613459
       1 0   |    1.234339   .3807449     3.24   0.001     .4880927    1.980585
       1 1   |    .7584357   .3758859     2.02   0.044     .0217128    1.495159
             |
       _cons |     -1.0267   .0744555   -13.79   0.000    -1.17263    -.8807696
------------------------------------------------------------------------------
```

Les résultats sont équivalents, mais ne s'interprètent pas de la même façon ...

Intervalle de confiance des odds ratios en présence d'interaction

Logit P = $\alpha + \beta E + \gamma X + \delta EX$

Ln $OR_0 = \beta$ Ln $OR_1 = \beta + \delta$

IC de β : $\hat{\beta} \pm 1{,}96\sqrt{var(\hat{\beta})}$

IC de $\beta+\delta$: $\hat{\beta} + \hat{\delta} \pm 1{,}96\sqrt{var(\hat{\beta} + \hat{\delta})}$

$= \hat{\beta} + \hat{\delta} \pm 1{,}96\sqrt{var(\hat{\beta}) + var(\hat{\delta}) + 2\,cov(\hat{\beta}, \hat{\delta})}$

L'IC de β (et donc de OR_0) se lit directement dans les résultats de la commande logit (ou logistic).

Ce n'est pas le cas de l'IC de $\beta+\delta$ qui fait intervenir la covariance entre β et δ. Il faut utiliser la commande lincom qui permet de calculer l'IC d'une combinaison linéaire des coefficients (et donc de l'OR correspondant).

Avec E=clomid et X=age30, on obtient les résultats suivants :

OR_0 = 3,44 [1,63 ; 7,25]

OR_1 = 1,11 [0,53 ; 2,33]

```
. logistic ct i.clomid##i.age30

Logistic regression                             Number of obs     =      1,683
                                                LR chi2(3)        =      44.67
                                                Prob > chi2       =     0.0000
Log likelihood =  -1046.124                     Pseudo R2         =     0.0209

------------------------------------------------------------------------------
          ct | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.clomid |   3.436107    1.30828     3.24   0.001     1.629206    7.246984
     1.age30 |   1.917694   .2856888     6.07   0.000     1.431186    2.569344
             |
clomid#age30 |
       1 1   |   .3239952   .1734734    -2.10   0.035     .1134459    .9253126
             |
       _cons |   .3581871    .026669   -13.79   0.000     .3095518    .4144638
------------------------------------------------------------------------------


. lincom 1.clomid + 1.age30#1.clomid

 ( 1)  [ct]1.clomid + [ct]1.age30#1.clomid = 0

------------------------------------------------------------------------------
          ct | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   1.113282   .4190843     0.29   0.776     .5323271    2.328262
------------------------------------------------------------------------------
```

Interaction et analyses séparées pour X=0 et X=1

1. Avec terme d'interaction

```
. logit ct i.clamid##i.age30

Logistic regression                             Number of obs   =       1683
                                                LR chi2(3)      =      44.67
                                                Prob > chi2     =     0.0000
Log likelihood =  -1046.124                     Pseudo R2       =     0.0209

------------------------------------------------------------------------------
          ct |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.clamid |    1.234339   .3807449     3.24   0.001      .4880927    1.980585
     1.age30 |    .6511233   .1072584     6.07   0.000      .4409006    .8613459
             |
clamid#age30 |
       1 1   |   -1.127027   .5354195    -2.10   0.035     -2.17643   -.0776236
             |
       _cons |     -1.0267   .0744555   -13.79   0.000     -1.17263   -.8807696
------------------------------------------------------------------------------
```

2. Analyses séparées

X=0 (moins de 30 ans)

```
. logit ct clamid if age30==0

Logistic regression                             Number of obs   =        958
                                                LR chi2(1)      =      10.35
                                                Prob > chi2     =     0.0013
Log likelihood = -555.90136                     Pseudo R2       =     0.0092
------------------------------------------------------------------------------
          ct |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      clamid |    1.234339   .3807449     3.24   0.001      .4880928    1.980585
       _cons |     -1.0267   .0744555   -13.79   0.000     -1.17263   -.8807696
------------------------------------------------------------------------------
```

X=1 (plus de 30 ans)

```
. logit ct clamid if age30==1

Logistic regression                             Number of obs   =        725
                                                LR chi2(1)      =       0.08
                                                Prob > chi2     =     0.7761
Log likelihood = -490.22259                     Pseudo R2       =     0.0001
------------------------------------------------------------------------------
          ct |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      clamid |    .1073125   .3764404     0.29   0.776    -.6304971     .845122
       _cons |   -.3755765   .0772059    -4.86   0.000    -.5268973   -.2242556
------------------------------------------------------------------------------
```

On peut remarquer que :

- n = 1683 = $n_0 + n_1$ = 958 + 725 LnV = -1046,1 = -555,9-490,2
- les coefficients de clomid sont les mêmes (0,107 = 1,234-1,127) ainsi que les intervalles de confiance

Donc : résultats identiques sans modification de puissance.

L'interaction est une notion symétrique entre les deux variables E et X.

```
. logit ct i.clomid##i.age30

Logistic regression                             Number of obs   =       1683
                                                LR chi2(3)      =      44.67
                                                Prob > chi2     =     0.0000
Log likelihood =  -1046.124                     Pseudo R2       =     0.0209

------------------------------------------------------------------------------
          ct |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.clomid |    1.234339   .3807449     3.24   0.001     .4880927    1.980585
     1.age30 |    .6511233   .1072584     6.07   0.000     .4409006    .8613459
             |
clomid#age30 |
       1 1   |   -1.127027   .5354195    -2.10   0.035    -2.17643    -.0776236
             |
       _cons |     -1.0267   .0744555   -13.79   0.000    -1.17263    -.8807696
------------------------------------------------------------------------------
```

Logit P = -1,027 + 0,651 AGE30 + 1,234 CLOMID
- 1,127 AGE30×CLOMID

Il y a **deux** odds-ratios associés au Clomid :

- chez les femmes de moins de 30 ans : OR = $e^{1,234}$ = 3,43
- chez les femmes de plus de 30 ans: OR = $e^{1,234-1,127}$ = 1,11

Il y a aussi **deux** odds-ratios associés à l'âge supérieur à 30 ans :

- chez les femmes sans clomid : OR = $e^{0,651}$ = 1,92
- chez les femmes avec clomid : OR = $e^{0,651-1,127}$ = 0,62

Interaction et facteur à plus de 2 classes

X en 3 classes -> décomposition en X_1 et X_2

Logit $P = \alpha + \beta E + \beta_1 X_1 + \beta_2 X_2 + \delta_1 E X_1 + \delta_2 E X_2$

Test de l'interaction : $H_o : \delta_1 = 0$ et $\delta_2 = 0$

Exemple : agec3 = 0 : < 30 ; 1 : 30-34 ; 2 : ≥ 35

```
. logistic ct i.clomid##i.agec3

Logistic regression                                     Number of obs   =       1683
                                                        LR chi2(5)      =      53.76
                                                        Prob > chi2     =     0.0000
Log likelihood = -1041.5811                             Pseudo R2       =     0.0252

------------------------------------------------------------------------------
          ct | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.clomid |   3.436107   1.30828     3.24   0.001     1.629206    7.246984
             |
       agec3 |
          1  |   1.690709   .1984507     4.47   0.000     1.343251    2.128043
          2  |   2.729796    .456932     6.00   0.000     1.966302    3.789746
             |
clomid#agec3 |
        1 1  |   .2746101   .1623754    -2.19   0.029     .0861799    .8750387
        1 2  |   .4960686   .4152704    -0.84   0.402     .0961571    2.559187
             |
       _cons |   .3581871    .026669   -13.79   0.000     .3095518    .4144638
------------------------------------------------------------------------------
```

Test de l'interaction

```
. testparm 1.clomid#1.agec3 1.clomid#2.agec3    //  ou testparm 1.c*#1.a* 1.c*#2.a*
 ( 1)  [ct]1.clomid#1.agec3 = 0
 ( 2)  [ct]1.clomid#2.agec3 = 0
           chi2(  2) =    4.82
         Prob > chi2 =    0.0898
```

Attention : 3,44 = OR_{Clomid} chez les femmes < 30 ans

Il y a 3 odds-ratios associés au Clomid :

Chez les femmes < 30 ans : $OR_1 = e^{\beta} = 3,44$

Chez les femmes 30-34 ans : $OR_2 = e^{\beta+\delta_1} = 3,44 \times 0,27 = 0,93$

Chez les femmes ≥ 35 ans : $OR_3 = e^{\beta+\delta_2} = 3,44 \times 0,50 = 1,72$

Interaction et variable continue

Logit P = $\alpha + \beta E + \gamma X + \delta EX$

Exemple
agea = âge en années

```
. logistic ct i.clomid##c.agea

Logistic regression                             Number of obs   =       1683
                                                LR chi2(3)      =      63.91
                                                Prob > chi2     =     0.0000
Log likelihood = -1036.5066                     Pseudo R2       =     0.0299

------------------------------------------------------------------------------
          ct | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.clomid |   112.8515   213.0682     2.50   0.012     2.788711    4566.787
       agea |   1.084319   .0119044     7.37   0.000     1.061237    1.107904
            |
clomid#c.agea |
          1 |   .8735553   .0537033    -2.20   0.028     .7743929    .9854156
            |
       _cons |   .0447012   .0147564    -9.41   0.000      .023406     .085371
------------------------------------------------------------------------------
```

Attention : 112,9 = OR_{Clomid} chez les femmes de 0 ans !

Si on remplace agea par ageabis=agea-25, les résultats deviennent plus facilement interprétables :
3,84 = OR_{Clomid} chez les femmes de 25 ans

```
. logistic ct i.clomid##c.ageabis

Logistic regression                             Number of obs   =      1,683
                                                LR chi2(3)      =      63.91
                                                Prob > chi2     =     0.0000
Log likelihood = -1036.5066                     Pseudo R2       =     0.0299

------------------------------------------------------------------------------
          ct | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.clomid |   3.843857   1.639776     3.16   0.002     1.665895    8.869248
     ageabis |   1.084319   .0119044     7.37   0.000     1.061237    1.107904
            |
clomid#c.ageabis |
          1 |   .8735553   .0537033    -2.20   0.028     .7743929    .9854156
            |
       _cons |   .3382596    .025155   -14.58   0.000     .2923814    .3913367
------------------------------------------------------------------------------
```

Exercice 3

1. Y a-t-il un lien entre d'une part le risque de GEU et d'autre part le nombre de partenaires sexuels (variable npart) et l'âge au premier rapport sexuel (variable aprs) ?

2. Y a-t-il une interaction entre ces deux variables (npart et aprs) ?

3. Quel est l'odds ratio (et son intervalle de confiance) associé à une "exposition" conjointe à plus de 5 partenaires sexuels et un âge au premier rapport sexuel inférieur à 14 ans ?

Note : Le nombre de partenaires sexuel est une variable en 3 classes codées 1 à 3 (1, 2-5, >5). L'âge au premier rapport sexuel est une variable en 4 classes codées 1 à 4 (≤14, 15-17, 18-20, >20)

Choix des variables à inclure dans un modèle logistique

Principes généraux

• C'est le cœur de l'analyse des enquêtes épidémiologiques

• Question difficile, toujours discutée aujourd'hui et sans solution uniformément meilleure que les autres

• Importance d'expliciter ce qu'on fait

◊ Le choix des variables suppose la définition de critères précisant ce qu'est le "meilleur" modèle, ce qui peut dépendre des objectifs de l'étude :

– prédire un événement d'intérêt (maladie, guérison, grossesse, ...)
  - on privilégie les qualités prédictives du modèle (≈ adéquation du modèle)
  - la limitation du nombre de variables n'est pas toujours un objectif prioritaire

– identifier les facteurs de risque importants de la maladie, comprendre et quantifier leur effet (plus fréquent en épidémiologie)
  - variables scientifiquement importantes prises en compte
  - phénomènes de confusion contrôlés
  - résultats suffisamment stables pour être extrapolables

◊ Le choix des variables repose sur un “mélange” :

– de connaissances scientifiques indépendamment des données ("subject-matter knowledge") (souvent limitées)
– d’utilisation de méthodes statistiques appliquées aux données ("data driven procedures")
– d’expérience et de bon sens

◊ Il faut trouver un équilibre entre :

– Trop de facteurs pris en compte
  - perte de puissance
  - sur-ajustement ou même "création" de biais de confusion
  - résultats instables, et moins interprétables
– Pas assez de facteurs pris en compte
  - possibilité de confusion résiduelle
  - moins bonne adéquation du modèle
  - interprétation plus facile

◊ On procède en général en deux étapes

1 . Choix des variables “candidates”

L’objectif est de faire un premier tri parmi l’ensemble des variables disponibles dans un fichier de données ou une enquête

2 . Sélection des variables à inclure dans le modèle final

A l’aide de méthodes statistiques lorsqu’il reste encore trop de variables à l’étape précédente.

Nombre maximum de variables dans un modèle

Le nombre de sujets doit être assez grand pour que :
- la méthode du maximum de vraisemblance donne des résultats (estimations, intervalles de confiance et tests) non biaisés
- le processus de sélection des variables soit “stable”

La limite habituellement retenue porte sur le nombre d’ “évènements par variable” (EPV).

Le nombre d’événements est le nombre de malades (ou de non malades s’il est plus petit).

Règle générale pour modèle logistique : EPV ≥ 10

(pour modèle de Cox : EPV≥10, pour modèle linéaire : EPV≥2)

Cette règle est discutée :
- pour certains auteurs EPV ≥ 5 suffit dans la plupart des cas
- pour d'autres, EPV ne résume pas tout. La corrélation entre les variables, des coefficients élevés demandent que EPV augmente

Dans le cas de la GEU, il y a 574 cas. Il ne faut donc pas dépasser 57 variables dans le modèle (ce qui laisse de la marge ...).

NB : il existe des méthodes “exactes” d’estimation adaptées aux petits échantillons.
- elles résolvent (en partie au moins) les problèmes statistiques d’estimation
- elles ne résolvent pas l’instabilité des modèles et de la sélection des variables

Choix des variables “candidates”

Au moment du protocole

Le choix est en partie fait à ce moment (les variables doivent figurer dans le questionnaire ...)

- connaissance scientifique de la question
- hypothèses nouvelles à tester ou résultats anciens à confirmer
- revue bibliographique
- facteurs de risque reconnus de la maladie (souvent inclus de façon systématique dans les analyses ultérieures)

Au moment de l'analyse

◊ Commencer par une analyse univariée détaillée et complète pour :

- examiner les liens entre les facteurs de risque reconnus et la maladie
  ces facteurs seront de toute façon retenus dans les variables candidates, mais il est utile de vérifier si on retrouve les associations connues
- identifier les autres facteurs de risque
  => étudier de façon systématique les liens entre les différentes variables de l'enquête et la maladie
  retenir celles dont le degré de signification est ≤ 20 ou 25% (5% ne suffit pas)
- choisir le codage des variables
  regroupement de catégories des variables qualitatives nominales
  représentation des variables qualitatives ordinales et quantitatives
  (ces choix pourront être rediscutés dans la suite de l'analyse)

- nb :
  • la pertinence de cette étape est discutée en raison de la multiplication des tests et de l'imprécision générée sur le modèle final
  • elle doit être guidée par les connaissances biblio sur la question étudiée

◊ Si besoin, générer des variables synthétiques pour limiter les phénomènes de colinéarité

◊ Eliminer les variables qui ont trop de données manquantes ou d'erreurs de mesure, ou qui ont trop peu de variabilité

◊ Rechercher les interactions à prendre en compte

Ce choix pourra aussi être revu dans la suite de l'analyse après ajustement sur les autres variables

- rechercher les interactions au cas par cas (et quand il y a "une certaine logique") par analyse "bivariée"
- en pratique, on ne prend en compte que rarement les interactions d'ordre supérieur à 2
- ne retenir que le minimum de termes d'interaction car ils rendent l'interprétation du modèle difficile voire impossible

◊ Dans les étapes précédentes :

- ne pas conserver deux variables qui représentent à peu près la même chose (colinéarité)

- porter une attention particulière aux variables dont la variabilité est faible ou dont la répartition est déséquilibrée, notamment les variables pour lesquelles les tableaux croisés avec la maladie comportent des 0.

  Elles sont généralement à l'origine de problèmes numériques et d'instabilité dans les analyses ultérieures.

  Ces problèmes sont d'autant plus fréquents que la taille de l'échantillon est petite

  Exemple : intervention chirurgicale sur les trompes (ctub:chirurgie tubaire) et antécédent de GEU (ageu).

  Les GEU étaient presque toutes traitées chirurgicalement

```
. tab ctub ageu

    chir |       atcd geu
  tubaire |         0          1 |     Total
----------+----------------------+----------
        0 |     1,540          1 |     1,541
        1 |        70        114 |       184
----------+----------------------+----------
    Total |     1,610        115 |     1,725
```

```
. logistic ct ctub
Logistic regression                             Number of obs   =       1725
                                                LR chi2(1)      =     177.23
                                                Prob > chi2     =     0.0000
Log likelihood = -1008.6796                     Pseudo R2       =     0.0808
------------------------------------------------------------------------------
          ct | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        ctub |   9.301394   1.744355    11.89   0.000     6.440451    13.43321
       _cons |   .3870387   .0219819   -16.71   0.000     .3462664    .4326118
------------------------------------------------------------------------------
```

```
. logistic ct ageu
Logistic regression                             Number of obs   =       1725
                                                LR chi2(1)      =     148.57
                                                Prob > chi2     =     0.0000
Log likelihood = -1023.0071                     Pseudo R2       =     0.0677
------------------------------------------------------------------------------
          ct | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        ageu |   14.78487   4.064826     9.80   0.000      8.62572    25.34192
       _cons |   .4185022     .02287   -15.94   0.000     .3759949    .4658151
------------------------------------------------------------------------------
```

```
. logistic ct ctub ageu
Logistic regression                             Number of obs   =       1725
                                                LR chi2(2)      =     188.82
                                                Prob > chi2     =     0.0000
Log likelihood = -1002.8848                     Pseudo R2       =     0.0860
------------------------------------------------------------------------------
          ct | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        ctub |   4.823154   1.230297     6.17   0.000     2.925536    7.951641
        ageu |   3.390356   1.235502     3.35   0.001      1.65979     6.92528
       _cons |   .3866787   .0219646   -16.73   0.000     .3459387    .4322165
------------------------------------------------------------------------------
```

La forte variation des OR lorsque les deux variables sont incluses ensembles doit alerter sur l'existence possible d'un problème (ici le fort lien entre ctub et ageu).

En pratique :

impossible de séparer les rôles de ctub et ageu

il faut - en choisir une

ou - faire une variable combinée "ctub ou ageu" (qui est ici presque identique à ctub).

Sélection des variables à inclure dans le modèle final

Les différentes méthodes de sélection

- Utilisation des connaissances scientifiques existantes sur la question

- Examen de la variation de OR entre OR_{brut} et OR_a

- Méthodes de sélection pas à pas (stepwise)

- Autres méthodes
  - Examen de tous les modèles possibles
  - MFP (multivariable fractional polynomial, voir plus loin, variables quantitatives)
  - Approche bayésienne : prise en compte de la probabilité a priori que chaque modèle soit possible
  - DAG (Directed acyclic graphs) : moyen pratique et utile de présenter les connaissances scientifiques sur une question

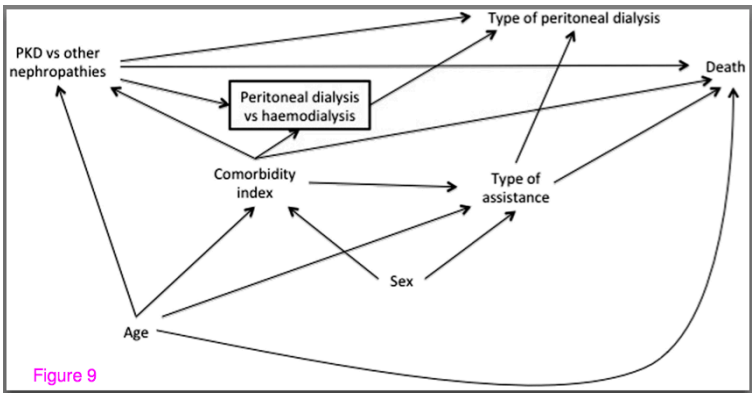


Figure 9

  - random forest : méthode utilisée pour repérer les variables influentes
  - shrinkage (contraction): méthode visant à ramener les coefficients vers 0 pour éviter une surestimation liée au processus de choix du modèle

Fréquence d'utilisation des différentes méthodes

Analyse de 4 revues majeures d'épidémiologie en 2008 (300 articles)[2] (American Journal of Epidemiology, Epidemiology, European Journal of Epidemiology and the International Journal of Epidemiology)

- Utilisation des connaissances a priori : 28%

(mais la moitié sans donner de références pour justifier les choix)

- Changement de l'estimation de OR : 15%

- Méthodes de sélection pas à pas (stepwise) : 20%

- Autres méthodes : 3%

- méthode non précisée : 35%

(souvent parce que la partie matériel et méthodes est trop imprécise)

Comparaison entre les méthodes

Pas de comparaison possible avec l'utilisation des connaissances a priori

pourtant c'est la méthode recommandée par la majorité des ouvrages de référence

Dans l'ensemble stepwise fait moins bien que des méthodes plus "modernes", mais l'écart est faible

Le pourcentage de variables incluses à tort ou non incluses à tort est de l'ordre de 10 à 15% avec stepwise et 5% avec des méthodes plus sophistiquées.

[2] Walter S, Tiemeier H. Variable selection: current practice in epidemiological studies. Eur J Epidemiol. 2009;24(12):733-6

Sélection selon la modification du coefficient de la variable d'intérêt

Particulièrement utile s'il y a une exposition d'intérêt E, les autres variables X_i étant des facteurs de confusion potentiels (ce n'est pas la situation la plus fréquente)

Le modèle complet s'écrit : Logit P = $\alpha + \beta E + \sum \beta_i X_i$

Principe général : on retire une (ou des) variables X_i du modèle à condition que les effets de confusion soient correctement pris en compte c'est-à-dire que le retrait des variables ne modifie pas (trop) la valeur du coefficient β de l'exposition d'intérêt principal.

Critère de décision :

- variation de β inférieure à une valeur seuil (par exemple moins de 10%), parce qu'alors le phénomène de confusion dû à X_i est négligeable et X_i peut être retirée.

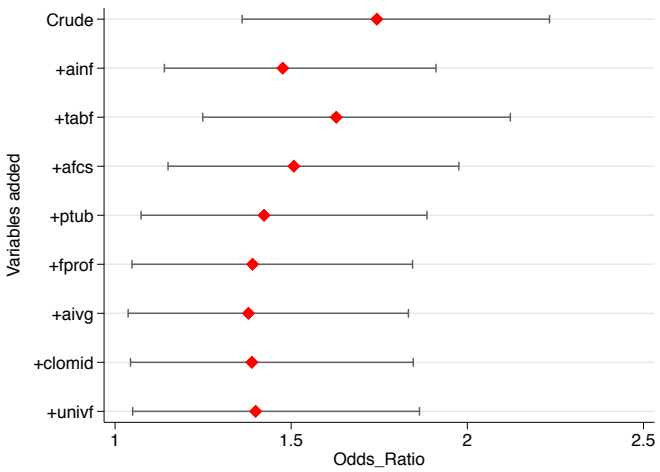
- test de la variation de β qui présente plusieurs problèmes
    - peu présent dans les logiciels
    - discutable dans son principe car l'hypothèse que X_i est facteur de confusion est peu intéressante (voire non pertinente) à tester. Ce qui compte c'est l'ampleur du phénomène de confusion.
    - Il faudrait aussi s'intéresser à la modification de l' IC

Avec Stata, la variation de β peut être étudiée avec le module chest.

Exemple : association entre GEU et age>30ans ajustée sur les facteurs de confusion potentiels

```
. logistic ct age30 tabf univf fprof afcs aivg  ainf clomid ptub

Logistic regression                             Number of obs     =      1,187
                                                LR chi2(9)        =     212.42
                                                Prob > chi2       =     0.0000
Log likelihood = -631.06752                     Pseudo R2         =     0.1441

------------------------------------------------------------------------------
          ct | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       age30 |   1.399087   .2048355     2.29   0.022     1.050082    1.864087
        tabf |   2.059245   .2983083     4.99   0.000     1.550245    2.735367
       univf |   .9197903    .158741    -0.48   0.628     .6558207    1.290008
       fprof |   1.392854   .2212318     2.09   0.037      1.02025    1.901538
        afcs |   1.538804   .2524503     2.63   0.009     1.115677    2.122405
        aivg |   1.096733   .2121842     0.48   0.633     .7506189    1.602442
        ainf |   2.473358   .4513877     4.96   0.000     1.729588    3.536968
      clomid |    1.39017   .4958898     0.92   0.356     .6909304    2.797057
        ptub |   4.074246    .643337     8.90   0.000     2.989783    5.552068
       _cons |   .1193269   .0212106   -11.96   0.000     .0842239     .16906
------------------------------------------------------------------------------
```

```
. chest age30,eform format(%5.3g)
Change-in-estimate
logistic regression.                    Outcome:  ct
number of obs = 1187                    Exposure: age30

------------------------------------------------------------------
Variables |
added     | exp(Coef.)  [95% Conf.  Interval]   Change, %
----------+-------------------------------------------------------
    Crude |       1.74        1.36        2.23
   +ainf  |       1.48        1.14        1.91         -15.3
   +tabf  |       1.63        1.25        2.12          10.3
   +afcs  |       1.51        1.15        1.98         -7.39
   +ptub  |       1.42        1.07        1.89         -5.62
  +fprof  |       1.39        1.05        1.84         -2.28
   +aivg  |       1.38        1.04        1.83         -.843
  +clomid |       1.39        1.04        1.85          .699
  +univf  |        1.4        1.05        1.86          .771
------------------------------------------------------------------
```




```
graph export chest.png
```

Examen de tous les modèles possibles

- peu fréquemment programmé dans les logiciels classiques

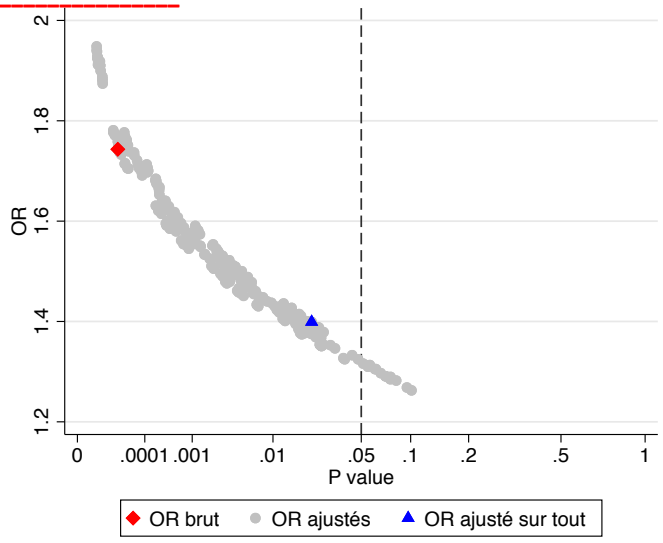
p variables => 2^p modèles (p=20 donne 1 048 576 modèles ...)

Stata : module confall

```
confall age30, eform table
```

Var.adj.	Exp(Coef.)	[95% Conf.	Interval]
Null	1.743234	1.360727	2.233264
tabf	1.911759	1.482228	2.465761
univf	1.777161	1.383996	2.282015
.....			
tabf fprof afcs aivg ainf clomid ptub	1.388378	1.043871	1.846581
univf fprof afcs aivg ainf clomid ptub	1.289524	.9733552	1.708393
tabf univf fprof afcs aivg ainf clomid p	1.399087	1.050083	1.864085

```
256 sets of confounders
Outcome variable: ct      Exposure: age30

graph export confall.png,replace
```



- sélection des modèles basée sur la minimisation d'un critère d'information de la forme IC = -2 ln(V) + a dim(M)

dim(M) = nombre de paramètres ; a = constante de "pénalité"

critère d'Aikaïke (AIC) : a = 2

critère de Bayes (BIC) : a = ln(n)

Procédures pas à pas

Particulièrement utile lorsqu'on s'intéresse à l'ensemble des facteurs de risque de la maladie (c'est la situation la plus fréquente, mais il ne s'agit pas alors seulement de s'occuper d'une question de confusion).

Le modèle complet s'écrit : Logit P = $\alpha + \sum \beta_i X_i$

Principe général : on retire une variable X_i si son coefficient est non significativement différent de 0 (ou avec un degré de signification supérieur à une limite fixée, par exemple 0,10).

On retire donc successivement les variables qui ne sont pas facteur de risque de la maladie.

C'est ce que proposent les procédures pas à pas ("stepwise") de la plupart des logiciels.

• Procédure descendante

1. On part du modèle "complet" contenant toutes les variables X_i retenues.

2. On retire successivement la variable X_k dont le degré de signification p du lien avec la maladie (ajusté sur les autres variables) est le plus grand, tant que $p \geq \alpha$ (constante fixée).

procédure descendante "mixte"
2bis. A partir du moment où 2 variables ont été retirées :
- si la variable exclue la plus significative vérifie $p < \alpha'$, on la réintègre
- si la variable inclue la moins significative vérifie $p \geq \alpha$, on la retire

• Procédure ascendante

Même principe ("à l'envers") en partant du modèle vide

• Possibilité de “forcer des variables”

• Le choix des constantes α et α’ est important. Plus elles sont grandes, plus le modèle final comprendra de variables. On prend souvent α et α’ de l’ordre de 0,10 à 0,15.

• Les modèles obtenus ne sont pas toujours les mêmes avec les procédures descendantes et ascendantes. En général, moins de variables sélectionnées avec procédure ascendante.

• Procédure descendante a l’avantage de partir du modèle complet qui est une meilleure référence que le modèle vide.

Avantages et inconvénients des procédures pas à pas

Avantages

- utilisation facile
- examen systématique et standardisé de l'ensemble des variables
- pas de réelle alternative en l'absence de connaissances scientifiques fortes des mécanismes gouvernant la question étudiée (ce qui est fréquent)
- résultats obtenus sont satisfaisants

Inconvénients

- fondées sur des critères uniquement statistiques
- absence de considérations épidémiologiques (méthodes "pas politiquement correctes")
- violation de conditions statistiques (risque d'erreur, distribution)
- sous estimation des IC des coefficients qui ne tient pas compte de l'incertitude ajoutée par le choix du modèle
- biais possibles dans la sélection des variables
  - la sélection est basée sur les coefficients estimés des variables (pas sur leur vraie valeur)
  - une variable est plus vraisemblablement incluse si son coefficient est surestimé

    - les "prédicteurs forts" (facteurs fortement liés à la maladie) sont presque toujours inclus que leur coefficient soit sous-estimé ou pas

    - les "prédicteurs faibles" ne sont inclus que si leur coefficient est sur-estimé

Exemples de procédure pas à pas

Modèle complet

```
. logistic ct ctub i.agec tabf univf aacc afcs aivg ainf clomid

Logistic regression                             Number of obs     =      1,620
                                                LR chi2(11)       =     302.66
                                                Prob > chi2       =     0.0000
Log likelihood = -863.89877                     Pseudo R2         =     0.1491

------------------------------------------------------------------------------
          ct | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        ctub |   7.013603   1.494838     9.14   0.000     4.618715    10.65029
             |
        agec |
          1  |   .9546271   .1684659    -0.26   0.792     .6754911    1.349112
          2  |   1.129204   .2221671     0.62   0.537     .7678939    1.660517
          3  |   1.511209   .3402527     1.83   0.067      .972018    2.349497
             |
        tabf |   2.870179   .3520093     8.60   0.000     2.256912    3.650089
       univf |   1.119296   .1614581     0.78   0.435     .8436442    1.485014
        aacc |   1.164726   .1544326     1.15   0.250     .8981763    1.510378
        afcs |   1.422089   .2010925     2.49   0.013     1.077857    1.876256
        aivg |   1.153113   .1871852     0.88   0.380     .8388724    1.585067
        ainf |   2.322488   .3787415     5.17   0.000     1.687109    3.197156
      clomid |   1.372395   .4458498     0.97   0.330     .7260211    2.594232
       _cons |   .1534527   .0252264   -11.40   0.000     .1111844     .21179
------------------------------------------------------------------------------
```

• Procédure descendante

(dans Stata, on ne peut pas utiliser "factor variable", il faut utiliser "l'ancienne" commande xi)

```
. xi : stepwise,pr(.15) pe(.1) : logistic ct ctub (i.agec) tabf univf aacc afcs aivg ainf clomid
i.agec            _Iagec_0-3          (naturally coded; _Iagec_0 omitted)
                  begin with full model
p = 0.4346 >= 0.1500  removing univf
p = 0.3838 >= 0.1500  removing aivg
p = 0.3023 >= 0.1500  removing clomid
p = 0.3393 >= 0.1500  removing aacc

Logistic regression                               Number of obs     =      1,620
                                                  LR chi2(7)        =     299.33
                                                  Prob > chi2       =     0.0000
Log likelihood = -865.56246                       Pseudo R2         =     0.1474

------------------------------------------------------------------------------
          ct | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        ctub |   6.869142   1.459751     9.07   0.000     4.529132    10.41814
    _Iagec_1 |   1.017915   .1745255     0.10   0.918      .727394     1.42447
    _Iagec_2 |   1.259479   .2299469     1.26   0.206     .8806118    1.801347
    _Iagec_3 |   1.721392   .3596479     2.60   0.009     1.142983    2.592508
        tabf |   2.870837   .3463653     8.74   0.000     2.266269    3.636685
        ainf |    2.35047   .3631412     5.53   0.000     1.736386    3.181728
        afcs |   1.437321   .2023726     2.58   0.010     1.090702    1.894095
       _cons |   .1643953   .0258905   -11.46   0.000     .1207353    .2238436
------------------------------------------------------------------------------
```

• Procédure ascendante

```
. xi : stepwise,pr(.15) pe(.1) forw : logistic ct ctub (i.agec) tabf univf aacc afcs aivg ainf clomid
i.agec            _Iagec_0-3          (naturally coded; _Iagec_0 omitted)
                  begin with empty model
p = 0.0000 <  0.1000  adding   ctub
p = 0.0000 <  0.1000  adding   tabf
p = 0.0000 <  0.1000  adding   ainf
p = 0.0012 <  0.1000  adding   afcs
p = 0.0180 <  0.1000  adding   _Iagec_1 _Iagec_2 _Iagec_3

Logistic regression                               Number of obs     =      1,620
                                                  LR chi2(7)        =     299.33
                                                  Prob > chi2       =     0.0000
Log likelihood = -865.56246                       Pseudo R2         =     0.1474

------------------------------------------------------------------------------
          ct | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        ctub |   6.869142   1.459751     9.07   0.000     4.529132    10.41814
        tabf |   2.870837   .3463653     8.74   0.000     2.266269    3.636685
        ainf |    2.35047   .3631412     5.53   0.000     1.736386    3.181728
        afcs |   1.437321   .2023726     2.58   0.010     1.090702    1.894095
    _Iagec_1 |   1.017915   .1745255     0.10   0.918      .727394     1.42447
    _Iagec_2 |   1.259479   .2299469     1.26   0.206     .8806118    1.801347
    _Iagec_3 |   1.721392   .3596479     2.60   0.009     1.142983    2.592508
       _cons |   .1643953   .0258905   -11.46   0.000     .1207353    .2238436
------------------------------------------------------------------------------
```

En résumé

• Si
- l’étape l'étape du choix des variables candidates a été faite soigneusement (et cela peut prendre du temps ...)
- on a une bonne connaissance du problème et/ou une expérience suffisante
- les données s’y prêtent ...,

alors, le nombre de variables retenues est limité, elles sont toutes pertinentes et on les inclut donc toutes dans le modèle final.

Cette procédure est généralement considérée comme la méthode de référence (la "meilleure").

• Sinon (et c’est la situation la plus fréquente en raison du manque de connaissances solides a priori)

on peut utiliser des procédures statistiques s’appuyant sur les données observées et des règles “automatiques” de sélection des variables.

- modification du coefficient de la variable d'intérêt (ne pas oublier de s'intéresser aussi à l'IC du coefficient)
- procédures pas à pas (les plus courantes)

Les procédures pas à pas peuvent être associées à d'autres méthodes :

◊ "Garde-fous" statistiques
– faire un test global préalable avec l'ensemble des variables candidates (si NS, on s'arrête)
– recourir à des méthodes de shrinkage pour limiter la surestimation des coefficients
– vérifier la stabilité des résultats (sur une autre partie de l'échantillon ou par boostrap)

◊ Introduction des connaissances a priori
– forcer certaines variables
– regrouper les variables en groupes homogènes
– examiner soigneusement chaque étape de la sélection

◊ Groupes homogènes de variables
Si les variables sont nombreuses, on peut les réunir en groupes homogènes qu'on commence par analyser séparément

- cela revient à introduire de la connaissance du problème (matter-knowledge) dans les procédures automatiques
- cela permet aussi d'ordonner sa réflexion et la présentation des résultats
- les procédures précédentes sont appliquées à chaque groupe

- certaines variables peuvent intervenir dans plusieurs sous-analyses : par exemple, on peut ajuster systématiquement l'analyse des antécédents sur l'âge

- les variables retenues dans chaque groupe sont ensuite incluses dans le modèle final

Exemple : Y = Grossesse extra-utérine

Variables X_i

Carac. socio-prof.	Exposition potentielle aux MST
Age	Infection gynécologique
Tabac	Sérologie positive pour C. Trachomatis
Niveau d'études	Age aux premiers rapports sexuels
	Nombre de partenaires
Antécédents chir. et obstét.	
Appendicectomie	**Fertilité**
Chirurgie tubaire	Antécédent de contraception
Antécédent de GEU	Dernier mode de contraception
Antécédent de FCS	Antécédent d'infécondité
Antécédent d'accouchement	Grossesse induite
Antécédent d'IVG	Délai depuis la dernière grossesse

◊ Examiner soigneusement chaque étape de la sélection

– Pour s'assurer qu'on est capable d'interpréter les opérations réalisées (codage des variables, termes d'interaction, retrait ou ajout de variable ...)

– Chaque fois que c'est nécessaire, il ne faut pas hésiter à revenir à l'analyse univariée pour bien comprendre les résultats des analyses multivariées

– Sur le modèle "final", il est possible de revoir (remettre en cause ?) le codage des variables et/ou la présence d'interactions

◊ Ne pas oublier que

– Le choix des variables ne se pose vraiment que pour celles qui ont un lien faible avec la maladie

– Quand l'échantillon est petit, les procédures de sélection sont toutes instables

Régression logistique multinomiale

Y = 0, 1, …, k-1 (variable nominale à k classes)

Le modèle logistique multinomial s'écrit :

$$\text{Ln}\left(\frac{P(Y=j\,|\,X)}{P(Y=0\,|\,X)}\right)=\alpha_j+g_j(x)=\alpha_j+\sum\beta_{ji}x_i$$

Cela revient en fait à prendre la catégorie Y = 0 comme référence et à faire k-1 régressions logistiques dichotomiques.

Comme $\sum P(Y=j)=1$, le modèle peut s'écrire :

$$P(Y=0|X)=\frac{1}{1+\exp\{\alpha_1+g_1(x)\}+\ldots+\exp\{\alpha_{k-1}+g_{k-1}(x)\}}$$

$$=\frac{1}{1+\sum_{l=1}^{k-1}\exp\{\alpha_l+g_l(x)\}}$$

$$P(Y=j|X)=\frac{\exp\{\alpha_j+g_j(x)\}}{1+\sum_{l=1}^{k-1}\exp\{\alpha_l+g_l(x)\}}\quad\text{pour j = 1, … , k-1}$$

L'écriture peut se synthétiser en :

$$P(Y=j|X)=\frac{\exp\{\alpha_j+g_j(x)\}}{\sum_{l=0}^{k-1}\exp\{\alpha_l+g_l(x)\}}$$

avec $\alpha_o=\beta_{oi}=0$

et j = 0, …, k-1

Le modèle logistique multinomial est utilisable dans des enquêtes de type cas-témoins.

Soit f_j la fraction de sondage de la catégorie Y = j.

On a : $P(Y = j|X) = f_j \frac{\exp(\alpha_j + g_j(x)}{\sum_{l=0}^{k-1} \exp\{\alpha_l + g_l(x)\}}$ avec $\alpha_o = \beta_{oi} = 0$

D'où :

$$\frac{P(Y = j|X)}{P(Y = 0|X)} = \frac{f_j \exp\{\alpha_j + g_j(x)\}}{f_o \exp\{\alpha_o + g_o(x)\}} = \frac{f_j}{f_o} \exp\{\alpha_j + g_j(x)\}$$

et donc :

$$\text{Ln}\left(\frac{P(Y = j|X)}{P(Y = 0|X)}\right) = \text{Ln}\left(\frac{f_j}{f_o}\right) + \{\alpha_j + g_j(x)\} = \alpha'_j + \sum \beta_{ji} x_i$$

Exemple

Y = issue de grossesse (après FIV)

0 : à terme (référence)
1 : prématurée (rupture des membranes)
2 : prématurée (spontanée)
3 : prématurée (provoquée)

1. Coefficients

```
. mlogit cprema  age35

Multinomial logistic regression                 Number of obs   =       7601
                                                LR chi2(3)      =       8.70
                                                Prob > chi2     =     0.0336
Log likelihood = -3025.1826                     Pseudo R2       =     0.0014

------------------------------------------------------------------------------
      cprema |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
0            |  (base outcome)
-------------+----------------------------------------------------------------
1            |
       age35 |   .1761545   .1993624     0.88   0.377    -.2145885    .5668976
       _cons |  -4.211474   .1221634   -34.47   0.000     -4.45091   -3.972038
-------------+----------------------------------------------------------------
2            |
       age35 |    .084688   .1188679     0.71   0.476    -.1482887    .3176648
       _cons |  -3.088647   .0707297   -43.67   0.000    -3.227275   -2.950019
-------------+----------------------------------------------------------------
3            |
       age35 |     .35869   .1277462     2.81   0.005     .1083121    .6090679
       _cons |  -3.407101   .0824436   -41.33   0.000    -3.568688   -3.245514
------------------------------------------------------------------------------
```

2. Odds ratios

```
. mlogit cprema  age35,rrr

Multinomial logistic regression                 Number of obs   =       7601
                                                LR chi2(3)      =       8.70
                                                Prob > chi2     =     0.0336
Log likelihood = -3025.1826                     Pseudo R2       =     0.0014

------------------------------------------------------------------------------
      cprema |        RRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
0            |  (base outcome)
-------------+----------------------------------------------------------------
1            |
       age35 |   1.192622    .237764     0.88   0.377     .8068734     1.76279
       _cons |   .0148245    .001811   -34.47   0.000     .0116679     .018835
-------------+----------------------------------------------------------------
2            |
       age35 |   1.088377   .1293731     0.71   0.476     .8621822    1.373916
       _cons |   .0455635   .0032227   -43.67   0.000     .0396654    .0523387
-------------+----------------------------------------------------------------
3            |
       age35 |   1.431453   .1828626     2.81   0.005     1.114396    1.838717
       _cons |   .0331371   .0027319   -41.33   0.000     .0281928    .0389485
------------------------------------------------------------------------------
```

On retrouve les mêmes résultats avec des régressions logistiques “simples”

- A terme / rupture prématurée des membranes

```
. logistic rupmemb age35
Logistic regression                               Number of obs   =       7015
                                                  LR chi2(1)      =       0.77
                                                  Prob > chi2     =     0.3805
Log likelihood = -561.69019                       Pseudo R2       =     0.0007

------------------------------------------------------------------------------
     rupmemb | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       age35 |   1.192622    .237764     0.88   0.377     .8068734    1.76279
       _cons |   .0148245    .001811   -34.47   0.000     .0116679    .018835
------------------------------------------------------------------------------
```

- A terme / prématurité spontanée

```
. logistic spont age35
Logistic regression                               Number of obs   =       7230
                                                  LR chi2(1)      =       0.50
                                                  Prob > chi2     =     0.4779
Log likelihood = -1322.4787                       Pseudo R2       =     0.0002

------------------------------------------------------------------------------
       spont | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       age35 |   1.088378   .1293732     0.71   0.476      .862183    1.373917
       _cons |   .0455636   .0032227   -43.67   0.000     .0396654    .0523387
------------------------------------------------------------------------------
```

- A terme / prématurité provoquée

```
. logistic prov age35
Logistic regression                               Number of obs   =       7168
                                                  LR chi2(1)      =       7.71
                                                  Prob > chi2     =     0.0055
Log likelihood = -1120.2642                       Pseudo R2       =     0.0034

------------------------------------------------------------------------------
        prov | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       age35 |   1.431453   .1828626     2.81   0.005     1.114396    1.838717
       _cons |   .0331371   .0027319   -41.33   0.000     .0281928    .0389485
------------------------------------------------------------------------------
```

- Régression multinomiale

```
. mlogit cprema  age35,rrr
Multinomial logistic regression                   Number of obs   =       7601
                                                  LR chi2(3)      =       8.70
                                                  Prob > chi2     =     0.0336
Log likelihood = -3025.1826                       Pseudo R2       =     0.0014

------------------------------------------------------------------------------
      cprema |        RRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
0            |  (base outcome)
-------------+----------------------------------------------------------------
1            |
       age35 |   1.192622    .237764     0.88   0.377     .8068734    1.76279
       _cons |   .0148245    .001811   -34.47   0.000     .0116679    .018835
-------------+----------------------------------------------------------------
2            |
       age35 |   1.088377   .1293731     0.71   0.476     .8621822    1.373916
       _cons |   .0455635   .0032227   -43.67   0.000     .0396654    .0523387
-------------+----------------------------------------------------------------
3            |
       age35 |   1.431453   .1828626     2.81   0.005     1.114396    1.838717
       _cons |   .0331371   .0027319   -41.33   0.000     .0281928    .0389485
------------------------------------------------------------------------------
```

Dans le cas de plusieurs variables X, les résultats sont différents, mais légèrement seulement

(voir Begg CB, Gray R : Calculation of polychotomous logistic regression parameters using individualized regressions. Biometrika 1984; 71 : 11-18)

• Régression multinomiale

```
. mlogit cprema age35 hta,rrr

Multinomial logistic regression                 Number of obs   =       7601
                                                LR chi2(6)      =      71.91
                                                Prob > chi2     =     0.0000
Log likelihood = -2993.5779                     Pseudo R2       =     0.0119

------------------------------------------------------------------------------
      cprema |        RRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
0            |  (base outcome)
-------------+----------------------------------------------------------------
1            |
       age35 |   1.193357   .2379296     0.89   0.375      .8073451    1.763931
         hta |   .8710343   .5135622    -0.23   0.815      .2742625     2.76633
       _cons |   .0148815   .0018322   -34.17   0.000      .0116908    .0189429
-------------+----------------------------------------------------------------
2            |
       age35 |   1.085837   .1291046     0.69   0.489      .8601182    1.370792
         hta |   1.499127    .409747     1.48   0.139      .8773738    2.561486
       _cons |   .0448984   .0032195   -43.28   0.000      .0390116    .0516735
-------------+----------------------------------------------------------------
3            |
       age35 |   1.404867   .1808795     2.64   0.008      1.091544    1.808128
         hta |   5.511777   1.022349     9.20   0.000      3.831836    7.928232
       _cons |   .0292238   .0025245   -40.90   0.000      .0246721    .0346152
------------------------------------------------------------------------------
```

• A terme / rupture prématurée des membranes

```
. logistic rupmemb age35 hta

Logistic regression                             Number of obs   =       7015
                                                LR chi2(2)      =       0.83
                                                Prob > chi2     =     0.6614
Log likelihood = -561.66125                     Pseudo R2       =     0.0007

------------------------------------------------------------------------------
     rupmemb | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       age35 |   1.193451   .2379533     0.89   0.375      .8074022    1.764084
         hta |     .87036     .51318    -0.24   0.814      .2740407    2.764285
       _cons |   .0148814    .001832   -34.18   0.000      .0116911    .0189423
------------------------------------------------------------------------------
```

- A terme / prématurité spontanée

```
. logistic spont age35 hta

Logistic regression                               Number of obs   =       7230
                                                  LR chi2(2)      =       2.47
                                                  Prob > chi2     =     0.2909
Log likelihood = -1321.4958                       Pseudo R2       =     0.0009

------------------------------------------------------------------------------
       spont | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       age35 |   1.085099   .1290305     0.69   0.492     .8595124    1.369894
         hta |   1.497894   .4094452     1.48   0.139     .8766122    2.559497
       _cons |   .0449104    .003219   -43.29   0.000     .0390243    .0516842
------------------------------------------------------------------------------
```

- A terme / prématurité provoquée

```
. logistic prov age35 hta

Logistic regression                               Number of obs   =       7168
                                                  LR chi2(2)      =      69.83
                                                  Prob > chi2     =     0.0000
Log likelihood = -1089.2052                       Pseudo R2       =     0.0311

------------------------------------------------------------------------------
        prov | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       age35 |   1.407678   .1813277     2.65   0.008     1.093596    1.811963
         hta |   5.518765   1.023717     9.21   0.000     3.836596    7.938486
       _cons |   .0291983   .0025242   -40.88   0.000     .0246474    .0345895
------------------------------------------------------------------------------
```

Intérêt de la régression logistique multinomiale : comparaison des OR selon les valeurs de Y

```
. mlogit cprema age35 hta,rrr

Multinomial logistic regression                 Number of obs   =       7601
                                                LR chi2(6)      =      71.91
                                                Prob > chi2     =     0.0000
Log likelihood = -2993.5779                     Pseudo R2       =     0.0119

------------------------------------------------------------------------------
      cprema |        RRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
0            |  (base outcome)
-------------+----------------------------------------------------------------
1            |
       age35 |   1.193357   .2379296     0.89   0.375      .8073451    1.763931
         hta |   .8710343   .5135622    -0.23   0.815      .2742625     2.76633
       _cons |   .0148815   .0018322   -34.17   0.000      .0116908    .0189429
-------------+----------------------------------------------------------------
2            |
       age35 |   1.085837   .1291046     0.69   0.489      .8601182    1.370792
         hta |   1.499127    .409747     1.48   0.139      .8773738    2.561486
       _cons |   .0448984   .0032195   -43.28   0.000      .0390116    .0516735
-------------+----------------------------------------------------------------
3            |
       age35 |   1.404867   .1808795     2.64   0.008      1.091544    1.808128
         hta |   5.511777   1.022349     9.20   0.000      3.831836    7.928232
       _cons |   .0292238   .0025245   -40.90   0.000      .0246721    .0346152
------------------------------------------------------------------------------
```

	Rupt préma	Préma spont.	Préma prov.
Age ≥ 35	1,2 [0,81 ; 1,8]	1,1 [0,86 ; 1,4]	1,4 [1,1 ; 1,8]
HTA	0,87 [0,27 ; 2,8]	1,5 [0,88 ; 2,6]	5,5 [3,8 ; 7,9]

```
. test [1=2] : age35

 ( 1)  [1]age35 - [2]age35 = 0.0

           chi2(  1) =    0.17
         Prob > chi2 =    0.6805
```

```
. test [1=3] : age35

 ( 1)  [1]age35 - [3]age35 = 0.0

           chi2(  1) =    0.48
         Prob > chi2 =    0.4868
```

```
. test ([1=2] : age35) ([1=3] : age35)

 ( 1)  [1]age35 - [2]age35 = 0
 ( 2)  [1]age35 - [3]age35 = 0

           chi2(  2) =    2.27
         Prob > chi2 =    0.3210
```

```
. test [1=2] : hta

 ( 1)  [1]hta - [2]hta = 0.0

           chi2(  1) =    0.71
         Prob > chi2 =    0.3980
```

```
. test [1=3] : hta

 ( 1)  [1]hta - [3]hta = 0.0

           chi2(  1) =    9.14
         Prob > chi2 =    0.0025
```

```
. test ([1=2] : hta) ([1=3] : hta)

 ( 1)  [1]hta - [2]hta = 0
 ( 2)  [1]hta - [3]hta = 0

           chi2(  2) =   22.88
         Prob > chi2 =    0.0000
```