# Modèles linéaires généralisés

Valérie Monbet

IRMAR, Université de Rennes 1

# Références, supports

Ce cours est essentiellement une compilation des cours de G. Rodriguez (Princeton University) et A. Lavenu (Université de Rennes 1).

- Hardin, J. and Hilbe, J. (2012). Generalized Linear Models and Extensions, 3rd Edition. College Station, Texas : Stata Press. Un livre avec des exemples et des applications incluant des analyses avec Stata.
- Notes de cours de G. Rodriguez et exemples de codes R : `http://data.princeton.edu/wws509/`
- Notes de cours de L. Rouvière `http://perso.univ-rennes2.fr/system/files/users/rouviere_l/poly_logi`
- Notes de cours de F. Bertrand `http://www-irma.u-strasbg.fr/~fbertran/enseignement/Ecole_Doctorale_SVS_Automne_2008/ED_RegLog.pdf`
- Hosmer, D.W. and Lemeshow, S. (2013). Applied Logistic Regression, 3rd Edition. New York : John Wiley and Sons. Une discussion détaillée sur le modèle logistique avec des applications.
- McCullagh, P. and Nelder, J.A. (1989). Generalized Linear Models, 2nd Edition. London : Chapman and Hall. La "bible" des modèles linéaires généralisés. Très intéressant, mais plutôt destinée à des étudiants avancés.

# Outline

# Outline

# Modèles linéaires pour les données continues

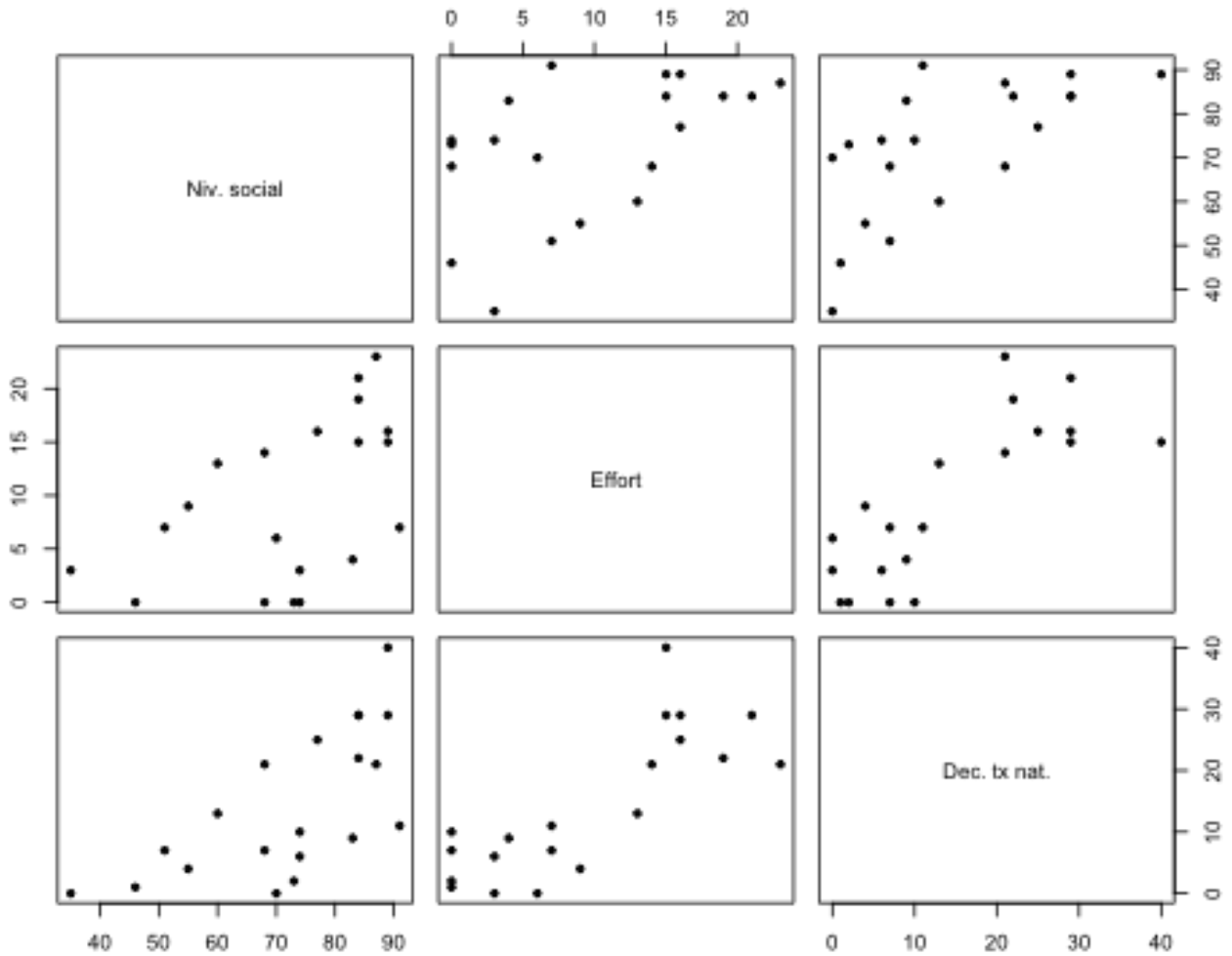
- Les modèles linéaires sont utilisés pour étudier comment une variable continue dépend d'un ou plusieurs prédicteurs ou variables explicatives. Les prédicteurs peuvent être quantitatifs ou qualitatifs.
- Exemple : données des efforts des plannings familiaux en Amérique du Sud (Mauldin and Berelson, 1978)
  Le niveau social et les efforts des plannings familiaux sont mesurés par une combinaison d'indices. Plus l'indice est élevé plus le niveau social (resp. l'effort) est élevé.

| | Niv. social | Effort | Déclin du tx nat. |
|---|---|---|---|
| Bolivia | 46 | 0 | 1 |
| Brazil | 74 | 0 | 10 |
| Chile | 89 | 16 | 29 |
| Colombia | 77 | 16 | 25 |
| CostaRica | 84 | 21 | 29 |
| Cuba | 89 | 15 | 40 |
| Dominican Rep | 68 | 14 | 21 |
| Ecuador | 70 | 6 | 0 |
| El Salvador | 60 | 13 | 13 |
| ⋮ | ⋮ | ⋮ | ⋮ |

- Dans cet exemple on cherche à comprendre comment le niveau social et les efforts de planification influent sur le taux de natalité.
- Notons $Y$ le taux de natalité et $X_1$ et $X_2$ respectivement le niveau social et l'effort de planification.
- On dispose de $n = 20$ observations $\{(x_{i1}, x_{i2}, y_i)\}$, $i = 1, \cdots, n$.

# Exemple (suite)

- On observe que le déclin du taux de natalité augmente avec le niveau social et avec l'effort de planification.



# Exemple (suite) et vocabulaire

- Si on suppose que $Y_i$ est une variable gaussienne de moyenne $\mu_i$ et de variance $\sigma^2$, on pourra écrire le modèle

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} = \beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}$$

▸ Calcul matriciel

- La densité de probabilité de $Y_i$ est donnée par

$$\varphi(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(y-\mu_i)^2}{\sigma^2}\right)$$

- Une manière plus standard d'écrire ce modèle linéaire est

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon = \beta_0 + \mathbf{X}^T \boldsymbol{\beta} + \epsilon$$

avec $\epsilon$ une variable aléatoire de loi de Gauss centrée et de variable $\sigma^2$.

- $\mathbf{X}$ est la matrice de design et $\mathbf{X}^T\boldsymbol{\beta}$ est le prédicteur linéaire.

# Inférence

- L'estimation des paramètres du modèle $\beta_0$, $\boldsymbol{\beta}$, $\sigma$ se fait par minimisation d'un critère des moindres carrées (OLS)

$$(\beta_0^*, \boldsymbol{\beta}^*) = arg \min_{(\beta_0, \boldsymbol{\beta})} \sum_{i=1}^{n} (y_i - \beta_0 - \mathbf{x}_i^T \boldsymbol{\beta})^2$$

$$\sigma^* = Var(y_i - \beta_0^* - \mathbf{x}_i^T \boldsymbol{\beta}^*)$$

  ou par maximum de vraisemblance (ML).
- Dans le cas du modèle linéaire, ces deux estimateurs sont identiques et

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T y$$

- On dispose ensuite de tests (test de Fisher, test de Wald, etc) pour valider et interpréter le modèle.

# Régression linéaire simple

```
> m1<- lm(DecTxNat ~ NivSocial,data=fpe)
> summary(m1)

Call:
lm(formula = DecTxNat ~ NivSocial, data = fpe)

Residuals:
    Min      1Q  Median      3Q     Max
-13.239  -6.260   0.787   6.678  17.162

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -22.1254     9.6416  -2.295  0.03398 *
NivSocial     0.5052     0.1308   3.863  0.00114 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.973 on 18 degrees of freedom
Multiple R-squared:  0.4532,Adjusted R-squared:  0.4228
F-statistic: 14.92 on 1 and 18 DF,  p-value: 0.001141
```

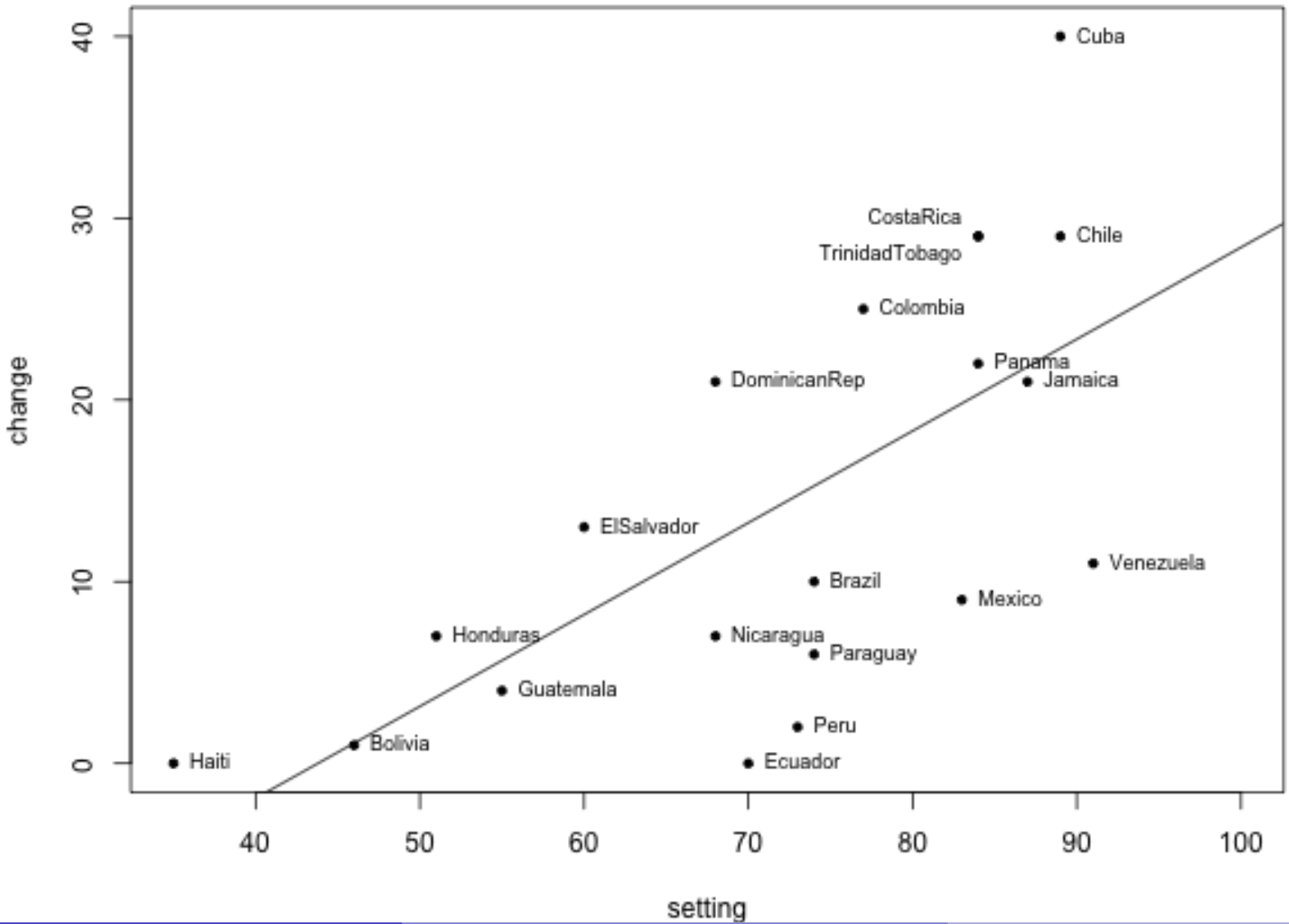
# Régression linéaire simple

- Le test `t` permet de tester l'hypothèse $H_0 : \beta_j = 0$ pour chaque variable $j$.
- Le test de Fisher permet de tester plusieurs paramètres simultanément. C'est notamment important quand on travaille avec des variables catégorielles recodées en variables binaires.
- Ici les deux tests conduisent à la même conclusion : la variable explicative a un effet significatif.

```
> anova(m1)
Analysis of Variance Table

Response: DecTxNat
           Df Sum Sq Mean Sq F value   Pr(>F)
NivSocial   1 1201.1 1201.08  14.919 0.001141 **
Residuals  18 1449.1   80.51
---
```

# Régression linéaire simple



# Régression linéaire multiple

```
> m2 <- lm(DecTxNat ~ NivSocial+Effort,data=fpe)
> summary(m2)

Call:
lm(formula = DecTxNat ~ NivSocial + Effort, data = fpe)

Residuals:
     Min       1Q   Median       3Q      Max
-10.3475  -3.6426   0.6384   3.2250  15.8530

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -14.4511     7.0938  -2.037 0.057516 .
NivSocial     0.2706     0.1079   2.507 0.022629 *
Effort        0.9677     0.2250   4.301 0.000484 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.389 on 17 degrees of freedom
Multiple R-squared:  0.7381,Adjusted R-squared:  0.7073
F-statistic: 23.96 on 2 and 17 DF,  p-value: 1.132e-05
```

# Régression linéaire simple

```
> anova(m2)
Analysis of Variance Table

Response: DecTxNat
           Df  Sum Sq Mean Sq F value    Pr(>F)
NivSocial   1 1201.08 1201.08  29.421 4.557e-05 ***
Effort      1  755.12  755.12  18.497 0.0004841 ***
Residuals 17  694.01   40.82
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

# Cas des prédicteurs qualitatifs : analyse de la variance

- Dans certains cas, les prédicteurs sont tous qualitatifs. Par exemple, si le niveau social est donné par une variable discrète à 3 modalités (inf à 70, entre 70 et 80, sup à 80).
- Les notations sont un peu différentes de celles du modèle de régression.
  On note $K$ le nombre de niveaux dans le facteur et $n_k$ le nombre d'observations dans le niveau $k$. $y_{ki}$ est la réponse de l'individu $i$ au niveau $k$.
- Dans l'exemple, $y_{ki}$ sera donc le déclin du taux de natalité du pays $i$ pour le niveau social $k$. $k = 1, \cdots, 3$, $K = 3$, ...
- On écrit alors le modèle de l'analyse de la variance à un facteur : $Y_{ki} \sim \mathcal{N}(\mu_{ki}, \sigma^2)$ avec

$$\mu_{ki} = \mu + \alpha_k$$

où $\mu$ est un effet moyen (commun à taux les niveaux du facteur) et $\alpha_k$ représente l'effet spécifique du niveau $k$.

# ANOVA, un facteur

```
> m1g <- lm(DecTxNat ~ NivSocial.g)
>  summary(m1g)

Call:
lm(formula = DecTxNat ~ NivSocial.g)

Residuals:
    Min      1Q  Median      3Q     Max
-14.750  -6.579  -1.161   5.250  16.400

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)          7.571      3.498   2.164  0.04497 *
NivSocial.gMedium    1.029      5.420   0.190  0.85173
NivSocial.gHigh     16.179      4.790   3.377  0.00358 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.256 on 17 degrees of freedom
Multiple R-squared:  0.4505,Adjusted R-squared:  0.3858
F-statistic: 6.967 on 2 and 17 DF,  p-value: 0.006167
```

# ANOVA, un facteur

- Dans ce cas, on voit bien la différence entre le test `t` et le test de Fisher.
- Le test de Fisher a plus de sens car il considère la variable explicative dans son ensemble et non modalité par modalité.

```
> anova(m1g)
Analysis of Variance Table

Response: DecTxNat
             Df Sum Sq Mean Sq F value   Pr(>F)
NivSocial.g   2 1193.8  596.89  6.9672 0.006167 **
Residuals    17 1456.4   85.67
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Ce petit exemple permet de rappeler les différentes étapes de la modélisation linéaire
  1. Analyse descriptive, visualisation des données
  2. Choix du modèle : transformation de variables, variable à introduire dans le modèle
  3. Inférence : estimation des paramètres du modèle, tests, intervalles de confiance
  4. Sélection des variables explicatives, choix de leur paramétrisation (continues/discrétisées)
  5. Interprétation du modèle
  6. Validation et comparaisons de modèle(s)

- On boucle souvent sur les étapes 2 à 6 jusqu'à obtenir un bon modèle.

# Outline

# Introduction d'une fonction de lien

- Quand la variable à prédire est discrète, elle ne suit plus une loi de Gauss mais une loi de Bernoulli, une loi binomiale (ou multinomiale), une loi de Poisson, etc.
- De façon similaire au modèle linéaire, on va écrire que le paramètre de localisation de la loi varie avec les prédicteurs.

- Exemple des variables dichotomiques : loi de Bernoulli
  $Y_i \in \{0, 1\}$ alors $Y_i \sim B(\pi_i)$ avec

$$g(\pi_i) = \beta_0 + \mathbf{X}_i^T \boldsymbol{\beta} \text{ ou } \quad \pi_i = g^{-1}\left(\beta_0 + \mathbf{X}_i^T \boldsymbol{\beta}\right)$$

  La fonction $g$ est une fonction de lien définie de $[0, 1]$ sur $\mathbb{R}$. En effet, $\beta_0 + \mathbf{X}_i^T \boldsymbol{\beta} \in \mathbb{R}$ alors que $0 < \pi_i < 1$ est une probabilité.

- Exemple des variables de comptage : loi de Poisson
  $Y_i \in \{0, 1, 2, 3, \cdots\}$ alors $Y_i \sim P(\lambda_i)$ avec

$$\exp(\lambda_i) = \beta_0 + \mathbf{X}_i^T \boldsymbol{\beta}$$

  En effet, $\beta_0 + \mathbf{X}_i^T \boldsymbol{\beta} \in \mathbb{R}$ alors que $\lambda_i \in \mathbb{R}^+$.

# Introduction d'une fonction de lien

- Pour choisir un modèle linéaire généralisé (GLM) il faut
  - choisir la loi de $Y|X = x$ dans la famille exponentielle des GLM.
  - choisir une fonction de lien inversible $g$.
- Pour utiliser un modèle GLM il faudra donc estimer les paramètres $x^T\beta$ . Une fois cette estimation réalisée, on disposera d'une estimation de $\eta(x)$ ainsi que de $E(Y|X = x)) = g^{-1}(x^T\beta)$.
- A chaque loi (ou type de variable $Y$), on associe une fonction de lien canonique

| Modèle | Logistique | Log-linéaire | Linéaire |
|---|---|---|---|
| Loi de $Y\|X = x$ | Bernoulli | Poisson | Gauss |
| $E(Y\|X = x)$ | $logit(E(Y\|X = x)) = x^T\beta$ | $log(E(Y\|X = x)) = x^T\beta$ | $E(Y\|X = x)$ |

# Alternative : introduction d'une variable latente

- Une alternative à l'approche précédente consiste à introduire une variable latente qui suit un modèle linéaire.
- Exemple de la loi de Bernoulli
  On définit alors

$$Y_i^* = \beta_0 + \mathbf{X}_i^T \boldsymbol{\beta} + u_i$$

avec $u_i$ une variable à densité symétrique et de fonction de répartition $F$. On n'observe pas directement $Y_i^*$ mais on observe $Y_i$ et

$$Y_i = 1 \text{ si } Y_i^* > 0$$

$$Y_i = 0 \text{ si } Y_i^* \leq 0$$

- Les deux formulations sont équivalentes. En effet, ce second modèle permet d'écrire la probabilité

$$\pi_i = P(Y_i = 1) = P(Y_i^* > 0) = P(u_i > -\beta_0 - \mathbf{X}_i^T \boldsymbol{\beta}) = 1 - F(-\beta_0 - \mathbf{X}_i^T \boldsymbol{\beta})$$

$F$ définit de façon unique la fonction de lien.

# Données censurées (ou tronquées)

- Remarque : dans le cas de données censurées on utilise naturellement la formulation avec variable latente.
- Exemple de données censurées
  Un exemple typique de données censurées à droite sont des données de suivi de cohorte (ex : cancer) avec lesquelles on cherche à estimer une durée de survie. On n'observe pas la date de décès des patients encore vivant à la fin de l'étude.

# Inférence

- Dans les modèles linéaires généralisés, on n'a pas accès aux moindres carrés ($Y_i^*$ n'est pas observée) et l'estimation des paramètres du modèle se fait par maximum de vraisemblance.
- Comme dans le cas du modèle linéaire, on dispose d'outils statistiques pour valider et interpréter un modèle ou pour comparer des modèles entre eux.

# Outline

# Outline

# Variables catégorielles

- Quand la variable à expliquer est catégorielle, le modèle naturel est la régression logistique.

- Variable binaire ou dichotomique : variable qualitative qui ne peut prendre que deux valeurs (généralement codée 0 ou 1).
  Exemple : malade (oui/non), sexe (H/F), ...

- Variable polytomique : variable qualitative qui peut prendre comme valeur une modalité parmi $K$.
  Exemple : CSP, partis politiques, symptômes d'une maladie...

- Variable ordinale : variable qualitative qui peut prendre comme valeur une modalité parmi $K$ qui sont ordonnées.
  Exemple : intensité d'une douleur, classes d'âge, ...

# Outline

# Exemple : diabète

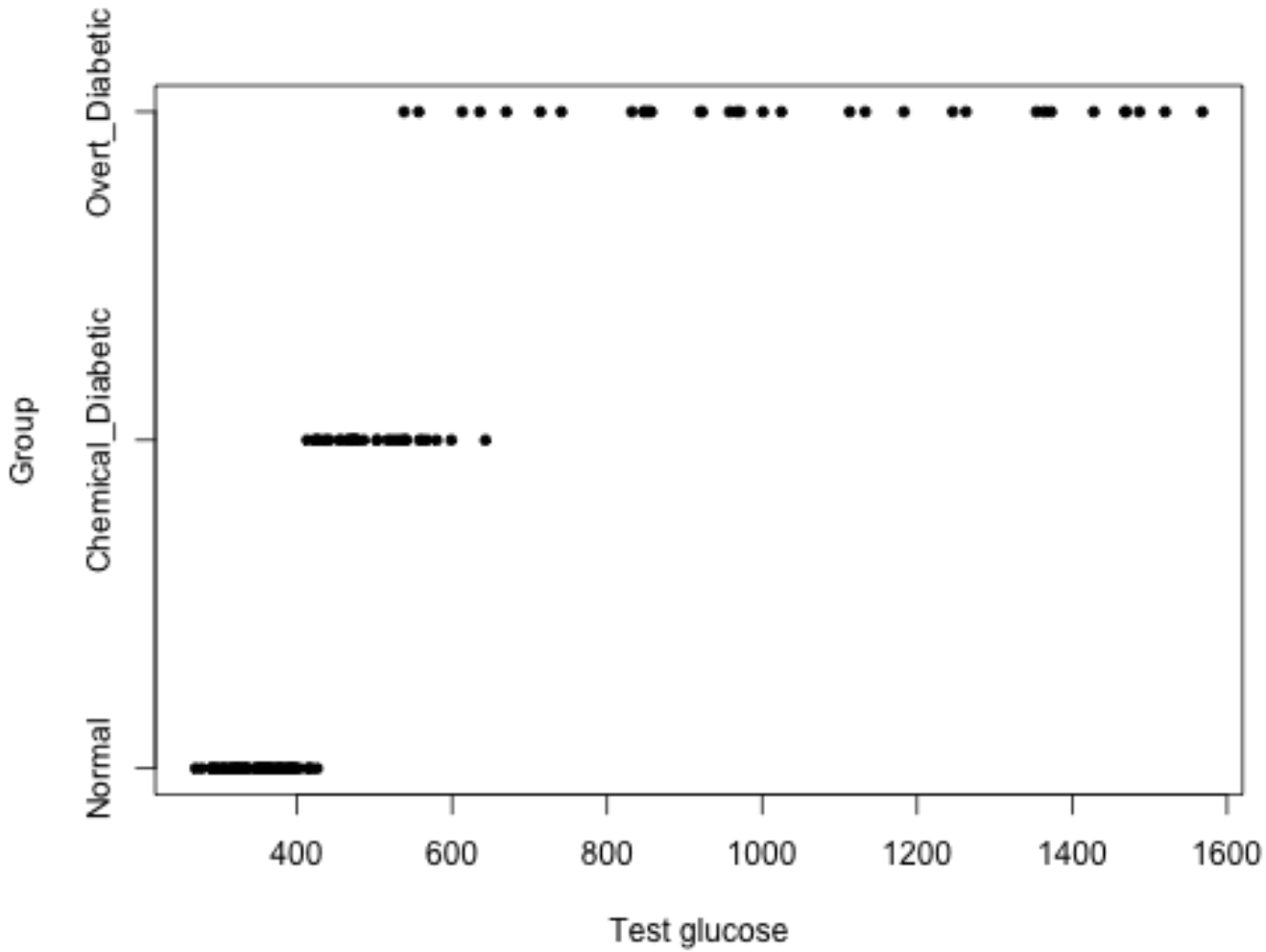
- Reaven et Miller (1979) ont collecté des données de 145 adultes non obèses diagnostiqués "chimical diabetic", "overt diabetic" ou "normal". L'étude avait pour objectif de déterminer des relations entre des données biologiques (prise de sang) et le statut diabétique du patient.
- Les variables mesurées sont
  - rw : poids relatif, ie ratio entre le poids et le poids attendu sachant la taille
  - fpg : glycémies à jeun
  - ga : glycémies test (mesure de l'intolérance au glucose)
  - ina : insuline durant le test (mesure de la réponse insuline au glucose administré par voie orale )
  - sspg : concentration de glucose plasmatique à l'équilibre
  - cc : groupe clinique (3 =overt diabetic", 2= chemical diabetic, 1=normal)

| ident | rw | fpg | ga | ina | sspg | cc |
|---|---|---|---|---|---|---|
| 1 | 0.81 | 80 | 356 | 124 | 55 | Normal |
| 2 | 0.95 | 97 | 289 | 117 | 76 | Normal |
| 3 | 0.94 | 105 | 319 | 143 | 105 | Normal |
| 4 | 1.04 | 90 | 356 | 199 | 108 | Normal |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

Reaven, G. M. and Miller, R. G. (1979). An attempt to define the nature of chemical diabetes using a multidimensional analysis. Diabetologia 16, 17-24.

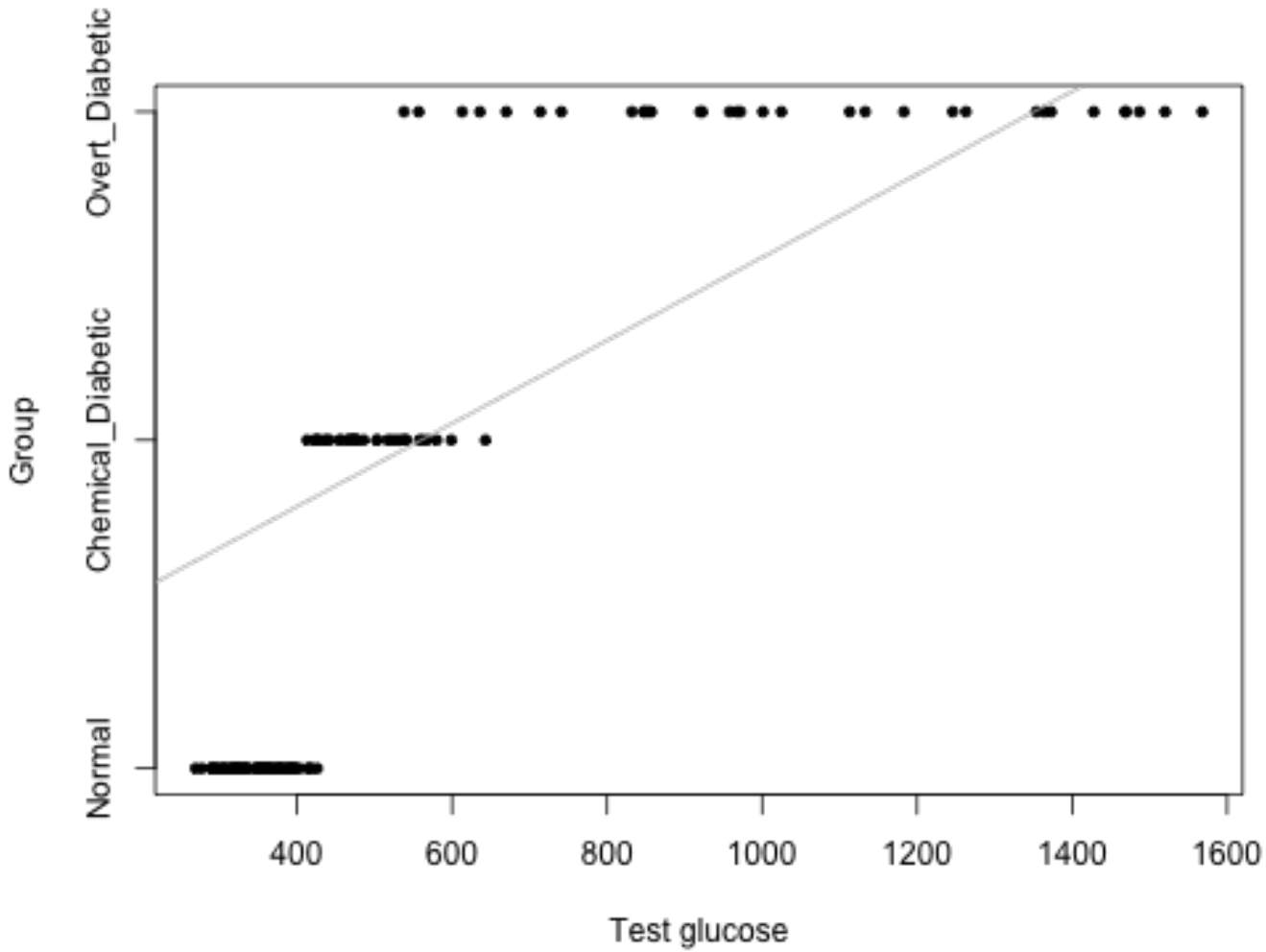
# Exemple : diabète

- On observe une dépendance claire entre les glycémies test et le groupe clinique.



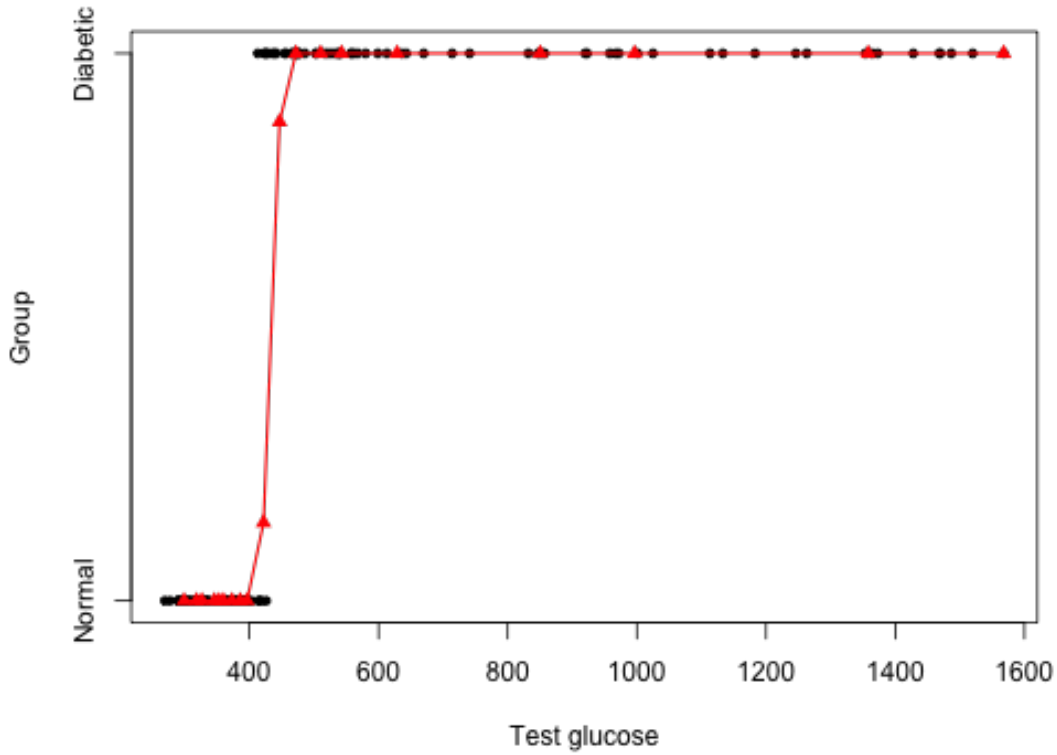
# Exemple : diabète

- Ajuster un modèle de régression linéaire (courbe grise) n'a pas de sens !



# Exemple : diabète, cas binaire

- On recode la variable groupe pour se ramener à une variable binaire.
- Sur la figure, on trace en rouge la proportion de patients diabétiques en fonction de la mesure d'intolérance au glucose.
  Pour l'obtenir on calcule la moyenne des $y_i$ sur des groupes de d'individus ayant un test de glucose similaire.
- On observe que la courbe rouge croit selon une forme logistique entre 0 et 1.



# Exemple : diabète, cas binaire

- Notons $Y$ la variable de groupe (variable à prédire).
- Pour un patient $i$, $Y_i$ suit une loi de Bernoulli de paramètre $\pi_i$.
- La figure montre bien que $\pi_i$ varie en fonction des covariables.
- En pratique, on va modéliser la relation entre les covariables $\mathbf{X_i}$ et $\pi_i = E(Y_i)$.

$$g(\pi_i) = \beta_0 + \mathbf{X}_i^T \boldsymbol{\beta}$$

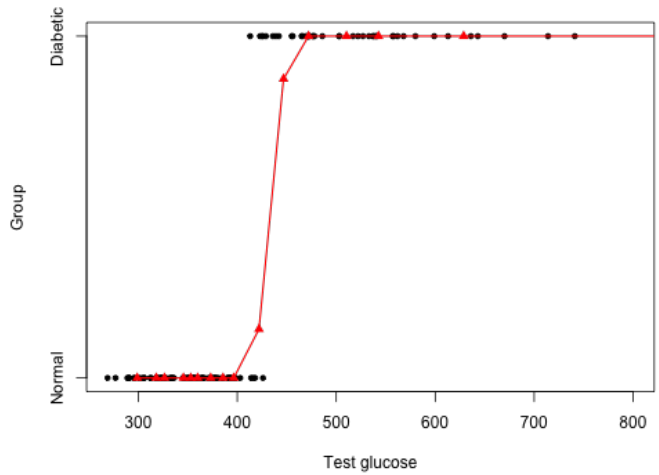
La fonction $g(.) = logit(.)$ est une fonction de lien définie de $[0, 1]$ sur $\mathbb{R}$.

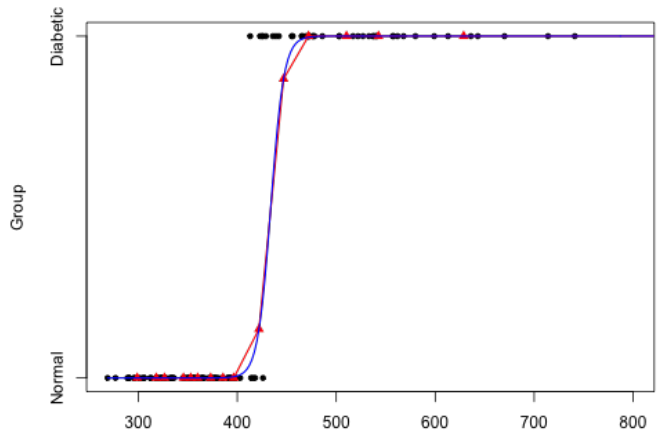
- $g$ est l'inverse de la fonction logistique (en bleu) . Cette fonction est parfois appelée fonction sigmoide.

$$g^{-1}(x) = \frac{\exp(x)}{1 + \exp(x)}$$

- Autrement dit,

Zoom sur les glycémies<800



# Outline

# D'où vient le lien logit ?

- On a vu qu'on ne peut pas modéliser directement l'évolution de $\pi_i$ comme une fonction linéaire des covariables $\mathbf{X}_i$.
- On va donc transformer ces probabilités pour ne plus avoir la contrainte d'être dans un intervalle restreint.
- On définit alors le ratio des cas favorables sur les cas défavorables (risque relatif).

$$odds_i = \frac{\pi_i}{1 - \pi_i}$$

Il est parfois plus naturel de parler en terme de risque relatif plutôt que de probabilité.

- Ensuite, on prend le log avant de s'affranchir de la contrainte de positivité.

$$\eta_i = logit(\pi_i) = \log \frac{\pi_i}{1 - \pi_i}$$

On note que si $\pi_i = 1/2$, $logit(\pi_i) = 0$.

- La transformation $logit()$ est bijective. Son inverse est la fonction sigmoide

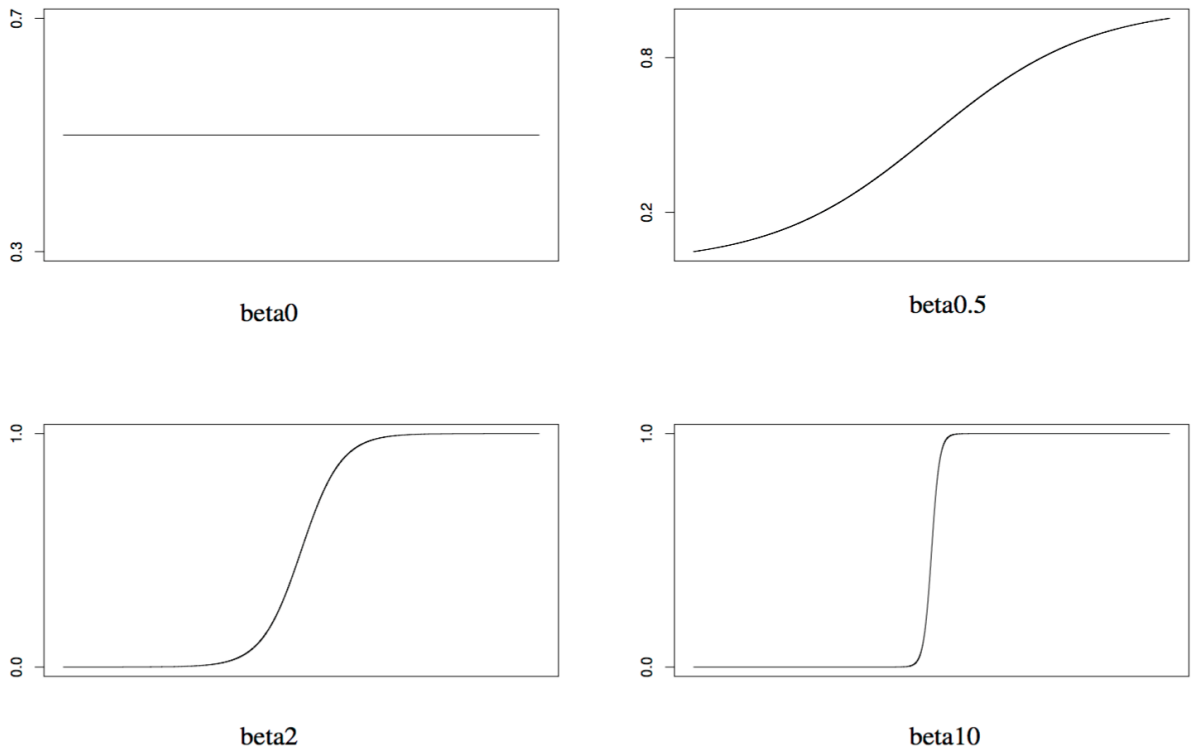
$$\pi_i = logit^{-1}(\eta_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}}$$

# Le lien logit

- Dans un premier temps, on peut retenir le modèle suivant

$$P(Y = 1|\mathbf{X} = \mathbf{x}) = \frac{e^{\mathbf{x}^T\beta}}{1 + e^{\mathbf{x}^T\beta}}$$

- Allure de la courbe pour différentes valeurs de $\beta$




- Lorsque $\beta$ augmente, $P(Y = 1|\mathbf{X} = \mathbf{x})$ est souvent proche de 0 ou 1.

# Odds et odds ratio

- On peut être tenté de dire : plus $\beta$ est grand, mieux on discrimine.
- Prudence : tout dépend de l'échelle de $x$ (si $x$ change d'échelle, $\beta$ va également changer...)
- Les coefficients du modèle logistique sont souvent interprétés en terme d'odds ratio.
- L'odds (chance) pour un individu x d'obtenir la réponse $Y = 1$ est défini par :

$$odds(x) = \frac{P(Y = 1|\mathbf{X} = \mathbf{x})}{1 - P(Y = 1|\mathbf{X} = \mathbf{x})}$$

- L'odds ratio (rapport des odds) (rapport des chances) entre deux individus $x$ et $\tilde{x}$ est

$$OR(x, \tilde{x}) = \frac{odds(x)}{odds(\tilde{x})}$$

# Odds et odds ratio : exemple

- Supposons qu'à un moment donné un cheval $x$ a une probabilité $p(x) = 3/4$ de perdre. Cela signifie que sur 4 courses disputées, il peut espérer en gagner une et en perdre 3. L'odds vaut 3/1 (3 défaites contre 1 victoire, on dit également que ce cheval est à 3 contre 1).
- Pour la petite histoire, si l'espérance de gain était nulle, cela signifierait que pour 10 euros joués, on peut espérer 30 euros de bénéfice si le cheval gagne. Le rapport $p(x)/(1-p(x))$ varie entre 0 (que des victoires) et l'infini (que des défaites) en passant par 1 (une victoire pour une défaite).

# Odds et odds ratio

- Il faut être prudent avec l'interprétation des OR : ils sont très souvent utilisés mais pas toujours bien interprétés.
- Comparaison de probabilités de succès entre deux individus :

$$OR(x, \tilde{x}) > 1 \Leftrightarrow P(Y = 1|X = x) > P(Y = 1|X = \tilde{x})$$

$$OR(x, \tilde{x}) = 1 \Leftrightarrow P(Y = 1|X = x) = P(Y = 1|X = \tilde{x})$$

$$OR(x, \tilde{x}) < 1 \Leftrightarrow P(Y = 1|X = x) < P(Y = 1|X = \tilde{x})$$

- Interprétation en termes de risque relatif : dans le cas où $P(Y = 1|X = x)$ et $P(Y = 1|X = \tilde{x})$ sont très petits par rapport à 1, on peut faire l'approximation

$$OR(x, \tilde{x}) \approx \frac{P(Y = 1|X = x)}{P(Y = 1|X = \tilde{x})}$$

  et interpréter "simplement".
- Mesure de l'impact d'une variable : pour le modèle logistique

$$p_\beta(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

  il est facile de vérifier que

$$OR(x, \tilde{x}) = \exp(\beta_1(x_1 - \tilde{x}_1)) \exp(\beta_2(x_2 - \tilde{x}_2)) \cdots \exp(\beta_p(x_p - \tilde{x}_p))$$

  Si on considère 2 observation qui diffèrent seulement par la $j$ième variable, alors $OR(x, \tilde{x}) = \exp(\beta_j(x_j - \tilde{x}_j))$ mesure l'influence de cette variable.

# Outline

# Exemple : données de contraception

- Little (1978) étudie des données de sondage aux Fidji portant 1607 femmes. Il cherche à modéliser l'utilisation (ou non) d'une contraception en fonction de l'âge, du niveau d'éducation et du désir d'enfant.
- Les variables sont toutes discrètes. On observe donc le nombre de femmes utilisant ou non une contraception dans les différents groupes.

| Age | Education | Desires More Children ? | Contraceptive use No | Yes | Total |
|---|---|---|---|---|---|
| < 25 | Lower | Yes | 53 | 6 | 59 |
| | | No | 10 | 4 | 14 |
| | Upper | Yes | 212 | 52 | 264 |
| | | No | 50 | 10 | 60 |
| 25-29 | Lower | Yes | 60 | 14 | 74 |
| | | No | 19 | 10 | 29 |
| | Upper | Yes | 155 | 54 | 209 |
| | | No | 65 | 27 | 92 |
| 30-39 | Lower | Yes | 112 | 33 | 145 |
| | | No | 77 | 80 | 157 |
| | Upper | Yes | 118 | 46 | 164 |
| | | No | 68 | 78 | 146 |
| 40-49 | Lower | Yes | 35 | 6 | 41 |
| | | No | 46 | 48 | 94 |
| | Upper | Yes | 8 | 8 | 16 |
| | | No | 12 | 31 | 43 |
| Total | | | 1100 | 507 | 1607 |

# Loi binomiale

- On peut définir

$$Y_i = \begin{cases} 1 & \text{si la femme i utilise une contraception} \\ 0 & \text{sinon} \end{cases}$$

$Y_i$ suit une loi de Bernoulli de paramètre $\pi_i$ et on a

$$P(Y_i = y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

- Supposons maintenant qu'on s'intéresse à un groupe $i$, on définit alors

$$Y_i = \text{ nombre de femmes utilisant une contraception dans le groupe } i.$$

$Y_i$ prend ses valeurs dans $0,1,\cdots,n_i$. Si les $n_i$ observations de chaque groupe sont indépendantes et ont la même probabilité de succès $\pi_i$ , alors $Y_i$ suit une loi binomiale de paramètres $\pi_i$ et $n_i$ : $Y_i \sim B(n_i, \pi_i)$

$$P(Y_i = y_i) = \begin{pmatrix} n_i \\ y_i \end{pmatrix} \pi_i^{y_i}(1-\pi_i)^{n_i-y_i}$$

- On remarque que si $n_i = 1$ on retrouve la loi de Bernoulli.

# Données individuelles ou agrégées

- Données individuelles

| | X | Y |
|---|---|---|
| individu | désir d'enfant (oui/non) | Contraception (oui/non) |

- Loi de Bernoulli

$$P(Y_i = y_i) = \pi_i$$

$$E(Y_i) = \pi_i$$

$$Var(Y_i) = \pi_i(1 - \pi_i)$$

$$P(Y_i = y_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}$$

- Données agrégées

| X | Y | n |
|---|---|---|
| désir d'enfant (oui/non) | Contraception (oui/non) | Nombre de femmes dans ce groupe |

- Loi de Binomiale

$$Y_i \sim B(n_i, \pi_i)$$

$$E(Y_i) = n_i\pi_i$$

$$Var(Y_i) = n_i\pi_i(1 - \pi_i)$$

$$P(Y_i = y_i) = \begin{pmatrix} n_i \\ y_i \end{pmatrix} \pi_i^{y_i}(1 - \pi_i)^{n_i - y_i}$$

# Modèle logistique

- Soient $y_1, \cdots, y_k$, $k$ observations indépendantes d'une variable aléatoire $Y_i$ telle que $Y_i \sim B(n_i, \pi_i)$, $n_i \geq 1$.
- On suppose de plus que

$$logit(\pi_i) = \mathbf{X}_i^T \boldsymbol{\beta}$$

avec $\mathbf{X}_i$ le vecteur des covariables (incluant la variable constante égale à 1 pour l'intercept).

- Ce modèle est appelé, modèle linéaire généralisé avec réponse binomiale et lien logit.
- Les coefficients $\boldsymbol{\beta}$ peuvent être interprétés de façon similaire à ceux modèle linéaire, mais il faut garder en tête qu'ils donnent la variation du logit et non directement de la moyenne.
- On remarque que

$$\frac{\pi_i}{1-\pi_i} = exp(\mathbf{X}_i^T \boldsymbol{\beta})$$

Imaginons que la variable $j$ augmente de 1 : $\mathbf{x}_i^T \boldsymbol{\beta} \rightarrow \mathbf{x}_i^T \boldsymbol{\beta} + \beta_j$. En prenant l'exponentiel, on a $e^{\mathbf{x}_i^T \boldsymbol{\beta}} e^{\beta_j}$

$$e^{\beta_j} = \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}} e^{\beta_j}}{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}$$

représente un *odds ratio* ie l'évolution du risque relatif quand la variable augmente de 1.

# Modèle logistique

- On peut aussi se rappeler que

$$\pi_i = \frac{e^{\mathbf{x}_i^T \beta}}{1 + e^{\mathbf{x}_i^T \beta}}$$

  Le terme de gauche est une moyenne, comme on en a l'habitude. Mais le terme de droite est plus difficile à interpréter.

- Pour les prédicteurs continus, on peut regarder la dérivée de cette expression

$$\frac{\partial \pi_i}{\partial x_{ij}} = \beta_j \pi_i (1 - \pi_i)$$

  On voit que l'effet de la variable $j$ dépend de la valeur du prédicteur et de celle de la probabilité.

- En pratique, dans le cas où on prédit une variable binaire, on interprète les odds ratio (rapport de risques relatif) et/ou la probabilité associée à la moyenne des variables explicatives dans chaque groupe.

# Exemple des données de diabète

- Dans le cas des données diabète,

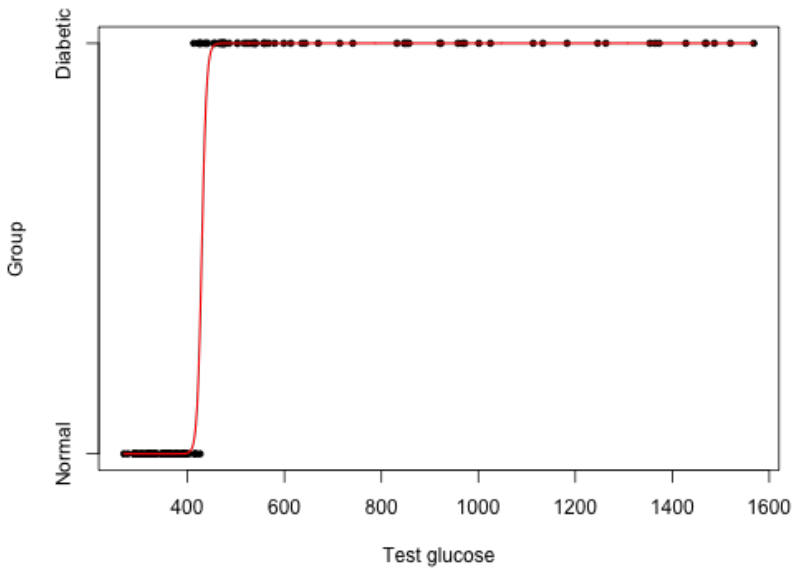
$$logit(P(Y = \text{"}Diabete\text{"})) = -90 + 0.21ga$$

et $exp(0.21) = 1.23$ signifie que quand le taux de glycémie test augmente de 1 point, le rapport des risques relatifs augmente de 23%.

- On peut aussi calculer $P(Y = \text{"}Diabete\text{"})$ pour la moyenne de la glycémie test dans chaque groupe.

$$P(Y = \text{"}Diabete\text{"} | ga = \bar{ga}_{\text{Diabet}}) = 1$$

$$P(Y = \text{"}Diabete\text{"} | ga = \bar{ga}_{\text{Normal}}) = 10^{-5}$$




La pente de la courbe rouge est d'autant plus forte que la partie linéaire est grande.

# Identifiabilité : cas des variables qualitatives

- Tout comme pour le modèle d'analyse de variance, une variable qualitative est représentée par les indicatrices associées aux différentes modalités.
- Considérons un modèle où la seule variable explicative est le `sexe`

$$\pi(x) = \beta_0 + \beta_F \mathbf{I}_F(x) + \beta_H \mathbf{I}_H(x)$$

ce qui s'écrit aussi

$$\pi(x) = \beta_0 + \beta_F + (\beta_H - \beta_F)\mathbf{I}_H(x)$$

et il y a une infinité d'écriture possible.

- La 1ère écriture correspond à une matrice de design $\mathbf{X}$ à 3 colonnes. Les colonnes 2 et 3 sont liées ce qui rend l'estimation impossible.
- Une solution pour pallier cette difficulté consiste à mettre une contrainte sur les coefficients $\beta_F$ et $\beta_H$.
- La solution souvent utilisée par les logiciels est de supprimer une des colonnes de la matrice $\mathbf{X}$, ce qui revient à considérer que le coefficient de la modalité associée à cette colonne est nul. Cette modalité est alors prise comme modalité de référence.

# Exemple du modèle à un facteur

- Revenons à l'exemple des données d'usage de contraception. On considère le facteur âge.
- Chaque groupe a son propre logit

$$logit(\pi_i) = \eta + \alpha_i$$

Pour éviter la redondance, on adopte la méthode qui consiste à supposer que $\alpha_1 = 1$.

- Dans ce cas, $\eta$ correspond au logit du groupe de référence et $\alpha_i$ mesure la différence de logit entre le groupe de référence et le groupe $i$.

| Paramètre | Symbole | Estimation | Std. err. |
|---|---|---|---|
| Constant | $\eta$ | -1.507 | 0.130 |
| 25-29 | $\alpha_2$ | 0.461 | 0.173 |
| 30-39 | $\alpha_3$ | 1.048 | 0.154 |
| 40-49 | $\alpha_4$ | 1.425 | 0.194 |

- Le logit de la constante, -1.51, pour les femmes de moins de 25 ans correspond à un *odds* de 0.22 ie que le modèle prédit que 22% des femmes de moins de 25 ans utilisent une contraception. La valeur empirique est aussi 22%.
- En prenant l'exponentiel des coefficients des groupes d'âge, on obtient des *odds ratio* de 1.59, 2.85 et 4.16 ie que le risque d'utiliser une contraception augmente avec l'âge jusqu'à être multiplié par 4.

# Outline

# Outline

# Méthode du maximum de vraisemblance

- Soit $Y_i$ une variable aléatoire de loi $f(.; \boldsymbol{beta})$ : $P(Y_i = y) = f(y; \boldsymbol{beta})$
- Soient $y_1, \cdots, y_n$, $n$ réalisations indépendantes de $Y_1, \cdots, Y_n$
- La vraisemblance du modèle $f(.; \boldsymbol{beta})$ sachant l'échantillon $y_1, \cdots, y_n$ est la probabilité d'observer cet échantillon pour un vecteur $\beta$ donné
- Autrement dit,

$$\mathcal{L}(\beta) = P_\beta(Y_1 = y_1, \cdots, Y_n = y_n)$$

Par l'indépendance des variables $Y_1, \cdots, Y_n$, on a

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} P_\beta(Y_i = y_i) = \prod_{i=1}^{n} f(y_i; \beta_i)$$

Et en prenant le logarithme, on obtient

$$\log \mathcal{L}(\beta) = \sum_{i=1}^{n} \log P_\beta(Y_i = y_i) = \sum_{i=1}^{n} \log f(y_i; \boldsymbol{beta})$$

On note que le log est une fonction croissante qui ne change pas la position des optima locaux d'une fonction.

- Le vecteur $\widehat{\beta}$ qui réalise le maximum de vraisemblance est celui qui rend l'échantillon le plus vraisemblable.

# Comment trouver le maximum de vraisemblance

- L'approche standard pour trouver le maximum d'une fonction de plusieurs variables consiste à annuler son gradient (dérivée première) et à vérifier que son hessien (dérivée seconde) est défini négatif.
- Pour obtenir l'estimateur du maximum de vraisemblance on résout donc le système d'équations suivant

$$\begin{cases} \frac{\partial \log \mathcal{L}(\boldsymbol{\beta})}{\partial \beta_1} = 0 \\ \vdots \\ \frac{\partial \log \mathcal{L}(\boldsymbol{\beta})}{\partial \beta_p} = 0 \end{cases}$$

  Ce système est appelée système d'équations du score.
- C'est un système de $p$ équations à $p$ inconnues.
- Dans le cas où les équations sont non linéaires (comme pour le modèle logistique), ce système n'admet pas de solution analytique et on approche sa solution en utilisant un algorithme itératif.

# Propriétés de l'estimateur du maximum de vraisemblance

- Pour les modèles linéaires généralisés, on peut montrer que l'estimateur du maximum de vraisemblance existe et qu'il est unique.
- Il a les propriétés suivantes
  - Il est consistent : il tend vers la vraie valeur du paramètre quand le nombre d'observations tend vers l'infini.
  - Il est asymptotiquement efficace : c'est l'estimateur qui a la plus petit variance possible, sous réserve que le nombre d'observations soit assez grand.
  - Il est asymptotiquement distribué suivant une loi de Gauss : quand le nombre d'observations tend vers l'infini, l'estimateur du maximum de vraisemblance tend, en loi, vers une variable gaussienne.

# Modèle logistique

- Dans le modèle logistique, l'estimation se fait par maximum de vraisemblance.
- Soient $(y_1, \mathbf{x}_1), \cdots, (y_n, \mathbf{x}_n)$, $n$ observations indépendantes du vecteur $(Y, \mathbf{X})$.
- La log vraisemblance est donnée par

$$\log \mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \log(\pi_i) + (n_i - y_i) \log(1 - \pi_i)$$

où $\pi_i$ dépend des covariables $\mathbf{x}_i$ et d'un vecteur de paramètres $\boldsymbol{\beta}$.

▸ Vraisemblance

- Pour maximiser la log-vraisemblance, on peut calculer les dérivées premières et secondes. Et utiliser l'algorithme de Newton-Raphson pour approcher l'estimateur du maximum de vraisemblance. On obtient ainsi le score et l'information de Fisher.

▸ Newton

- En pratique, on utilise l'algorithme du "iteratively re-weighted least square" (IRLS). On peut montrer que c'est une façon d'écrire l'algorithme de Newton-Raphson.

# IRLS pour le modèle logistique

- Algorithme
  Choisir un $\boldsymbol{\beta_0}$ initial
  Répéter jusqu'à convergence
    Calculer $\widehat{\boldsymbol{\eta}} = \mathbf{X}^T \boldsymbol{\beta}_{k-1}$ avec $k$ le numéro d'itération
    Calculer $\widehat{\boldsymbol{\mu}} = logit^{-1}(\widehat{\boldsymbol{\eta}})$, probabilité prédite de $Y_i = 1$
    Calculer $z_i = \hat{\eta}_i + (y_i - \hat{\mu}_i)\frac{d\eta_i}{d\mu_i} = \hat{\eta}_i + \frac{n_i}{\hat{\mu}_i(n_i - \hat{\mu}_i)}(y_i - \hat{\mu}_i)$
    Régresser $\mathbf{z}$ sur les covariables

$$\boldsymbol{\beta}_k \leftarrow (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{z}$$

avec une matrice de poids diagonale de terme

$$w_{ii} = \hat{\mu}_i(n_i - \hat{\mu}_i)/n_i$$

- L'estimateur obtenu est consistent (ie qu'il tend en moyenne vers la vraie valeur du paramètre quand $n$ tend vers l'infini) et sa variance est donnée par

$$Var(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$$

# Exemple sur données simulées

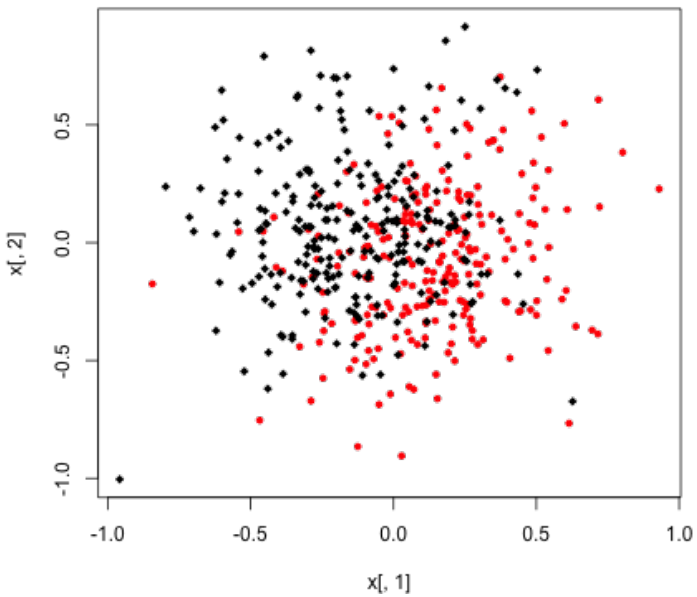
- On suppose que pour tout $Y_i$ soit une loi de Bernoulli de paramètre

$$\pi_i = 5X_{1i} - 2X_{2i}$$

avec $X_1$ et $X_2$ des variables aléatoires indépendantes de loi de Gauss centrée et de variance 0.09.

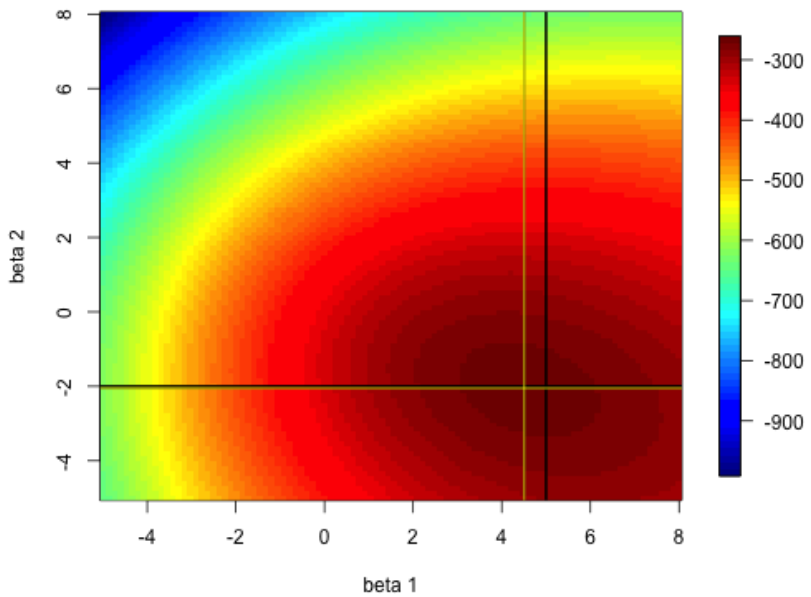
- On simule 500 réalisations indépendantes.

Scatter plot




Y=0 (noir), Y=1 (rouge)

Vraisemblance en fonction de $\beta$




Vraies valeurs de $\beta$ (noir), estimation (jaune)

# Exemple sur données simulées

- Code R

```
> n = 500
> x=matrix(rnorm(2*n),n,2)
> beta = matrix(c(5,-2),2,1)*.3
> pi = exp(x%*%beta)/(1+exp(x%*%beta))
> u = runif(n)
> Y = u<pi
> mod.glm = glm(Y~x,family=binomial,control=list(trace=1))
Deviance = 534.4288 Iterations - 1
Deviance = 532.3176 Iterations - 2
Deviance = 532.3024 Iterations - 3
Deviance = 532.3024 Iterations - 4
> summary(mod.glm)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.00211    0.10581   0.020    0.984
x1           4.50960    0.46758   9.645  < 2e-16 ***
x2          -2.05054    0.36946  -5.550 2.86e-08 ***
---
> pi.hat = exp(mod.glm$coefficients[1]+
+ x%*%mod.glm$coefficients[2:3])/(1+exp(x%*%mod.glm$coefficients[2:3]))
> table(pi>.5,pi.hat>.5)
       FALSE TRUE
  FALSE   244    5
  TRUE      1  250
```

- L'algorithme itératif converge en 4 itérations et on retrouve des valeurs de coefficients proches des vrais paramètres.
- En prédiction, 5 individus sont mal classés.

# Cas (pathologique) des nuages de points complètement séparables (1/2)

- Dans le cas où le nuage de points $(y_1, \mathbf{x}_1)$ ,$\cdots$,$(y_n, \mathbf{x}_n)$ est complètement séparable l'algorithme IRLS ne converge pas.
- On dit qu'un nuage de points $(y_1, \mathbf{x}_1)$ ,$\cdots$,$(y_n, \mathbf{x}_n)$, avec $x_i \in \mathbb{R}^p$ pour tout $i$, est complètement séparable s'il existe un $\boldsymbol{\beta} \in \mathbb{R}^p$ tel que
  pour tout $i$ tel que $Y_i = 1$, $\mathbf{x}_i^T \boldsymbol{\beta} > 0$ et
  pour tout $i$ tel que $Y_i = 0$, $\mathbf{x}_i^T \boldsymbol{\beta} < 0$
- On dit qu'un nuage de points $(y_1, \mathbf{x}_1)$ ,$\cdots$,$(y_n, \mathbf{x}_n)$, avec $x_i \in \mathbb{R}^p$ pour tout $i$, est quasi-complètement séparable s'il existe un $\boldsymbol{\beta} \in \mathbb{R}^p$ tel que
  pour tout $i$ tel que $Y_i = 1$, $\mathbf{x}_i^T \boldsymbol{\beta} \geq 0$ et
  pour tout $i$ tel que $Y_i = 0$, $\mathbf{x}_i^T \boldsymbol{\beta} \leq 0$

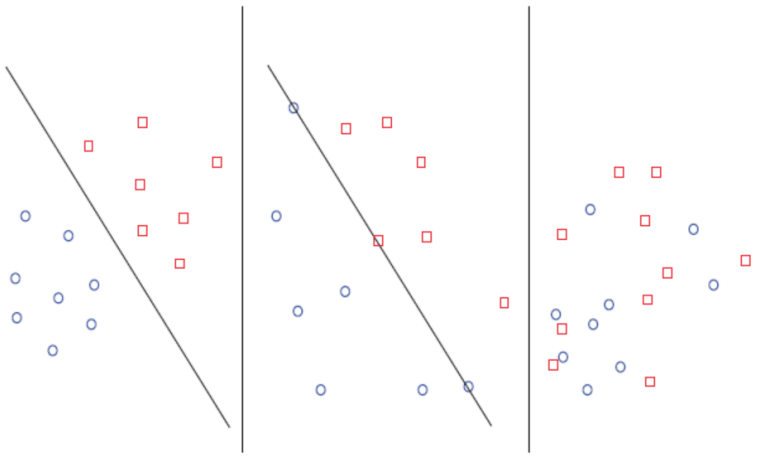


**Illustration**: Exemple de séparabilité complète (gauche), quasi-complité (milieu) et de recouvrement (droite).

# Cas (pathologique) des nuages de points complètement séparables (2/2)

- Dans le modèle logistique, la valeur de $\widehat{\beta}$ est liée à la pente de la frontière entre les groupes.
- On voit bien que dans le cas complètement séparable, on peut trouver une infinité de frontières possibles.
- Dans ce cas, l'estimateur du maximum de vraisemblance n'existe pas.
- Quand on travaille avec des données réelles, il est rare que les nuages de points soient complètement séparables.

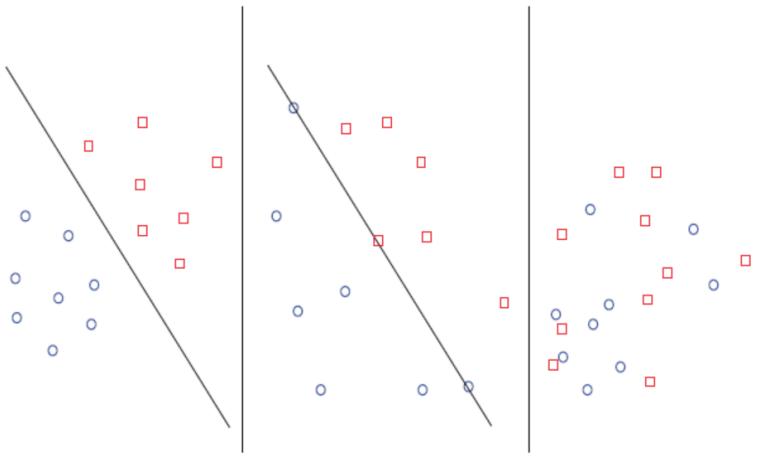
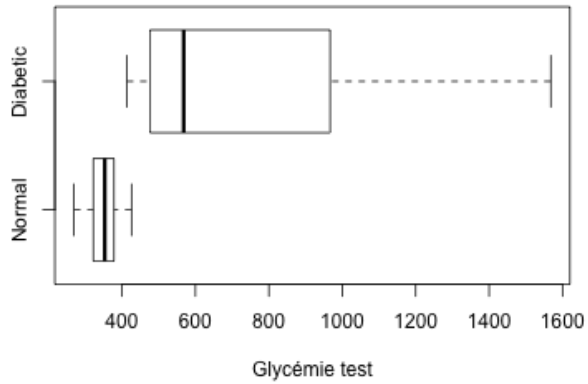


**Illustration**: Exemple de séparabilité complète (gauche), quasi-complité (milieu) et de recouvrement (droite).

```
# On regroupe les cas de diabète
> w = which(chemdiab$cc!="Normal")
> chemdiab$cc <- ordered(chemdiab$cc, levels = c("Normal", "Diabetic"))
> chemdiab$cc[w] = "Diabetic"

# On ajuste un modèle de régression logistique
> mod = glm(cc~ga,data=chemdiab,family=binomial)
Warning message:
glm.fit: des probabilités ont été ajustées numériquement à 0 ou 1

> boxplot(chemdiab$ga~chemdiab$cc)
```



# Outline

# Prédiction

- Sachant la valeur des variables explicatives $\mathbf{x}$, l'estimation de la moyenne de $y$ est $\hat{\mu}$ avec $g(\hat{\mu}) = \mathbf{x}^T\widehat{\boldsymbol{\beta}}$ pour une fonction de lien $g$.
- Par exemple, pour le lien logit

$$logit(\hat{\mu}) = \mathbf{x}^T\boldsymbol{\beta}$$

- On peut construire un intervalle de confiance de cet estimateur pour indiquer sa précision. Pour obtenir cet intervalle de confiance, on a besoin de la loi de $\hat{\mu}$
- La variance du prédicteur linéaire est $\mathbf{x}^T\widehat{\boldsymbol{\beta}}$ est

$$Var(\mathbf{x}^T\widehat{\boldsymbol{\beta}}) = \mathbf{x}^T\left(\mathbf{x}^T W\mathbf{x}\right)^{-1}\mathbf{x}$$

- Un intervalle de confiance $[\mu_-, \mu_+]$ au risque $\alpha$ est obtenu pour $\hat{\mu}$ via

$$logit(\mu_-) = \mathbf{x}^T\widehat{\boldsymbol{\beta}} - \Phi^{-1}(1-\alpha/2)\sqrt{\mathbf{x}^T\left(\mathbf{x}^T W\mathbf{x}\right)^{-1}\mathbf{x}}$$

avec $\Phi^{-1}(1-\alpha/2)$ le quantile de la loi de Gauss centrée réduite. $\mu_+$ est obtenu en ajoutant $\Phi^{-1}(1-\alpha/2)\sqrt{\mathbf{x}^T\left(\mathbf{x}^T W\mathbf{x}\right)^{-1}\mathbf{x}}$

▸ Intervalle de confiance

# Outline

# Déviance

- Quand on ajuste un modèle, on souhaite savoir s'il décrit bien les données.
- La déviance mesure un écart entre les valeurs observées $y_i$ et $n_i - y_i$ et les valeurs estimées $\hat{\mu}_i$ et $n_i - \hat{\mu}_i$.

$$D = 2 \sum_{i=1}^{n} \left\{ y_i \log \left( \frac{y_i}{\hat{\mu}_i} \right) + (n_i - y_i) \log \left( \frac{n_i - y_i}{n_i - \hat{\mu}_i} \right) \right\}$$

  où $y_i$ est la valeur observée et $\hat{\mu}_i$ la valeur prédite pour l'observation $i$.
- Si l'ajustement est parfait le rapport des valeurs observées sur les valeurs prédites est égal à 1 et son log est nul. L'ajustement est donc d'autant meilleur que la déviance est faible.
- Dans le cas de données groupées, la distribution de la déviance tend en loi vers une variable de loi chi2 à $n - p$ degrés de liberté quand $n$ tend vers l'infini, où $n$ est le nombre de groupes et $p$ le nombre de paramètres.
  On peut en déduire un test de qualité d'ajustement.
- Pour des données individuelles, la déviance ne tend pas en loi vers un chi2 et elle sert uniquement d'indice de qualité.

# Test de Pearson

- Pour tester la qualité de l'ajustement, on peut aussi utiliser le test de Pearson.
- Dans le cas de données binomiales, sa statistique de test s'écrit

$$\chi^2_p = \sum_{i=1}^{n} \frac{n_i(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i(n_i - \hat{\mu}_i)}$$

  Chaque terme de la somme est un écart au carré entre la valeur observée et la valeur prédite divisée par la variance de $y_i$.
- Pour les données groupées, la statistique $\chi^2_p$ suit approximativement une loi du chi 2 à $n - p$ degrés de liberté.
- Pour les données individuelles, l'approximation n'est pas de très bonne qualité et le test de Pearson n'est donc pas considéré comme une bonne mesure.
- La statistique de Pearson est aussi utilisée pour calculer des résidus.

# Test d'Hosmer Lemeshow

- Ce test permet de vérifier l'adéquation d'un modèle en présence de données individuelles.
- L'idée est de se rapprocher du cas de données répétées en créant ces répétitions. Le test s'effectue de la manière suivante (voir Hosmer & Lemeshow (2000), chapitre 5 pour plus de précisions).
  1. Les probabilités $\mu_i$ sont ordonnées par ordre croissant
  2. Ces probabilités ordonnées sont ensuite séparées en $K$ groupes de taille égale (on prend souvent $K = 10$ si $n$ est suffisamment grand). On note
  - $m_k$ les effectifs du groupe $k$ ;
  - $o_k$ le nombre de succès ($Y = 1$) observé dans le groupe $k$ ;
  - $\mu_k$ la moyenne des $\mu_i$ dans le groupe $k$. La statistique de test est alors

$$C^2 = \sum_{k=1}^{K} \frac{(o_k - m_k \mu_k)}{m_k \mu_k (1 - \mu_k)}$$

  Le test se conduit de manière identique au test de déviance, la statistique $C^2$ suivant approximativement une loi du chi 2 à $K - 1$ degrés de liberté.
- Sous R : `hoslem.test {ResourceSelection}`

# Autres critères basés sur la vraisemblance

- Pour comparer des modèles, il est courant d'utiliser des critères basés sur une pénalisation de la vraisemblance.
- Le critère d'Akaike est défini par

$$AIC = -2 \log L + 2k$$

avec $k$ le nombre de paramètres à estimer.

- Si l'on considère un ensemble de modèles candidats, le modèle choisi est celui qui aura la plus faible valeur d'$AIC$. Ce critère repose donc sur un compromis entre la qualité de l'ajustement et la complexité du modèle.
- Le critère "Bayes Information criterion" est défini par

$$BIC = -2 \log L + \log(n)k$$

avec $n$ le nombre d'observations.

- L'AIC pénalise le nombre de paramètres moins fortement que le BIC.

# Tests d'hypothèses pour des modèles emboités

- Pour tester des modèles emboités, on utilise des tests basés sur des rapports de vraisemblance.
- Pour fixer les idées, on peut partitionner la matrice de design en deux composantes de dimension $p_1$ et $p_2$

$$\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2) \text{ et } \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}.$$

- On considère le test d'hypothèse nulle

$$H_0 : \beta_2 = 0$$

qui signifie que la composante $\mathbf{X}_2$ n'a pas d'effet sur la réponse.

- On note $D(\mathbf{X}_1)$ la déviance du modèle incluant uniquement $\mathbf{X}_1$ et $D(\mathbf{X}_1 + \mathbf{X}_2)$ la déviance du modèle complet. Alors la statistique

$$\chi^2 = D(\mathbf{X}_1) - D(\mathbf{X}_1 + \mathbf{X}_2)$$

tend en loi vers une variable de loi du chi 2 à $p_2$ degrés de liberté quand le nombre d'observations tend vers l'infini.

# Déviance pour les modèles emboités

$$H_0 : \beta_{rw} = 0$$

```
> logistic.fit1 = glm(cc~fpg,family = binomial,data=chemdiab)
> deviance(logistic.fit1)
[1] 121.944
> aic(logistic.fit1)
       lik       infl       vari        aic
-10.740278   7.470741   6.751109  36.422037
> logistic.fit2 = glm(cc~fpg+rw,family = binomial,data=chemdiab)
> deviance(logistic.fit2)
[1] 110.1438
> aic(logistic.fit2)
      lik      infl      vari       aic
-9.296681 13.896896 12.820934 46.387154
> 1-pchisq( deviance(logistic.fit1)- deviance(logistic.fit2),df=1)
[1] 0.00059
```

Selon les déviances, on conclut, au risque 5%, que l'ajout de la variable `rw` permet d'améliorer significativement le modèle.

```
> p = predict(logistic.fit2,chemdiab,typ="respons")
> table(chemdiab$cc=="Diabetic",p>.5)
           FALSE TRUE
  FALSE     66    10
  TRUE      20    49
```

Pourtant le critère AIC et les erreurs de prédictions sont dégradées : 30/145. ???

# Remarques sur la déviance

- Dans le modèle linéaire, la déviance est égale à la somme des carrés des résidus.
- Les tests de rapport de vraisemblance sont construit à partir des déviances normalisées. Dans le cas du modèle linéaire la normalisation est $\sigma^2$.
- Pour les données binomiales, la déviance joue encore le rôle d'une somme de carrés de résidus. Dans ce cas, la normalisation est égale à 1.
- Le test de Pearson ne peut pas être généralisé pour les modèles emboités. Ceci explique qu'on utilise plus facilement les déviances.

# Tests d'hypothèses pour les paramètres

- On considère le test dont l'hypothèse nulle est

$$H_0 : \beta_j = 0$$

- Comme dans le cas de la régression linéaire, on utilise le test de Wald dont la statistique de test est

$$z = \frac{\hat{\beta}_j}{\sqrt{\widehat{Var}(\hat{\beta}_j)}}.$$

  Cette statistique tend en loi vers une variable de loi de Gauss centrée et réduite quand le nombre d'observations tend vers l'infini.
- La statistique du test de Wald est aussi utilisée pour calculer des intervalles de confiance de niveau de confiance $100(1-\alpha)\%$ pour les paramètres

$$\hat{\beta}_j \pm \Phi^{-1}(1-\alpha/2)\sqrt{\widehat{Var}(\hat{\beta}_j)}$$

  avec $\Phi$ la fonction de répartition de la loi de Gauss centrée réduite [1].

  ▸ Intervalle de confiance
- Les intervalles de confiance des effets en terme de logit sont obtenus en prenant les exponentiels des bornes de l'intervalle précédent.

1. Pour $\alpha = .05$, $\Phi^{-1}(1-\alpha/2) = 1.96$

# Tests pour les paramètres

$$H_0 : \beta_j = 0$$

```
> logistic.fit3 = glm(cc~fpg+rw+sspg,family = binomial,data=chemdiab)
> summary(logistic.fit3)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -18.042959   3.359824  -5.370 7.86e-08 ***
fpg           0.104834   0.024475   4.283 1.84e-05 ***
rw            5.422535   2.195915   2.469  0.01353 *
ina           0.009150   0.003086   2.965  0.00303 **

Null deviance: 200.675  on 144  degrees of freedom
Residual deviance:  96.771  on 141  degrees of freedom
```

Ici on conclut que chacun des paramètres associé aux variables `fpg`, `rw` et `ina` est significatif.

# Outline

# Données de contraception et test d'homogénéité

- Considérons la table à 2 entrées suivante extraite de l'ensemble des données de contraception

$$Y_i \sim \mathcal{B}(n_i, \pi_i)$$

| Desires $i$ | Using $y_i$ | Not Using $n_i - y_i$ | All $n_i$ |
|---|---|---|---|
| Yes | 219 | 753 | 972 |
| No | 288 | 347 | 635 |
| All | 507 | 1100 | 1607 |

- Test d'homogénéité (les 2 groupes ont la même probabilité)
  Ce test correspond au modèle nul ie au modèle où les 2 groupes ont le même logit

$$logit(\pi_i) = \eta$$

```
# Préparation des données
> cuse <- data.frame(matrix(c(
+  0, 219, 753,
+  1, 288, 347), 2, 3, byrow=TRUE))
> names(cuse) <- c("nomore","using","notUsing")
> cuse$n <- cuse$using + cuse$notUsing
> cuse$Y <- cbind(cuse$using, cuse$notUsing)
```

# Ajustement du modèle nul

- Ajustement du modèle nul.

```
> m0 <- glm( Y ~ 1, data=cuse, family=binomial)
> m0

Call:  glm(formula = Y ~ 1, family = binomial, data = cuse)

Coefficients:
(Intercept)
    -0.7746

Degrees of Freedom: 1 Total (i.e. Null);  1 Residual
Null Deviance:      91.67
Residual Deviance: 91.67  AIC: 107.5
```

- On vérifie que le paramètre (intercept) correspond à la proportion de femmes utilisant une contraception.

```
> p <- sum(cuse$using)/sum(cuse$n) # probabilité empirique
> log(p/(1-p)) # logit de la probabilité empirique
[1] -0.7745545
```

# Ecart-types et déviance

- Pour obtenir la déviance et les écart-types, on utilise la fonction `summary`

```
> summary(m0)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.77455    0.05368  -14.43   <2e-16 ***

Null deviance: 91.674  on 1  degrees of freedom
Residual deviance: 91.674  on 1  degrees of freedom
AIC: 107.54
> 1-pchisq(91.674,df=1)
[1] 0
```

  La déviance du modèle nul est 91.67 pour 1 ddl et la pvalue associée est nulle. On rejette donc clairement l'hypothèse selon laquelle les 2 groupes sont homogènes.

- On vérifie que l'écart type du logit observé est la racine carrée de 1/succès+1/échec

```
> sqrt( 1/sum(cuse$using) + 1/sum(cuse$notUsing) )
[1] 0.0536794
```

- Il peut être aussi instructif de calculer la déviance "à la main".

```
> obs <- cuse$Y
> fit <- cbind(p*cuse$n, (1-p)*cuse$n)
> 2*sum(obs*log(obs/fit))
[1] 91.6744
```

# Intervalles de confiance

- On peut obtenir les intervalles de confiance.
  Pour revenir en échelle "probabilité", on utilise la fonction inverse du logit `plogis`

```
> confint(m0)
     2.5 %     97.5 %
-0.8804716 -0.6700014
> plogis(confint(m0))
    2.5 %    97.5 %
0.2930801 0.3384965
```

# Odds-ratio (rapport de cotes), modèle à un facteur

- On ajuste maintenant le modèle à un facteur "ne veut plus d'enfant".

$$logit(\pi_i) = \eta + \beta_i$$

- Ce modèle est saturé pour ce jeu de données (2 paramètres pour 2 probabilités) il a donc une déviance nulle.

```
> m1 <- glm(Y ~ nomore, family=binomial, data=cuse)
> summary(m1)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.23499    0.07677 -16.086   <2e-16 ***
nomore       1.04863    0.11067   9.475   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Null deviance: 9.1674e+01  on 1  degrees of freedom
Residual deviance: 1.7986e-14  on 0  degrees of freedom
AIC: 17.87
```

- La constante correspond au log-odds de "utilise une contraception" parmi les femmes qui veulent encore des enfants et le coefficient "nomore" est la différence en log-odds entre les 2 groupes.

# Interprétation de l'odds-ratio

- L'estimation de $\eta$ est le logit de la proportion observée de femme utilisant une contraception parmi les femmes qui veulent encore des enfants logit(219/972) = -1.235. L'estimation de $\beta$ est la différence entre les logits des 2 groupes, logit(288/635) - logit(219/972) = 1.049.
- En prenant l'exponentiel, on trouve un rapport de cote d'environ 3.

```
> exp(coef(m1)["nomore"])
  nomore
2.853737
```

- Ca ne signifie pas que "les femmes qui ne veulent plus d'enfant ont 3 fois plus de chance d'utiliser une contraception que les autres"
- C'est le rapport des cotes qui est égal à 3 et non la probabilité.
  En effet, on a par définition

$$\frac{\pi_i}{(1-\pi_i)} = \exp(\beta)$$

- Ici on observe que les proportions sont 0.454 (= 288/635) et 0.225 (=213/972), et le rapport est 2.01. Ainsi les femmes qui ne veulent plus d'enfant ont 2 fois plus de chance d'utiliser une contraception que les autres.

# Test de Wald et intervalle de confiance

- Modèle à un facteur

```
> summary(m1)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.23499    0.07677 -16.086   <2e-16 ***
nomore       1.04863    0.11067   9.475   <2e-16 ***
```

- Le test du chi2 associé à la statistique de Wald égale à 89.78 permet de rejeter l'hypothèse selon laquelle le coefficient est nul.

```
> b <- coef(m1)
> se <- sqrt(diag(vcov(m1)))
> (b[2]/se[2])^2
  nomore
89.77765
```

- On peut aussi calculer un intervalle de confiance (qui ne contient pas 0) au risque 5%.

```
> exp(confint(m1,"nomore"))
   2.5 %   97.5 %
2.298942 3.548111
> exp(confint.default(m1,"nomore"))
          2.5 %   97.5 %
nomore 2.297258 3.545015
```

# Outline

# Données

- On considère encore les données de contraception

```
>  cuse <- data.frame(matrix(c(
+ 1,  72, 325,
+ 2, 105, 299,
+ 3, 237, 375,
+ 4,  93, 101), 4, 3, byrow=TRUE))
> names(cuse) <- c("age","using","notUsing")
> cuse$n <- cuse$using + cuse$notUsing
> cuse$age <- factor(cuse$age,
      labels=c("< 25","25-29","30-39","40-49"))
> cuse
    age using notUsing   n
1  < 25    72      325 397
2 25-29   105      299 404
3 30-39   237      375 612
4 40-49    93      101 194

> cuse$Y <- cbind(cuse$using, cuse$notUsing)
```

# Modèle à un facteur

- On considère le modèle en fonction de l'âge, variable à 4 modalités. Il est bien sûr saturé (4 observations pour 4 paramètres).

$$logit(\pi_i) = \eta + \beta_i$$

avec $\beta_1 = 0$. Chaque $i$ correspond à une classe d'âge.

- Ce modèle est analogue à un modèle d'analyse de la variance.

```
> mag <- glm( Y ~ age, family=binomial, data=cuse)
> summary(mag)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.5072     0.1303 -11.571  < 2e-16 ***
age25-29      0.4607     0.1727   2.667  0.00765 **
age30-39      1.0483     0.1544   6.788 1.14e-11 ***
age40-49      1.4246     0.1940   7.345 2.06e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Null deviance: 7.9192e+01  on 3  degrees of freedom
Residual deviance: 2.1405e-13  on 0  degrees of freedom
AIC: 32.647
```

- La déviance du modèle nul correspond à un test de rapport de vraisemblance et permet de rejeter l'hypothèse selon laquelle la probabilité d'utiliser une contraception est la même dans toutes les classes d'âge.

# Coefficients et test de Wald

- Le logit de la modalité de référence -1.51 correspond à un odds de 0.22.
- On observe que les chances de succès augmentent avec l'âge de 59% et 185% quand on passe aux classes d'âge 25-29 et 30-39, et sont quadruplés pour la classe 40-49, en comparaison de la première classe d'âge 25.

```
> b <- coef(mag)
> exp(b[-1])
age25-29 age30-39 age40-49
1.585145 2.852778 4.156353
```

- Test de Wald

$$H_0 : \beta_2 = \beta_3 = \beta_4 = 0$$

Statistique de test :

$$\widehat{\boldsymbol{\beta}}^T var^{-1}(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\beta}}$$

```
> V <- vcov(mag)
> t(b[-1]) %*% solve(V[-1,-1]) %*% b[-1]
          [,1]
[1,] 74.35663
```

La statistique de test vaut 74.35 pour un test du chi2 à 3 ddl. On rejette l'hypothèse nulle.

- Le test basé sur la déviance ($D_0 - D_1 = 79$) et le test de Wald ne sont pas équivalents mais ils conduisent à la même conclusion.

# Outline

# Données

- On considère encore les données de contraception mais on traite l'âge comme une variable continue : on attribue à chaque femme du groupe $i$ l'âge moyen de ce groupe.
- Le modèle s'écrit

$$logit(\pi_i) = \alpha + \beta x_i$$

```
> cuse$agem <- c(20, 27.5, 35, 45)[as.numeric(cuse$age)]
> cuse[,c("age","agem")]
    age agem
1  < 25 20.0
2 25-29 27.5
3 30-39 35.0
4 40-49 45.0
```

# Ajustement du modèle

- On estime alors les paramètres

```
> mam <- glm(Y ~ agem, family=binomial, data=cuse)
> summary(mam)
Deviance Residuals:
      1         2         3         4
-0.3697   -0.3739    1.0845   -0.9750

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.672667   0.233249 -11.458   <2e-16 ***
agem         0.060671   0.007103   8.541   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
    Null deviance: 79.1917  on 3  degrees of freedom
Residual deviance:  2.4034  on 2  degrees of freedom
AIC: 31.05

> b <- coef(mam)
> exp(b["agem"])
    agem
1.062549
```

- Le paramètre de pente indique que le logit de la probabilité d'utiliser une contraception augmente de 0.061 chaque année. Ca correspond à une augmentation du rapport de cote de exp(0.06) = 1.063 soit 6.3%.

# Test de significativité

- Pour tester la significativité de l'âge,

$$H_0 : \beta = 0$$

on peut utiliser le test de Wald dont la statistique vaut 8.54 (voir page précédente). Le carré de cette statistique peut être comparé au quantile du chi2 à un degré de liberté.

```
> qchisq(0.95,df=1) # quantile
[1] 3.841459

> 1-pchisq(8.54,df=1) # pvalue
[1] 0.003474256
```

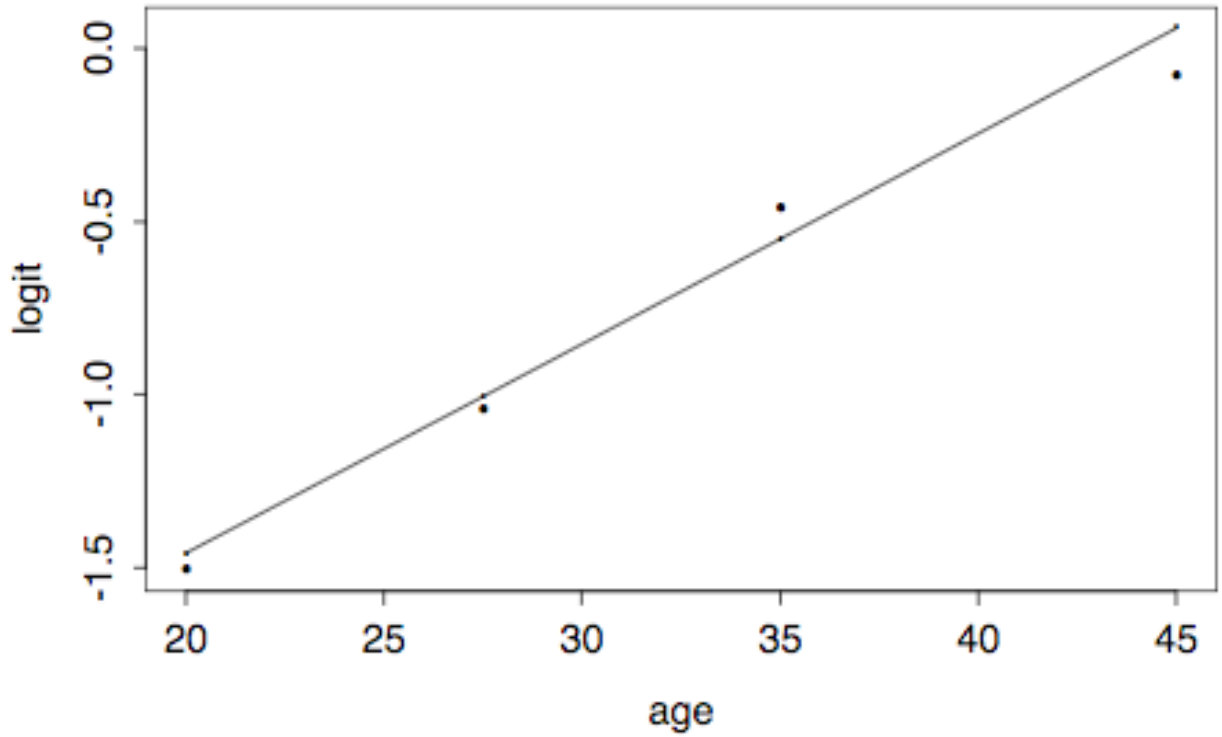
- On peut aussi utiliser un test de rapport de vraisemblance pour comparer ce modèle au modèle nul. La différence de déviance est égale à 76.8.
- Dans les 2 cas, on rejette l'hypothèse nulle.

# Test de linéarité

- On peut tester formellement l'hypothèse de linéarité (ce qui est équivalent à tester la significativité de l'âge) : $H_0 : \beta = 0$

  ```
  >79.2-2.4 = 76.8
  ```

- La statistique vaut 76.8 pour 1 degré de liberté, elle est donc très significative. Ceci indique qu'on ne peut pas rejeter l'hypothèse de linéarité.
- Ca implique aussi qu'on peu, si on le souhaite, modéliser l'âge par une variable quantitative et ainsi gagner 2 degrés de liberté.



# Outline

# Données

- On modélise l'évolution de la probabilité d'utiliser une contraception en fonction de l'âge et du désir d'enfant.

$$Y_{ij} \sim \mathcal{B}(n_{ij}, \pi_{ij})$$

$i \in \{1, \cdots, 4\}$ et $j \in \{1, 2\}$

| Age $i$ | Desires $j$ | Using $y_{ij}$ | Not Using $n_{ij} - y_{ij}$ | All $n_{ij}$ |
|---|---|---|---|---|
| <25 | Yes | 58 | 265 | 323 |
| | No | 14 | 60 | 74 |
| 25-29 | Yes | 68 | 215 | 283 |
| | No | 37 | 84 | 121 |
| 30-39 | Yes | 79 | 230 | 309 |
| | No | 158 | 145 | 303 |
| 40-49 | Yes | 14 | 43 | 57 |
| | No | 79 | 58 | 137 |
| Total | | 507 | 110 | 1607 |

# Table de déviance (1/3)

- En construisant successivement plusieurs modèles logistiques, on construit une table de déviance

| Modèle | Notation | $logit(\pi_{ij})$ | Déviance | dll |
|---|---|---|---|---|
| Nul | $\Phi$ | $\eta$ | 145.7 | 7 |
| Age | $A$ | $\eta + \alpha_i$ | 66.5 | 4 |
| Désir | $D$ | $\eta + \beta_i$ | 54.0 | 6 |
| Additif | $A + D$ | $\eta + \alpha_i + \beta_i$ | 16.8 | 3 |
| Saturated | $AD$ | $\eta + \alpha_i + \beta_i + (\alpha\beta)_{ij}$ | 0 | 0 |

- On note que le modèle nul ne permet pas de décrire les données : la déviance de 145.7 avec 7 ddl est très largement supérieure à 14.1, le quantile de la loi du chi2 à 7 ddl.
- L'introduction de l'âge réduit la déviance à 66.5 pour 4 ddl. La différence 145.7-66.5 = 79.2 pour 3 dll est une statistique de test pour l'effet âge. L'effet est très significatif. On obtient exactement le même résultat que pour les données contraception-âge de l'exemple 2.
- Cette équivalence souligne une propriété des données binomiales : l'effet de l'âge sur l'utilisation d'une contraception est complètement contenu dans les distributions marginales contraception-âge.
- Les déviances elles-mêmes varient car elles dépendent du contexte. Seule la différence reste identique.

# Table de déviance (2/3)

| Modèle | Notation | $logit(\pi_{ij})$ | Déviance | dll |
|---|---|---|---|---|
| Nul | $\Phi$ | $\eta$ | 145.7 | 7 |
| Age | $A$ | $\eta + \alpha_i$ | 66.5 | 4 |
| Désir | $D$ | $\eta + \beta_i$ | 54.0 | 6 |
| Additif | $A + D$ | $\eta + \alpha_i + \beta_i$ | 16.8 | 3 |
| Saturated | $AD$ | $\eta + \alpha_i + \beta_i + (\alpha\beta)_{ij}$ | 0 | 0 |

- Le modèle basé sur le désir d'enfant indique de nouveau un effet significatif du désir d'enfant sur l'utilisation d'une contraception.
- Le fait que la différence de déviance pour le désir d'enfant est égale à 91.7 pour 1 ddl alors qu'elle est seulement de 79.2 pour 3 ddl pour l'âge indique que l'effet du désir d'enfant est plus fort que celui de l'âge.
  Cependant cette comparaison est informelle : les modèles ne sont pas emboités.

# Table de déviance (3/3)

| Modèle | Notation | $logit(\pi_{ij})$ | Déviance | dll |
|---|---|---|---|---|
| Nul | $\Phi$ | $\eta$ | 145.7 | 7 |
| Age | $A$ | $\eta + \alpha_i$ | 66.5 | 4 |
| Désir | $D$ | $\eta + \beta_i$ | 54.0 | 6 |
| Additif | $A + D$ | $\eta + \alpha_i + \beta_i$ | 16.8 | 3 |
| Saturated | $AD$ | $\eta + \alpha_i + \beta_i + (\alpha\beta)_{ij}$ | 0 | 0 |

- Dans le modèle additif, on suppose $\alpha_1 = \beta_1 = 0$ et on interprète les coefficients de la façon suivante
  - $\eta$ est la probabilité qu'une femme de moins de 25 ans désirant plus d'enfant utilise une contraception ;
  - $\alpha_i$ , $i = 2, 3, 4$ représente l'effet net de la tranche d'âge comparé à la tranche -25 ans dans la même catégorie de désir d'enfant ;
  - $\beta_2$ représente l'effet net du désir d'enfant dans la même tranche d'âge.
- On peut comparer les déviances des modèles à un facteur à celle du modèle additif. Par exemple, quand on passe du modèle $D$ au modèle $A + D$, la déviance décroit de 37.2 et on perd 3 ddl. C'est la statistique de test observée de l'effet net de l'âge après avoir pris en compte l'effet du désir d'enfant

$$H_0 : \alpha_i = 0$$

et le test est très significatif.

# Paramètres du modèle additif

- La table de déviance permet de choisir le modèle.
- Ensuite, on doit interpréter les paramètres.
- Les paramètres du modèle additif sont donnés ci-dessous.

| Paramètre | Symbole | Estimation | Std err | $z$-ratio |
|---|---|---|---|---|
| Constant | $\eta$ | -1.694 | 0.135 | -12.53 |
| Age (25-29) | $\alpha_2$ | 0.368 | 0.175 | 2.10 |
| Age (30-39) | $\alpha_3$ | 0.808 | 0.160 | 5.06 |
| Age (40-49) | $\alpha_4$ | 1.023 | 0.204 | 5.01 |
| Désir (No) | $\beta_2$ | 0.824 | 0.117 | 7.04 |

- Les estimations des $\alpha_i$ montrent un effet monotone de l'âge. Cet effet pourrait varier avec la modalité de "Désir".
- De même $\hat{\beta}_2$ montre un fort effet du désir d'enfant. Cet effet pourrait lui aussi varier avec la tranche d'âge.

# Modèle avec interaction

- On considère maintenant le modèle qui inclut une interaction

$$logit(\pi_{ij}) = \eta + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

avec $\alpha_1 = \beta_1 = (\alpha\beta)_{1j} = (\alpha\beta)_{i1} = 0$)

- On interprète $(\alpha\beta)_{i2}$ pour $i = 2, 3, 4$ comme l'effet additionnel ne pas désirer d'enfant, comparé à celui de désirer un autre enfant, pour les femmes du groupe $i$ comparé aux femmes de moins de 25 ans.
On interprète parfois $\beta_2 + (\alpha\beta)_{i2}$ pour simplifier.

```
Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           -1.5193     0.1450 -10.480  < 2e-16 ***
age25-29               0.3682     0.2009   1.832  0.06691 .
age30-39               0.4507     0.1950   2.311  0.02082 *
age40-49               0.3971     0.3402   1.168  0.24298
nomoremore             0.0640     0.3303   0.194  0.84637
age25-29:nomoremore    0.2672     0.4091   0.653  0.51366
age30-39:nomoremore    1.0905     0.3733   2.921  0.00349 **
age40-49:nomoremore    1.3672     0.4834   2.828  0.00468 **
```

- Les résultats indiquent que l'usage d'une contraception, parmi les femmes qui désirent des enfants, varie peu avec l'âge. En revanche, l'effet de ne pas vouloir d'enfant augmente beaucoup avec l'âge.
- On peut résumer les résultats : la contraception utilisée pour espacer les naissances ne varie pas beaucoup avec l'âge au contraire de l'usage de la contraception pour limiter la fertilité qui augmente très fortement avec l'âge.

# Analyse de covariance

- Le modèle avec interaction est saturé, ce qui signifie que le modèle reproduit les données. Si on considère l'âge comme une variable continue, on peut espérer trouver un modèle plus parcimonieux.
- Les modèles proposés sont listés dans le tableau ci-dessous

| Modèle | Notation | $logit(\pi_{ij}$ | Déviance | ddl |
|---|---|---|---|---|
| Une ligne | X | $\alpha + \beta x_i$ | 68.88 | 6 |
| Lignes parallèles | X+D | $\alpha_j + \beta x_i$ | 19.99 | 5 |
| Deux lignes | XD | $\alpha_j + \beta_j x_i$ | 9.14 | 4 |

- Le dernier modèle inclut une interaction entre le désir d'enfant et l'âge, il permet donc à l'effet du désir d'enfant de varier avec l'âge.
- La réduction de déviance (de 9.9 pour 1 ddl) correspond au test (même pente)

$$H_0 : \beta_1 = \beta_2$$

Ce test est rejeté avec une p-value de 0.002.

- La déviance du dernier modèle, 9.14 pour 4 ddl, est juste en dessous du seuil critique 9.49 au risque 5%. On ne peut donc pas rejeter ce modèle.

# Modèle XD, choix de la paramétrisation

- Ici, il est utilise de réfléchir au choix de la paramétrisation du modèle.
- L'application directe de la méthode dans laquelle on choisit une modalité de référence conduit à une paramétrisation à 4 variables : une variable égale à 1, une variable $x$ pour l'âge, une variable $d$ qui prend la valeur 1 pour les femmes qui ne veulent plus d'enfant et une variable dx égale au produit de $d$ par $x$.
- On obtient alors les paramètres suivants : constante et pente pour les femmes qui veulent un autre enfant et la *différence* en constante et pente pour les femmes qui ne veulent plus d'enfant. La différence peut-être difficile à interpréter.
- L'alternative consiste à définir les variables : $d$, $1-d$, $dx$ et $(1-d)x$.

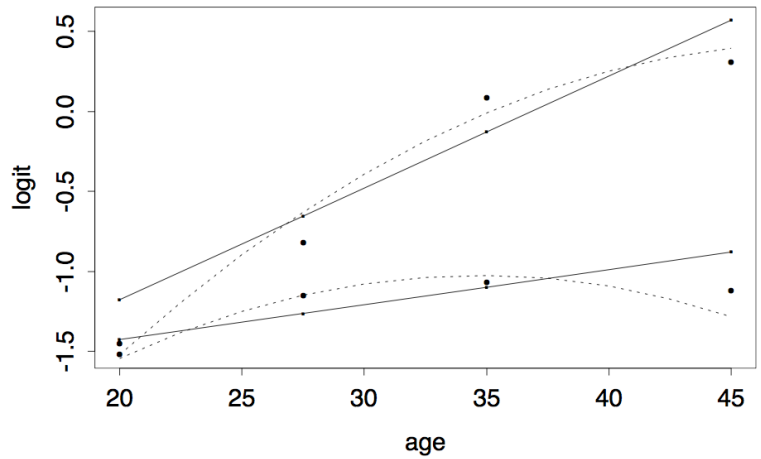
```
> cuse$agem <- c(20, 27.5, 35, 45)[as.numeric(cuse$age)]
> cuse$agec <- cuse$agem - 30.6
> ancova = list(
    + one       = glm(Y ~ agem, family=binomial, data=cuse),
parallel = glm(Y ~ agem+nomore, family=binomial, data=cuse),
    + two       = glm(Y ~ agec*nomore, family=binomial, data=cuse) )
> cuse$more <- 1 - cuse$nomore
> cuse$age.more   <- cuse$agec * cuse$more
> cuse$age.nomore <- cuse$agec * cuse$nomore
> alt <- glm(Y ~ more + age.more + nomore + age.nomore - 1,
    +   family=binomial, data=cuse)
```

# Modèle XD, paramètres

- On obtient finalement les estimations suivantes

| Désir | Age | Symbole | Estimation | Std. err. | P(>\|z\|) |
|---|---|---|---|---|---|
| More | Constante | $\alpha_1$ | -1.194 | 0.079 | < 2e-16 *** |
| | Pente | $\beta_1$ | 0.0218 | 0.0103 | 0.035 * |
| Nomore | Constante | $\alpha_2$ | -0.437 | 0.093 | 2.70e-06 *** |
| | Pente | $\beta_2$ | 0.06981 | 0.01144 | 1.04e-09 *** |

- On retrouve que l'effet de l'âge est d'autant plus fort que la femme ne veut plus d'enfant.
- En comparant les logit théoriques aux logits empiriques, on observe qu'un modèle quadratique serait pertinent.
- En effet, l'introduction d'un effet quadratique de l'âge en interaction avec le désir d'enfant conduit à un très bon ajustement.

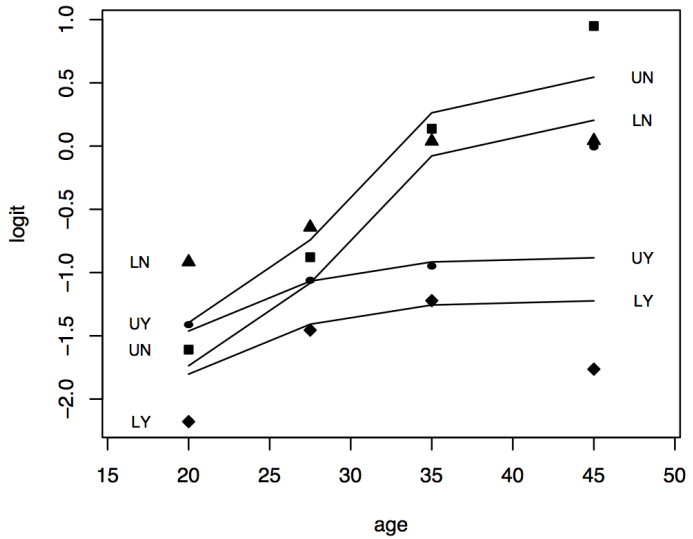


# Modèle multifacteurs

- G. Rodriguez présente une analyse complète des données de contraception dans ses notes de cours `http://data.princeton.edu/wws509/notes/c3.pdf`, section 3.6.
- Il propose deux approches
  1. Une sélection "forward" qui consiste à partir du modèle nul et à le complexifier petit à petit en ajoutant des effet individuels puis des interactions.
  2. Une sélection "backward" qui consiste à partir du modèle contenant toutes les interactions et à éliminer pas à pas les effets non significatifs.

  Dans les deux cas, le modèle retenu est le modèle avec une interaction entre l'âge et de désir d'enfant + un effet de l'éducation.
- Dans le graphique ci dessous L="low education", U="Upper education", Y = desir for more children, N = no desir for more children.

On observe l'interaction Y/N-age et pas d'interaction L/U-age (lignes parallèles). L'effet du non désir d'enfant augmente fortement avec l'âge (pente forte).



# Autres critères basés sur la vraisemblance

- Pour comparer des modèles, il est courant d'utiliser des critères basés sur une pénalisation de la vraisemblance.
- Le critère d'Akaike est défini par

$$AIC = -2 \log L + 2k$$

avec $k$ le nombre de paramètres à estimer.

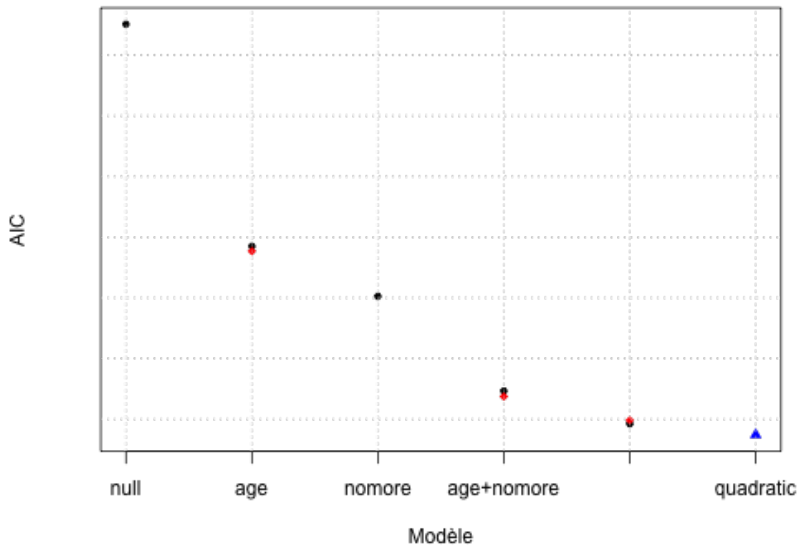
- Si l'on considère un ensemble de modèles candidats, le modèle choisi est celui qui aura la plus faible valeur d'$AIC$. Ce critère repose donc sur un compromis entre la qualité de l'ajustement et la complexité du modèle.
- Le critère "Bayes Information criterion" est défini par

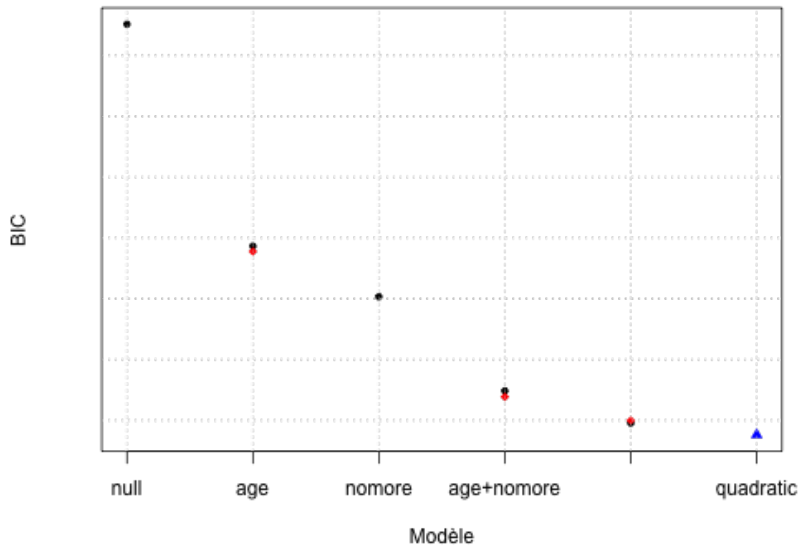
$$BIC = -2 \log L + \log(n) k$$

avec $n$ le nombre d'observations.

- L'AIC pénalise le nombre de paramètres moins fortement que le BIC.

- Comparaison des différents modèles pour les données de contraception.
- En noir pour l'âge sous forme de variable discrète, en rouge pour l'âge en variable continue.
- Les résultats des modèles BIC et AIC sont très proches pour ces modèles.
- Ils conduisent à choisir le modèle quadratique.



# Outline

# Diagnostiques de régression

- Les diagnostiques sont aussi importants pour la régression logistique que pour la régression classique.
- Les diagnostiques se basent aussi sur les résidus ie les différences entre les valeurs observées et les valeurs prédites.
- La principale différence par rapport au modèle linéaire est que pour les données binaires, on ne suppose plus que les données sont identiquement distribuées.
- On utilise deux types de résidus : les résidus de Pearson et les résidus basés sur la déviance.

# Résidus de Pearson

- L'approche la plus simple pour obtenir des résidus est de calculer la différence entre les valeurs observées et les valeurs prédites et de diviser par l'écart-type des valeurs observées

$$p_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i(n_i - \hat{\mu}_i)/n_i}}$$

  où la $\mu_i$ sont les valeurs prédites et le dénominateur est donné par $var(y_i) = n_i\pi_i(1-\pi_i) \simeq n_i\frac{\hat{\mu}_i}{n_i}(1-\frac{\hat{\mu}_i}{n_i})$ .
- Ces résidus sont appelés résidus de Pearson parce que la racine carrée de $p_i$ est la contribution de l'individu à la statistique du chi2 de Pearson.
- Pour les données groupées, les résidus de Pearson suivent approximativement une loi de Gauss. Ce n'est pas vrai pour les données individuelles.
- Dans les deux cas, un individu qui a un résidu $|p_i| > 2$ doit nécessiter une attention particulière.

# Résidus basés sur la déviance

- Une alternative consiste à considérer les résidus basés sur la déviance.
- Dans ce cas, on s'intéresse au rapport de $y_i/\hat{\mu}_i$ à la place de la différence $y_i - \hat{\mu}_i$.
- Les résidus sont alors définis par

$$d_i = \sqrt{2\left(y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{\hat{n}_i - \mu_i}\right)\right)}$$

- Si on prend le carré et qu'on somme tout ces résidus, on obtient la déviance.
- Les observations telles que $d_i > 2$ peuvent indiquer un défaut d'ajustement.

# Exemple

```
> additive <- glm(Y ~ age + education + nomore, family=binomial,
+ data=cuse)
> dr <- residuals(additive)
> sum(dr^2)
[1] 29.91722 # Deviance
> pr <- residuals(additive, type="pearson")
> sum(pr^2)
[1] 28.28834
> cbind(
+   cuse[,c("age","education","nomore")],
+   obs = cuse$using/cuse$n, fit = fitted(additive),
+   dr, pr)[abs(dr)>2,]
      age educ.   nomore obs fit dr pr
4     <25 high     1 0.167  0.308  -2.515 -2.375
8   25-29 high     1 0.293  0.397  -2.065  -2.026
13  40-49 low     0 0.146   0.315  -2.491  -2.325
```

- Le modèle additif est particulièrement mauvais pour les jeunes femmes éduquées qui veulent plus d'enfants et pour les femmes plus agées qui ont un faible niveau d'éducation.

# Résidus studentisés

- Les résidus définis ci-dessous ne sont pas standardisés.
- Ils prennent en compte le fait que les observations ont des variances différentes, mais ils ne tiennent pas compte de la variance supplémentaire due à l'estimation des paramètres ; dans le cas du modèle linéaire les résidus studentisés incluent cette variance supplémentaire.
- On peut approcher la variance des résidus

$$Var(y_i - \hat{\mu}_i) \approx (1 - h_{ii}) Var(y_i)$$

avec $h_{ii}$ l'effet levier qui est aussi le terme diagonal de la matrice

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}$$

et $\mathbf{W}$ une matrice diagonale de terme général $w_{ii} = \mu_i(n_i - \mu_i)/n_i$ évaluée à l'estimateur maximum de vraisemblance.

- Ainsi, un résidu studentisé est donné par

$$s_i = \frac{p_i}{\sqrt{1 - h_{ii}}}$$

- L' effet levier $h_{ii}$ mesurent l'impact de $y_i$ dans l'estimation de $\hat{\mu}_i$
- Un point $i$ a un effet levier si $H_{ii} > 2(p+1)/n$

# Effet levier et influence

- Les éléments diagonaux de la matrice **H** peuvent être interprétés comme des effets levier ie qu'ils mesurent l'importance du rôle que joue $y_i$ dans l'estimation de $\hat{\mu}_i$.
- On pourrait calculer les distances de Cook qui comparent l'estimateur du maximum de vraisemblance $\widehat{\boldsymbol{\beta}}$ à $\widehat{\boldsymbol{\beta}_{(i)}}$, l'estimateur obtenu sans l'observation $i$. Ce calcul est coûteux.
- Mais on peut approcher ces distances par

$$D_i = \frac{1}{p}(\widehat{\boldsymbol{\beta}_{(i)}} - \widehat{\boldsymbol{\beta}})^T \mathbf{X}^T \mathbf{W} \mathbf{X} (\widehat{\boldsymbol{\beta}_{(i)}} - \widehat{\boldsymbol{\beta}})^T \simeq p_i^2 \frac{h_{ii}}{(1-h_{ii})^2 p}$$

  en faisant juste une itération de l'algorithme IRLS sans l'observation $i$ en partant de $\widehat{\boldsymbol{\beta}}$.
- La distance de Cook mesure l'influence d'une observation sur l'ensemble des prévisions en prenant en compte l'effet levier et l'importance des résidus.

# Exemple

```
> obs <- cuse$Y
> p <- fitted(additive)
> pfit <- p # for clarity
> pobs <- cuse$using / cuse$n
> lev  <- hatvalues(additive)
> cd  <- cooks.distance(additive)
> i <- order(-lev) # sort descending
> cbind(cuse[,c("age","education","nomore")],
     + pobs,pfit,lev,cd)[i[1:3],]
age   educ nomore obs       pfit         lev            cd
3 <25      high 0   0.1969697 0.1623053  0.6696332 2.385867366
7 25-29 high 0   0.2583732 0.2223899  0.5774811 0.843656898
14   40-49 low 1   0.5106383 0.5140024  0.5601446 0.002054933
```

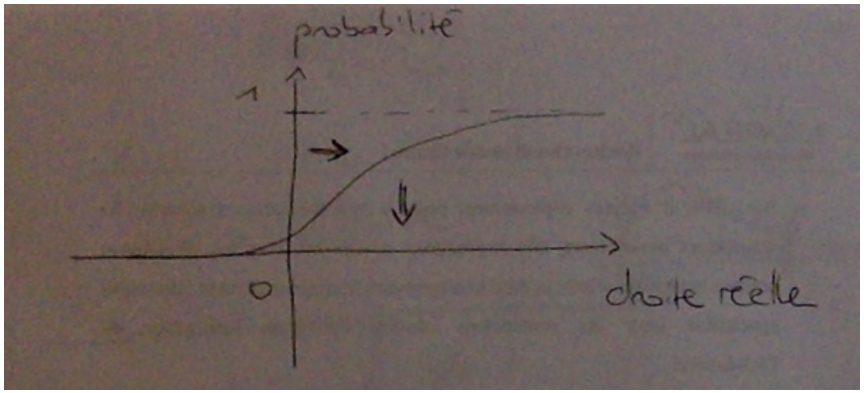
- Les 3 cellules qui ont potentiellement le plus d'influence sur l'estimation sont les jeunes femmes avec un haut niveau d'éducation et qui veulent des enfants, et les femmes plus âgées avec un faible niveau d'éducation et qui ne veulent plus d'enfant.
- Le groupe le plus jeune a le plus d'influence.
- Le groupe des femmes plus âgées n'a pas d'influence : distance de Cook nulle.

# Outline

# Outline

# Fonctions de lien

- Tous les modèles étudiés jusqu'ici ont une fonction de lien logit.
- En réalité, toutes les transformations qui envoient une probabilité sur la droite réelle peuvent être utilisées, sous réserve qu'elles soient bijectives, continues et différentiables.




- En particulier, si $F$ est la fonction de répartition d'une variable aléatoire définie sur tout $\mathbb{R}$, on peut écrire

$$\pi_i = F(\eta_i)$$

ou de façon équivalente

$$\eta_i = F^{-1}(\pi_i)$$

- Les choix les plus courants sont la loi de Gauss, la loi logistique et les lois de valeurs extrèmes.

# Formulation avec variable latente

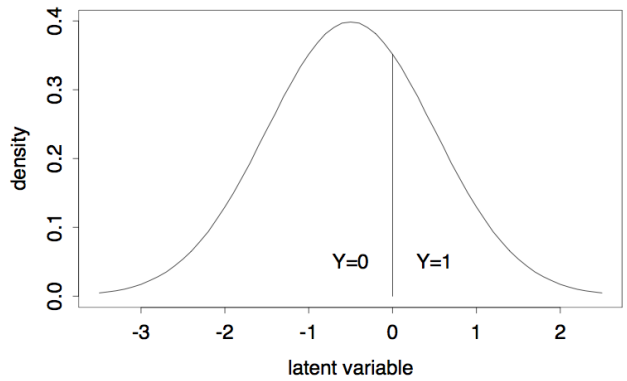
- On note $Y_i$ une variable aléatoire définie sur $\{0, 1\}$.
- Supposons qu'il existe une variable continue $Y_i^*$ (définie sur $\mathbb{R}$) non observable telle que

$$Y_i = \begin{cases} 1 & \text{si} \quad Y_i^* > c \\ 0 & \text{sinon} \end{cases}$$

avec $c$ un certain niveau.

- On dit que $Y_i^*$ est la réponse latente.
- Les biostatisticiens interprètent souvent $Y_i^*$ comme une dose et $Y_i$ comme une réponse. On parle parfois de modèle dose-réponse.
- On définit

$$\pi_i = P(Y_i = 1) = P(Y_i^* > c)$$



# Modèle dose-réponse

- Le modèle définit ci-dessus ne change pas si on multiplie $Y_i^*$ et $c$ par une même constante ou si on leur ajoute une même constante. Ainsi, pour que le modèle soit identifiable on suppose que $Y_i^*$ est une variable aléatoire centrée et réduite et on pose $c = 0$.
- Quand la réponse dépend d'un vecteur de covariables $\mathbf{x}$, on peut supposer

$$Y_i^* = \mathbf{x}_i^T \beta + U_i$$

avec $U_i$ un terme d'erreur qui a une fonction de répartition $F$.

- Avec ce modèle, la probabilité d'observer $Y_i = 1$

$$\pi_i = P(Y_i = 1) = P(U_i > -\eta_i) = 1 - F(-\eta_i)$$

avec $\eta_i = \mathbf{X_i}^T \beta$ el prédicteur linéaire.

- Si la distribution de l'erreur est symétrique autour de 0, $F(u) = 1 - F(-u)$ et

$$\pi_i = F(\eta_i)$$

- Cette expression définit un modèle linéaire généralisé avec une réponse de Bernoulli et le lien $\eta_i = F^{-1}(\pi_i)$

# Modèle Probit

- Le modèle probit correspond au cas où $U_i \sim \mathcal{N}(0,1)$.
- On parle alors de lien probit pour $\eta_i = \Phi^{-1}(\pi)$
  avec $\Phi$ la fonction de répartition de la loi de Gauss standard.
- On peut aussi considérer le cas plus général tel que $U_i \sim \mathcal{N}(0,\sigma^2)$

$$\begin{aligned}\pi_i &= P(Y_i^* > 0) = P(U_i > -\mathbf{x}_i^T\boldsymbol{\beta}) = P\left(\frac{U_i}{\sigma} > -\frac{\mathbf{x}_i^T\boldsymbol{\beta}}{\sigma}\right)\\ &= 1 - \Phi\left(1 - \frac{\mathbf{x}_i^T\boldsymbol{\beta}}{\sigma}\right) = \Phi\left(\frac{\mathbf{x}_i^T\boldsymbol{\beta}}{\sigma}\right)\end{aligned}$$

- On remarque qu'on ne peut pas identifier $\boldsymbol{\beta}$ et $\sigma$ séparément. On fixe donc $\sigma = 1$.
- Il y a un petit inconvénient à utiliser la distribution de Gauss pour la fonction de lien : elle n'a pas de forme explicite.

# Exemple des données de contraception

- Considérons un modèle probit avec comme covariables l'âge (v.a. continue) et le désir d'enfant et avec une interaction.

```
> probit.model <- glm(Y ~ agec * nomore,
+ family=binomial(link="probit"), data=cuse)
```

- On obtient les esdtimations suivantes

| Paramètre | Symbol | Estimation | Std. Err. | $P(>\|z\|)$ |
|---|---|---|---|---|
| Constante | $\alpha_1$ | -0.72 | 0.046 | < 2e-16 |
| Age | $\beta_1$ | 0.01 | 0.006 | 0.033 |
| Désir | $\alpha_2 - \alpha_1$ | 0.46 | 0.073 | 3.94e-10 |
| Age x Désir | $\beta_2 - \beta_1$ | 0.03 | 0.009 | 0.0009 |

- Pour interpréter ce modèle, on imagine une variable latente continue qui mesure la motivation de la femme à utiliser une contraception.
- Pour un âge moyen (30.6 ans), le fait de ne plus vouloir d'enfant accroit la motivation de presque 0.5 écart-type.
- Chaque année d'âge est associée à un accroissement de la motivation de 0.01 écart-type si la femme veut encore des enfants et à 0.03 si elle n'en veut plus.

# Modèle Logistique

- Une alternative au modèle Gaussien est le modèle logistique.

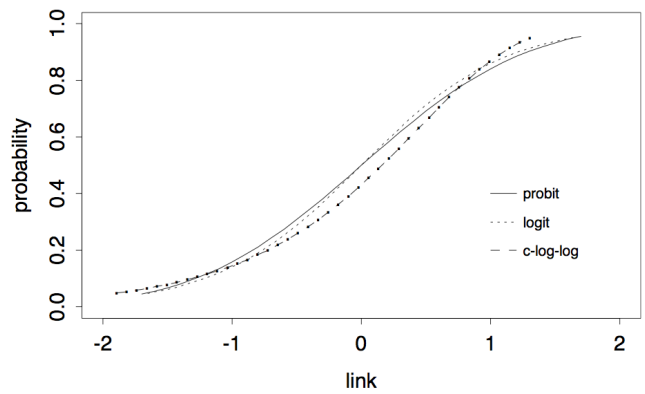
$$\pi_i = F(\eta_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}}$$

pour $-\infty < \eta_i < +\infty$

- La transfomatiopn inverse est la logit.

$$\eta_i = F^{-1}(\pi_i) = \log \frac{\pi_i}{1 - \pi_i}$$

- La distribution logistique standard est symétrique, elle a pour moyenne 0 et pour variance $\pi^2/3$. Sa forme est très proche de celle de la loi normale mais elle a des queues plus lourdes.
- Les deux sont presque linéaires entre 0.1 et 0.9 et elles conduisent à des résultats très proches.
- La valeurs des coefficients changent à cause de la variance standard qui change.



# Transformation log-log complémentaire

- Un autre choix est la transformation c-log-log

$$\eta_I = \log\left(-\log(1-\pi_i)\right).$$

- C'est la transformation inverse de la distribution de valeurs extrèmes (Gumbel)

$$F(\eta_i) = 1 - e^{-e_i^{\eta}}.$$

- Pour les petites valeurs de $\pi_i$ cette transformation est proche du logit.
- La transformation log-log correspond au cas où $-U_i$ a une distribution de valeur extrême reverse standard

$$F(u_i) = e^{-e^{-u_i}}.$$

Cette distribution est disymétrique avec une queue lourde vers à droite.

# Outline

# Quand la variable à prédire est multinomiale

- Dans certains problèmes, la variable à prédire est multinomiale.
- Dans l'exemple des données de diabète, la variable $Y$ prend 3 modalités "Normal", "Chemical Diabet" et "Overt Diabet".
  Dans l'exemple de données de contraceptions ci dessous, $Y$ prend les valeurs "Stérilisation", "Autre" et "Aucune".

| Age | Ster. | Autre | Aucune | Total |
|---|---|---|---|---|
| 15-19 | 3 | 61 | 232 | 296 |
| 20-24 | 80 | 137 | 400 | 617 |
| 25-29 | 216 | 131 | 301 | 648 |
| 30-34 | 268 | 76 | 203 | 547 |
| 40-44 | 150 | 24 | 164 | 338 |
| 45-49 | 91 | 10 | 183 | 284 |
| Total | 1005 | 489 | 1671 | 3165 |

Données du Salvador (1985) pour des femmes mariées.

- On cherche à modéliser les probabilités de $Y$ d'appartenir à chacun des groupes sachant la classe d'âge.

# Quand la variable à prédire est multinomiale

- La variable à prédire $Y_i$ prend alors ses valeurs dans $\{1, 2, \cdots, J\}$ avec $J \geq 2$ et on définit
$$\pi_{ij} = P(Y_i = j) \text{ avec } \sum_{j=1}^{J} \pi_{ij} = 1$$
- La contrainte $\sum_{j=1}^{J} \pi_{ij} = 1$ implique que le modèle a $J - 1$ vecteurs de paramètres.
- Loi de $Y$
$$P(Y_{ij} = y_{ij}, \cdots, Y_{iJ} = y_{iJ}) = \left( \begin{array}{c} n_i \\ y_{i1}, \cdots, y_{iJ} \end{array} \right) \pi_{i1}^{y_{i1}} \cdots \pi_{iJ}^{y_{iJ}}$$
- Le cas particulier où $J = 2$ correspond à la loi binomiale.

# Modèle logistique multinomial

- L'approche usuelle pour les données multinomiales consiste à choisir une catégorie de référence et à modéliser les log-odds pour les autres modalités.

$$\eta_{ij} = \log \frac{\pi_{ij}}{\pi_{iJ}} = \alpha_j + \mathbf{x}_i^T \boldsymbol{\beta}_j$$

avec $\alpha$ une constante et $\boldsymbol{\beta}_j$ un vecteur de paramètres, $j = 1, \cdots, J-1$.

- On obtient donc le modèle suivant

$$P(Y_i = j) = \frac{e^{\alpha_j + \mathbf{x}_i^T \boldsymbol{\beta}_j}}{1 + \sum_{j=1}^{J-1} e^{\alpha_k + \mathbf{x}_i^T \boldsymbol{\beta}_k}}$$

$$P(Y_i = J) = \frac{1}{1 + \sum_{j=1}^{J-1} e^{\alpha_k + \mathbf{x}_i^T \boldsymbol{\beta}_k}}$$

- On peut vérifier que

$$\sum_{j=1}^{J} P(Y_i = j) = 1$$

et on a bien

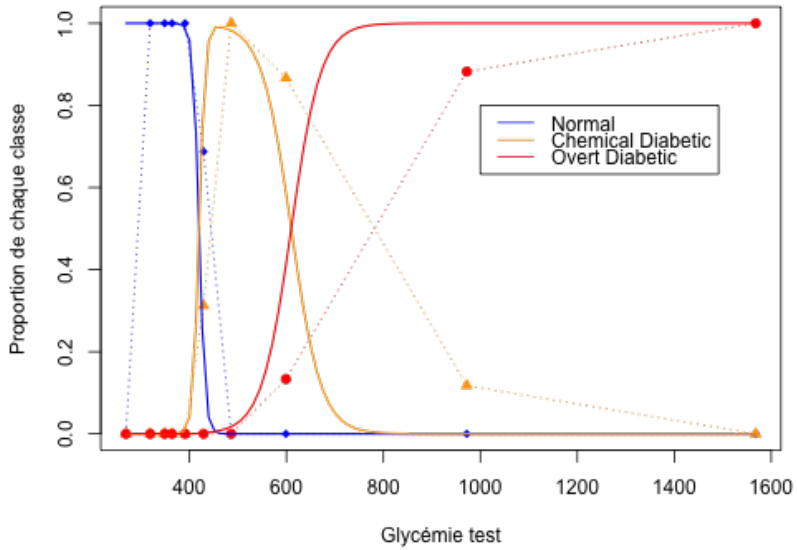
$$\eta_{ij} = \log \frac{P(Y_i = j)}{P(Y_i = J)} = \alpha_j + \mathbf{x}_i^T \boldsymbol{\beta}_j$$

# Exemple pour les données de diabète

- Pour les données de diabète, on trace en pointillés les probabilités empiriques et en trait plein les probabilités théoriques associées au modèle multinomial.

$$\log \frac{P(Y_i = "CD")}{P(Y_i = "No")} = -67.25 + 0.16 * ga$$

$$\log \frac{P(Y_i = "OD")}{P(Y_i = "No")} = -87.12 + 0.19 * ga$$




- Le modèle sous estime un peu les probabilités de "Chemical Diabetic" par rapport aux observations. mais on retrouve bien les tendances.

# Données de contraception

- Pour les modèles à réponse multinomiale, on considère un autre jeu de données de contraception.

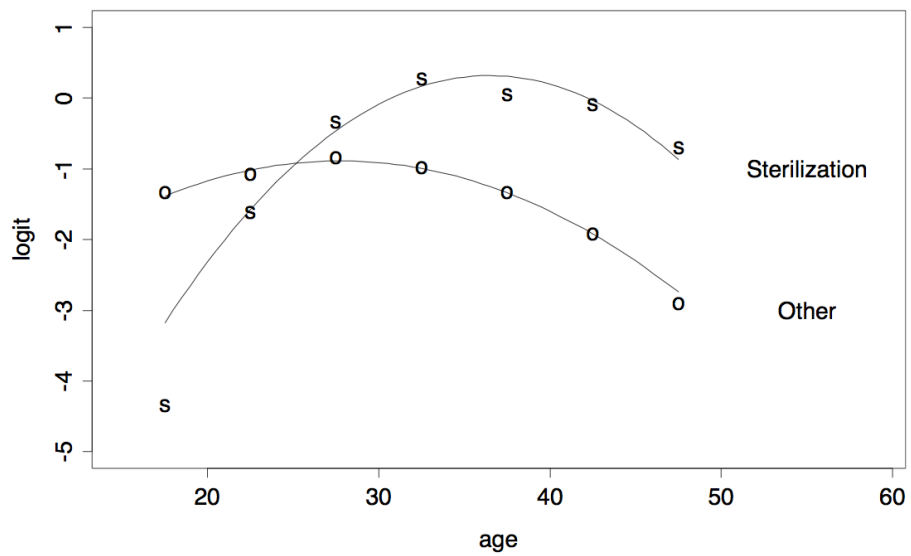
TABLE 6.1: Current Use of Contraception By Age Currently Married Women. El Salvador, 1985

| Age | Contraceptive Method | | | All |
|---|---|---|---|---|
| | Ster. | Other | None | |
| 15–19 | 3 | 61 | 232 | 296 |
| 20–24 | 80 | 137 | 400 | 617 |
| 25–29 | 216 | 131 | 301 | 648 |
| 30–34 | 268 | 76 | 203 | 547 |
| 35–39 | 197 | 50 | 188 | 435 |
| 40–44 | 150 | 24 | 164 | 338 |
| 45–49 | 91 | 10 | 183 | 284 |
| All | 1005 | 489 | 1671 | 3165 |

# Modèle logistique multinomial

- On peut aussi définir des modèles non linéaires.
- Dans l'exemple des données de contraception, les logits sont des fonctions quadratiques de l'âge. On pose le modèle

$$\eta_{ij} = \alpha_j + \beta_j a_i + \gamma_j a_i^2$$



# Inférence

- Pour estimer les paramètres du modèle par maximum de vraisemblance, on maximise la vraisemblance

$$P(Y_{ij} = y_{ij}, \cdots, Y_{iJ} = y_{iJ}) = \begin{pmatrix} n_i \\ y_{i1}, \cdots, y_{iJ} \end{pmatrix} \pi_{i1}^{y_{i1}} \cdots \pi_{iJ}^{y_{iJ}}$$

  avec les probabilités $\pi_{ij}$ qui dépendent des paramètres $\alpha_j$ et $\boldsymbol{\beta}_j$.
- L'optimisation requiert d'utiliser des procédures numériques. Les algorithmes de Fisher scoring et de Newton-Raphson sont généralement les plus performants.

# Exemple

- Pour les données de contraception, l'estimation par maximum de vraisemblance conduit à une déviance de 20.5 pour 8 ddl. La p-value est égale à 0.009 ce qui montre un mauvais ajustement.

```
> cuse$Y <- as.matrix(cuse[,c("none","ster","other")])
> msat <- multinom(Y ~ ageg, data=cuse); msat
```

- L'effet quadratique a un rapport de vraisemblance de 500.6 pour 4 ddl ce qui est très significatif.
- Les paramètres estimés sont

| | Contraste | |
|---|---|---|
| Paramètres | Ster. vs None | Other vs None |
| Constante | -12.62 | -4.552 |
| Linéaire | 0.7097 | 0.2641 |
| Quadratique | -0.0097 | -0.0047 |

Ils sont utilisés pour tracer la figure vue plus haut. Elle montre que l'ajustement n'est pas si mauvais à l'exception du groupe d'âge 15-19 pour lequel la probabilité de stérilisation est sur-estimée.

- On pourrait pousser plus loin en ajoutant un effet cubique. Vous pouvez essayer car ce modèle devrait passer les test !

# Outline

# Modèle logistique conditionnel

- Le modèle logistique conditionnel est une extension du modèle multinomial particulièrement bien adapté pour modéliser des comportements face à un choix.
- Les variables explicatives peuvent contenir des attributs du choix (prix) et des informations individuelles (salaire).
- Soit $Y_i$ une variable discrète qui représente un choix parmi $J$. Soit $U_{ij}$ la valeur ou l'utilité du $j$ème choix pour le $i$ème individu.
- Les $U_{ij}$ sont des variables indépendantes telles que

$$U_{ij} = \eta_{ij} + \epsilon_{ij}$$

  avec $\eta_{ij}$ un effet fixe et $\epsilon_{ij}$ un effet aléatoire.
- Les individus sont supposés avoir un comportement rationnel et faire leur choix de façon à maximiser son utilité ie que l'individu $i$ va choisir $j$ si $U_{ij}$ est le plus grand parmi $U_{i1}, \cdots, U_{iJ}$
- Mais $U_{ij}$ comporte un terme aléatoire, on traduit donc l'hypothèse en terme de probabilité

$$\pi_{ij} = P(Y_i = j) = P(U_{ij} = \max(U_{i1}, \cdots, U_{iJ}))$$

# Modèle logistique conditionnel

- On peut montrer que si $\epsilon_{ij}$ suit une loi de valeur extrême de type I (ie Gumbel), de densité

$$f(\epsilon) = \exp\left(-\epsilon - \exp(-\epsilon)\right)$$

alors

$$\pi_{ij} = \frac{\exp(\eta_{ij})}{\sum_{k=1}^{J} \exp(\eta_{ik})}$$

- On reconnait l'équation qui définit le modèle multinomial.
- Si $J = 2$, l'individu $i$ choisit $U_{i1}$ si $U_{i1} - U_{i2} > 0$. Et on peut montrer que la différence de deux utilités aléatoires qui suivent des lois de Gumbel indépendantes a une distribution logistique. On retrouve le modèle logistique standard.

# Quand le modèle logistique conditionnel ne marche pas bien...

- Il existe un cas de figure où le modèle multinomial ne marche pas bien appelé le "bus rouge ou bleu".
- Choix de transport : train/bus rouge/bus bleu.
- Hypothèse : la moitié des gens prennent le train et l'autre le bus ; et les voyageurs qui choisissent le bus sont indifférents à la couleur. Les probabilités de choix sont donc $\pi = (0.50, 0.25, 0.25)$ et l'espérance des utilités est $\eta = (\log 2, 0, 0)$.
- Hypothèse supplémentaire : le service du bus bleu est interrompu. On s'attend à ce que les voyageurs qui prennent le bus prennent tous le bus rouge et donc à une répartition 1 :1. Mais à cause des utilités $\log 2$ et $0$ ont obtient une répartition 2 :1.

# Logits

- Dans le cas des logit multinomials, les utilités moyennes $\eta_{ij}$ sont modélisées en fonction des caractéristiques individuelles

$$\eta_{ij} = \mathbf{x}_i^T \boldsymbol{\beta}_j$$

- Dans le cas des logit conditionnels, on modélise les utilités moyenne en fonction des caractéristiques des alternatives.
- Si $\mathbf{z}_j$ représente un vecteur des caractéristiques de la $j$ème alternative, alors

$$\eta_{ij} = \mathbf{z}_j^T \boldsymbol{\gamma}$$

- On peut combiner les deux modèles

$$\eta_{ij} = \mathbf{x}_i^T \boldsymbol{\beta}_j + \mathbf{z}_{ij}^T \boldsymbol{\gamma}$$

# Outline

# Modèle logistique hiérarchique

- Quand la réponse est multinomiale, on choisit une des modalités qui sert de référence.
- Une alternative consiste à hiérarchiser les modalités (ou les choix) en 2 sous ensembles et d'ajuster une modèle logistique ordinaire pour chaque comparaison.
- Pour les données de contraception, par exemple, on peut considérer
  1. le rapport de cote de "utiliser une forme de contraception" contre "ne pas utiliser de contraception" (1494 :1671)
  2. le rapport de cote de stérilisation contre toutes sortes de contraception (1005 :489).
- Cette approche est utile si on considère que les individus font leur choix de façon séquentielle. Par exemple, pour l'usage d'une contraception, la femme décide d'abord si elle va utiliser une contraception ou non. Puis, elle doit choisir un moyen de contraception.

# Exemple

- La figure ci-dessous montre les logarithmes des rapports de cote empiriques pour
  - l'utilisation d'une contraception contre aucune contraception (u) et
  - la stérilisation contre toutes sortes de contraception (s)

  en fonction de l'âge.
- On observe que l'usage d'une contraception augmente jusqu'à l'âge de 35 ans environ puis décroit. Alors que l'usage spécifique de la stérilisation continue de croître jusqu'à 50 ans.

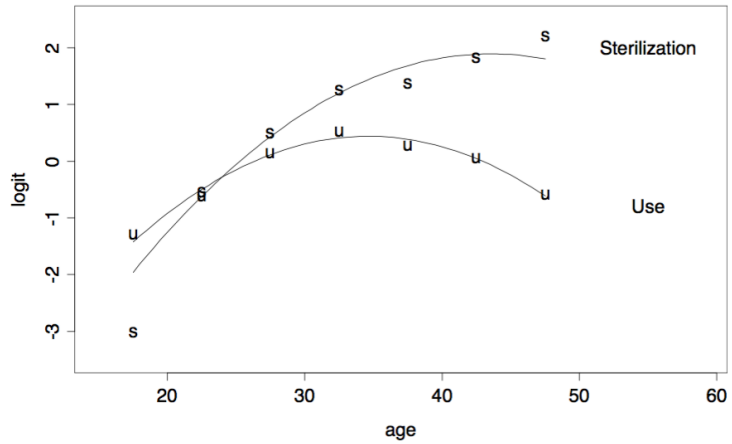



FIGURE 6.2: Log-Odds of Contraceptive Use vs. No Use and Sterilization vs. Other Method, by Age.

- Les données suggèrent donc le modèle

$$\eta_{ij} = \alpha_j + \beta_j a_i + \gamma_j a_i^2$$

où $a_i$ est le centre de la classe d'âge $i$. $j = 1$ pour l'équation correspondant à l'usage d'une contraception et $j = 2$ pour l'équation correspondant au choix de la méthode.

# Inférence

- Une caractéristique importante et intéressante du modèle logistique hiérarchique est que la vraisemblance s'écrit sous la forme d'un produit de vraisemblances binomiales qui peuvent être maximisées séparément.
- Prenons de nouveau l'exemple des données de contraception. La contribution de l'individu $i$ à la vraisemblance est

$$L_i = \pi_{i1}^{y_{i1}} \pi_{i2}^{y_{i2}} \pi_{i3}^{y_{i3}}.$$

où les $\pi_{ij}$ sont les probabilités et les $y_{ij}$ sont les nombres correspondants pour les femmes stérilisées, utilisant une autre méthode et n'utilisant aucune méthode respectivement.

- Ceci s'écrit aussi

$$L_i = \left(\frac{\pi_{i1}}{\pi_{i1}+\pi_{i2}}\right)^{y_{i1}} \left(\frac{\pi_{i2}}{\pi_{i1}+\pi_{i2}}\right)^{y_{i1}} (\pi_{i1}+\pi_{i2})^{y_{i1}+y_{i2}} \pi_{i3}^{y_{i3}}.$$

- En notant $\rho_{i1} = \pi_{i1} + \pi_{i2}$ la probabilité d'utiliser une contraception dans le groupe d'âge $i$ et $\rho_{i2} = \pi_{i1}/(\pi_{i1}+\pi_{i2})$ la probabilité conditionnelle d'être stérilisée sachant qu'on utilise une contraception,

$$L_i = \underbrace{\rho_{i2}^{y_{i1}}(1-\rho_{i2})^{y_{i2}}}_{\text{prob. binomiale}} \underbrace{\rho_{i1}^{y_{i1}+y_{i2}}(1-\rho_{i1})^{y_{i3}}}_{\text{prob. binomiale}}.$$

# Inférence

- On ajuste alors les 2 modèles logistiques séparément.

```
> cuse$A <- cbind(cuse[,"ster"] + cuse[,"other"], cuse[,"none"])
> sla <- glm(A ~ age + agesq, data=cuse, family=binomial)
> cuse$S.A <- cbind(cuse[,"ster"], cuse[,"other"])
> sls <- glm(S.A ~ age + agesq, data=cuse, family=binomial)
> x2 <- deviance(sla) + deviance(sls); x2
[1] 16.89298
> pchisq(x2, 8, lower.tail=FALSE)
[1] 0.03124295
```

- On obtient

| Paramètre | Contraste | |
|---|---|---|
| | Use vs No use | Ster vs Other |
| Constante | -7.180 | -8.869 |
| Linéaire | 0.4397 | 0.4942 |
| Quadratique | -.0063 | -0.0056 |

Le modèle est assez bien ajusté avec très peu de paramètres (p-value = .03).

# Outline

# Réponse ordinale

- Exemples de réponse ordinale
  - Qualité de vie d'un patient (excellente, bonne, assez bonne, mauvaise)
  - Douleur (aucune, faible, forte, sévère)
  - Diagnostique (absolument normal, probablement normal, équivoque, probablement anormal, absolument anormal)
  - Tendance politique (très libéral, légèrement libéral, modéré, légèrement conservatif, très légèrement)

- Les modèles multinomiaux pourraient être appliqués pour les réponses ordinales, mais on perdrait la notion d'ordre.

# Modèles pour réponse ordinale

- Dans le cas d'une réponse ordinale, on va donc construire des modèles spécifiques.
- Nous considérons l'exemple des données de diabètes. La réponse est ordonnée :

  (1) "normal" < (2) "chemical" < (3) "overt".
- Modéliser $Y_i = j$ peut être interprété comme la modélisation d'une variable continue $Y_i^*$ qui prend ses valeurs dans une suite d'intervalles

$$y_i = j \text{ si } \theta_{j-1} \leq y_i^* < \theta_j, \quad j = 2, \cdots, J+1$$

- En pratique, on va s'intéresser à la probabilité cumulée

$$\gamma_{ij} = P(Y_i \leq j) = P(Y_i^* < \theta_j).$$

  Et on va relier les probabilités cumulées aux variables explicatives $X$.
- Supposons que $Y^* = -\mathbf{X}^T\beta + \epsilon$ avec $E(\epsilon) = 0$. La moyenne de $Y^*$ sachant $\mathbf{X}^T = \mathbf{x}^T$ est donc $\mathbf{x}^T\beta$ et on a

$$\gamma_{ij} = P(\epsilon_i \leq \theta_j + \mathbf{X}_i^T\beta)$$

  La distribution de $\epsilon$ détermine la forme du modèle.

# Modèle cumulatif et lien logit (1/2)

- Si la loi de $\epsilon$ est logistique, on a

$$P(\epsilon \leq t) = \frac{1}{1+e^{-t}}$$

d'où

$$\gamma_{ij} = P(\epsilon_i \leq \theta_j + \mathbf{X}_i\boldsymbol{\beta}) = \frac{1}{1+e^{-(\theta_j+\mathbf{X}_i^T\boldsymbol{\beta})}}$$

et

$$\log(\gamma_{ij}) = \theta_j + \mathbf{X}_i^T\boldsymbol{\beta}, \quad j = 1, \cdots, J$$

- Comme on écrit explicitement la constante, on ne suppose plus que la matrice de design inclut une colonne de 1.
- $\theta_j$ est une constante qui représente la valeur de base de la probabilité cumulée transformée pour la catégorie $j$. C'est le seul paramètre qui dépend de $j$. En pratique ceci implique que les rapports de cote sont proportionnels (voir exemple ci-dessous).
- $\boldsymbol{\beta}$ représente l'effet des covariables sur la probabilité cumulée transformée.
- Le rapport de cote de $Y \leq j$ pour deux valeurs de $\mathbf{x}$ est

$$\frac{e^{\theta_j+\mathbf{x_1}^T\boldsymbol{\beta}}}{e^{\theta_j+\mathbf{x_2}^T\boldsymbol{\beta}}} = e^{(\mathbf{x_1}^T-\mathbf{x_2}^T)\boldsymbol{\beta}}$$

L'effet est le même pour toutes les modalités $j$.

# Modèle logistique cumulatif et lien logit (2/2)

- Si la fonction de lien est le logit

$$\begin{aligned} logit(\gamma_{ij}) &= logit(P(Y_i \leq j)) \\ &= \log\left(\frac{P(Y_i \leq j)}{1 - P(Y_i \leq j)}\right) \\ &= \log\left(\frac{\pi_{i1} + \cdots \pi_{ij}}{\pi_{ij+1} + \cdots + \pi_{iJ}}\right) \end{aligned}$$

On retrouve donc un modèle de régression logistique avec une variable réponse binaire pour laquelle une modalité est formée par les modalités 1 à $j$ de $Y$ et l'autre modalité est formée des modalités $j+1$ à $J$.

# Exemple : modèle dose-réponse

- On considère les données suivantes

Effect of intravenous medication doses on patients with subarachnoid hemorrhage trauma (p. 207, *OrdCDA*)

| | Glasgow Outcome Scale (y) | | | | |
|---|---|---|---|---|---|
| Treatment Group ($x$) | Death | Veget. State | Major Disab. | Minor Disab. | Good Recov. |
| Placebo | 59 | 25 | 46 | 48 | 32 |
| Low dose | 48 | 21 | 44 | 47 | 30 |
| Med dose | 44 | 14 | 54 | 64 | 31 |
| High dose | 43 | 4 | 49 | 58 | 41 |

`http://www.stat.ufl.edu/~aa/ordinal/R_examples.pdf`

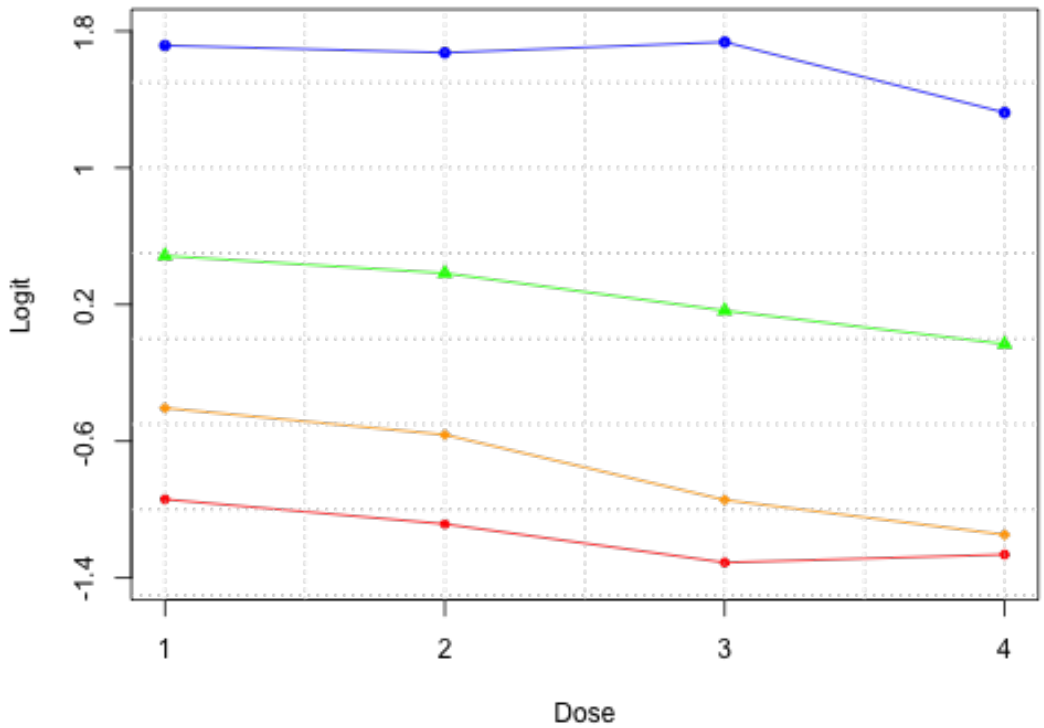
# Exemple : modèle dose-réponse

- Il est utile, quand c'est possible, de tracer les données. Par exemple, ici, on aimerait savoir si les rapports de cotes sont proportionnels ou non.

```
> trauma = matrix(c(1,2,3,4,59,48,44,43,25,21,14,4,46,44,54,
+ 49,48,47,64,58,32,30,31,41),4,6)
> colnames(trauma) <- c("dose","y1","y2","y3","y4","y5")
> trauma = as.data.frame(trauma)
> trauma
  dose y1 y2 y3 y4 y5
1    1 59 25 46 48 32
2    2 48 21 44 47 30
3    3 44 14 54 64 31
4    4 43  4 49 58 41
> plot(1:5,log( trauma[1,2:6]/apply(trauma[2:4,2:6],2,sum)),
+ pch=20,col="blue",ylim=c(-1.3,3),xlab="response",ylab="logit")
> lines(1:5,log( trauma[1,2:6]/apply(trauma[2:4,2:6],2,sum)),
+ col="blue")
> points(1:5, log( apply(trauma[1:2,2:6],2,sum)/apply(trauma[3:4,2:6
+ pch=18,col="green")
> lines(1:5,log(apply(trauma[1:2,2:6],2,sum)/apply(trauma[3:4,2:6],2
> points(1:5,log(apply(trauma[1:3,2:6],2,sum)/trauma[4,2:6]),
+ pch=17,col="red")
> lines(1:5,log(apply(trauma[1:3,2:6],2,sum)/trauma[4,2:6]),
+ col="red")
> grid()
```

# Exemple : modèle dose-réponse

- On observe que les rapports de cote sont assez proches de rapport proportionnels (ils évoluent de la même façon). L'hypothèse de linéarité semble raisonnable.



# Exemple : modèle dose-réponse

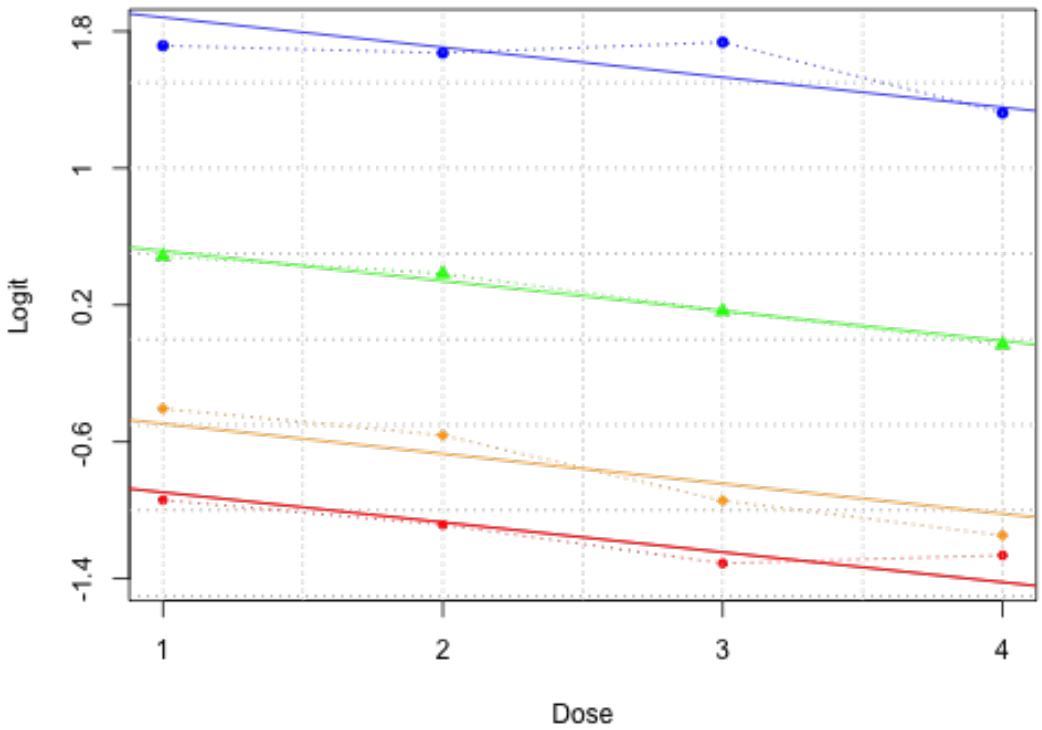
- On commence par ajuster un modèle logistique à rapports de cote proportionnels

$$\log\left(\frac{P(Y_i \leq j)}{1 - P(Y_i \leq j)}\right) = \theta_j + x_{ij}\beta$$

```
>library(VGAM)
> fit = vglm(cbind(y1,y2,y3,y4,y5)~dose,
   family = cumulative(parallel=TRUE),data=trauma)
> summary(fit)
Call:
vglm(formula = cbind(y1, y2, y3, y4, y5) ~ dose,
   family = cumulative(parallel = TRUE),  data = trauma)
Pearson residuals:
  logit(P[Y<=1]) logit(P[Y<=2]) logit(P[Y<=3]) logit(P[Y<=4])
1        -0.8742         1.4325       -0.17949        -0.9222
2        -0.6845         1.2073        0.19701        -0.2754
3        -0.1411        -0.8381       -0.08163         1.2159
4         1.9782        -1.9899        0.04382        -0.1565
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept):1 -0.71917    0.15881  -4.528 5.94e-06 ***
(Intercept):2 -0.31860    0.15642  -2.037  0.04167 *
(Intercept):3  0.69165    0.15793   4.380 1.19e-05 ***
(Intercept):4  2.05700    0.17369  11.843  < 2e-16 ***
dose          -0.17548    0.05632  -3.116  0.00183 **
```

# Exemple : modèle dose-réponse

- On obtient l'ajustement suivant



# Exemple : modèle dose-réponse

- On commence par ajuster un modèle logistique à rapports de cote proportionnels

$$\log\left(\frac{P(Y_i \leq j)}{1 - P(Y_i \leq j)}\right) = \theta_j + x_{ij}\beta$$

- On fait un test de rapport de vraisemblance pour comparer ce modèle au modèle nul.

```
>library(VGAM)
> fit = vglm(cbind(y1,y2,y3,y4,y5)~dose,
   family = cumulative(parallel=TRUE),data=trauma)
> fit = vglm(cbind(y1,y2,y3,y4,y5)~dose,family = cumulative(parallel=TRUE),dat
> fit0 = vglm(cbind(y1,y2,y3,y4,y5)~1,family = cumulative(parallel=TRUE),data=
> statRV = -2*(logLik(fit0)-logLik(fit))
> 1-pchisq(statRV,df=length(coef(fit))-length(coef(fit0)))
[1] 0.001932651
```

- On conclut que ...

# Exemple : modèle dose-réponse, modèle multinomial saturé

- Comparaison avec le modèle multinomial saturé (ie à rapports de cotes non proportionnels)
- Dans ce cas, on choisit une modalité de référence. Les rapports de cotes ne sont donc pas les mêmes et on ne peut pas comparer directement les paramètres estimés des deux modèles.

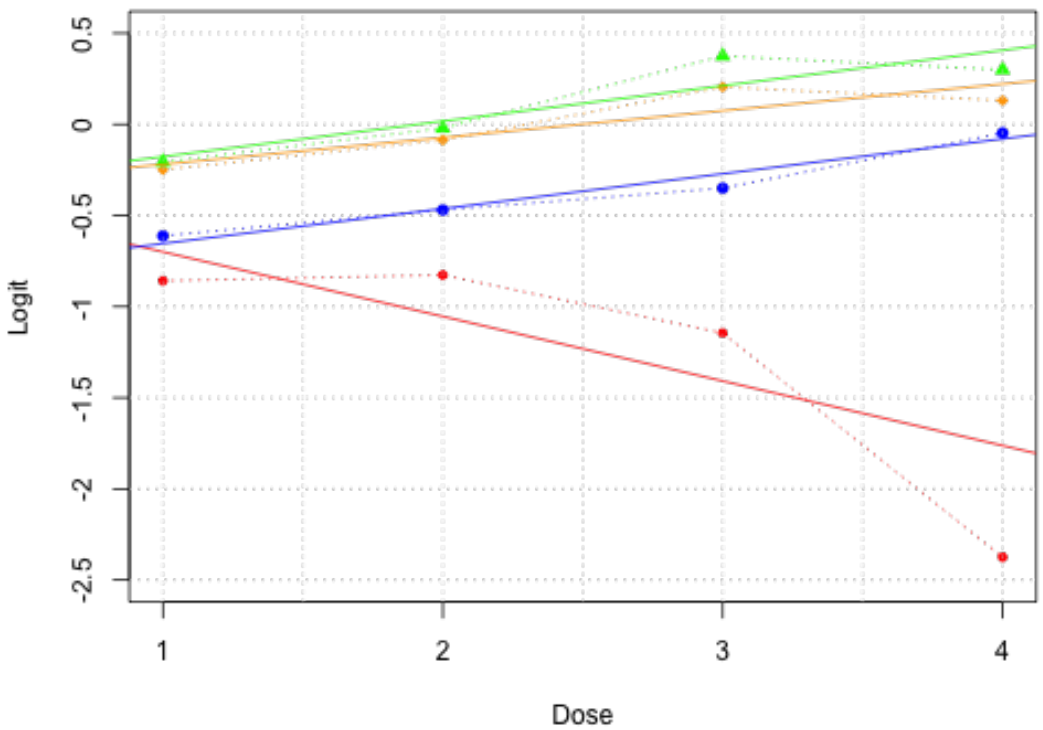
```
> trauma = matrix(c(1,2,3,4,59,48,44,43,25,21,14,4,46,44,54,49,48,
+ 47,64,58,32,30,31,41),4,6)
> count = c(t(trauma[,2:6]))
> dose = c(rep(1,5),rep(2,5),rep(3,5),rep(4,5))
> response = c(matrix(1:5,5,4))
> trauma2 = data.frame(dose=dose,response=response,count=count)
> y =  factor(trauma2$response)
> msat = multinom(y~dose,data=trauma2,weight=count)
> summary(msat)
Coefficients:
  (Intercept)       dose
2  -0.3454744 -0.3544391
3  -0.3664693  0.1470186
4  -0.3718410  0.1945529
5  -0.8458643  0.1914717
> statRV = -2*(logLik(msat)-logLik(fit))
> statRV
'log Lik.' 2351.399 (df=8)
```

# Exemple : modèle dose-réponse, modèle multinomial saturé

- Ajustement du modèle multinomial.
- On ne suppose plus que la réponse est ordinale.
- Le nombre de paramètres est plus important.

```
> -2*logLik(msat)+2*8 # modèle multinomial saturé
'log Lik.' 2465.145 (df=8)
> -2*logLik(fit)+2*5 # modèle à rapports de cotes proportionnels
[1] 107.7456
```

- Avantage : facile à mettre en oeuvre.



# Exemple : modèle dose-réponse, rapport de cotes non proportionnels

- Si on veut donner de la souplesse au modèle, il est plus naturel de comparer le modèle à rapports de cotes proportionnels au modèle à rapports de cotes non proportionnels qu'au modèle mutinomial.

```
> fit2 = vglm(cbind(y1,y2,y3,y4,y5)~dose,family = cumulative,data=trauma)
> summary(fit2)
...
Pearson residuals:
  logit(P[Y<=1]) logit(P[Y<=2]) logit(P[Y<=3]) logit(P[Y<=4])
1         0.5289       -0.48790       -0.13879       -0.35261
2        -0.2835        0.64231        0.20238       -0.06988
3        -0.8415        0.03446       -0.09066        1.07587
4         0.5233       -0.11885        0.03690       -0.67979

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept):1 -0.86459    0.19423  -4.451 8.53e-06 ***
(Intercept):2 -0.09374    0.17849  -0.525   0.5994
(Intercept):3  0.70625    0.17558   4.022 5.76e-05 ***
(Intercept):4  1.90867    0.23838   8.007 1.18e-15 ***
dose:1        -0.11291    0.07288  -1.549   0.1213
dose:2        -0.26890    0.06832  -3.936 8.29e-05 ***
dose:3        -0.18234    0.06385  -2.856   0.0043 **
dose:4        -0.11926    0.08470  -1.408   0.1592
> statRV = -2*(logLik(fit)-logLik(fit2))
> 1-pchisq(statRV,df=length(coef(fit2))-length(coef(fit)))
[1] 0.002487748
```

# Exemple : modèle dose-réponse, rapport de cotes non proportionnels

- On teste alors l'hypothèse des rapports de cotes proportionnels

$$H_0 : \beta_j = \beta, \quad \forall j \in \{1, \cdots, J\}$$

- La différence des déviances suit une loi du chi2 à $(J-1)p$ degrés de liberté.

```
> pchisq(deviance(fit)-deviance(fit2),
    df=df.residual(fit)-df.residual(fit2),lower.tail=FALSE)
[1] 0.002487748
```

- Une alternative consiste à utiliser la statistique de score ie $\nabla_\beta \ell(\tilde{\beta})$ où $\tilde{\beta}$ est la valeur de l'estimateur du maximum de vraisemblance.
  Cette statistique suit une loi du chi2 à $p(J-2)$ degrés de liberté.
  L'avantage de cette seconde approche est qu'on n'a pas besoin d'ajuster le modèle sous l'hypothèse alternative.
- AIC

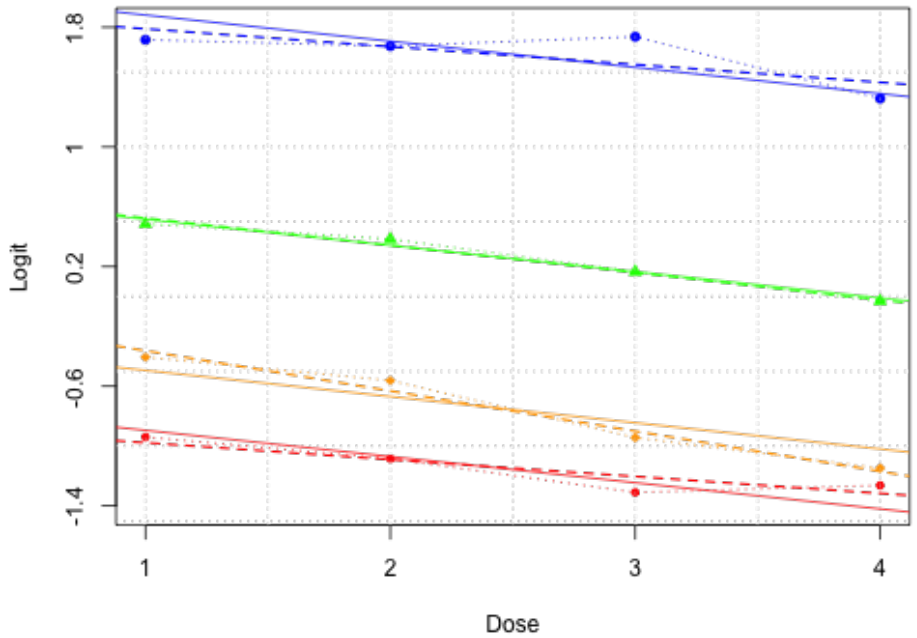
```
> -2*logLik(msat)+2*8
'log Lik.' 2465.145 (df=8)
> -2*logLik(fit)+2*5
[1] 107.7456
> -2*logLik(fit2)+2*8
[1] 99.41481
```

# Exemple : modèle dose-réponse

- L'amélioration est significative d'un point de vue statistique. Mais peut-être n'est-elle pas vraiment utile ? pvalue = 2e-3.

```
> sqrt(mean(sum((fitted(fit)-trauma)^2)))
[1] 190.9919
> sqrt(mean(sum((fitted(fit2)-trauma)^2)))
[1] 190.9899
```

- On le vérifie que la figure (rapports de cotes proportionnels : traits pleins ; rapports de cotes non proportionnels : tirets)



# Modèle cumulatif, autre formulation

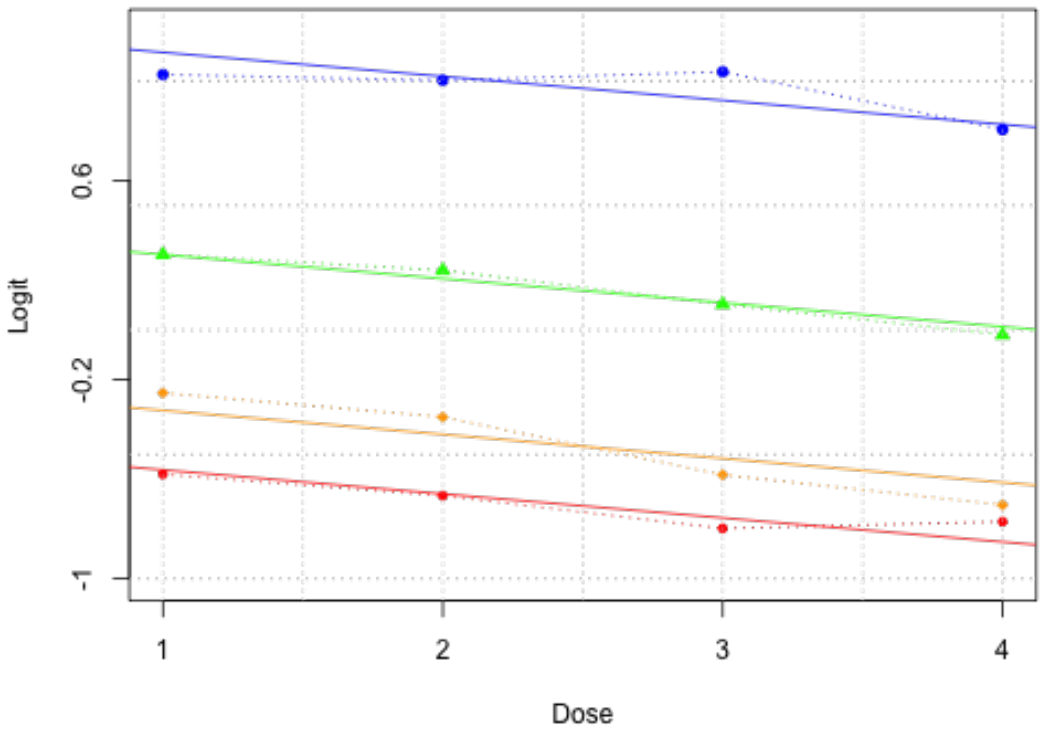
- On peut utiliser une définition basée sur une variable latente comme pour les modèles vus précédemment.
- On peut alors choisir d'autres fonctions de lien. Ici encore les liens usuels sont le lien probit et le lien c-log-log.

```
> fit2 = vglm(cbind(y1,y2,y3,y4,y5)~dose,family = cumulative,data=trauma)
> summary(fit2)
...
Pearson residuals:
  logit(P[Y<=1]) logit(P[Y<=2]) logit(P[Y<=3]) logit(P[Y<=4])
1         0.5289       -0.48790       -0.13879       -0.35261
2        -0.2835        0.64231        0.20238       -0.06988
3        -0.8415        0.03446       -0.09066        1.07587
4         0.5233       -0.11885        0.03690       -0.67979

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept):1 -0.86459    0.19423  -4.451 8.53e-06 ***
(Intercept):2 -0.09374    0.17849  -0.525   0.5994
(Intercept):3  0.70625    0.17558   4.022 5.76e-05 ***
(Intercept):4  1.90867    0.23838   8.007 1.18e-15 ***
dose:1        -0.11291    0.07288  -1.549   0.1213
dose:2        -0.26890    0.06832  -3.936 8.29e-05 ***
dose:3        -0.18234    0.06385  -2.856   0.0043 **
dose:4        -0.11926    0.08470  -1.408   0.1592
...
Residual deviance: 3.8516 on 8 degrees of freedom
> pchisq(deviance(fit)-deviance(fit2), df=df.residual(fit)-df.residual(fit2),
+lower.tail=FALSE)
[1] 0.002487748
```

# Exemple : modèle dose-réponse, lien probit

- Pour le modèle probit à rapports de cote proportionnels



# Réponse ordinale, lien c-log-log

- Si on suppose que $\epsilon$ suit une loi d'extrème, on obtient le modèle

$$\log\left(-\log(1-\gamma_{ij})\right) = \theta_j + \mathbf{x}_i^T\boldsymbol{\beta}$$

- Ce modèle s'appelle aussi le modèle de Cox pour données groupées (discrete proportional hazards model) et il est utilisé pour les données de survie.

# Modèle à risques adjacents

- Dans le cas d'une réponse ordinale, une autre approche usuelle consiste à considérer les rapports de cotes des modalités adjacentes

$$g\left(\frac{\pi_{ij}}{\pi_{ij+1}}\right) = \theta_i + \mathbf{x}_i^T \boldsymbol{\beta}$$

```
> fit.adj = vglm(cbind(y1,y2,y3,y4,y5)~dose,
     family = acat(parallel=TRUE),data=trauma)
> summary(fit.adj)
Pearson residuals:
  logit(P[Y=2]/P[Y=1]) logit(P[Y=3]/P[Y=2]) logit(P[Y=4]/P[Y=3]) logit(P[Y=5]/
1              1.05417              -1.2784               0.11029              0.8
2              0.80369              -1.1253              -0.25909              0.2
3              0.07975               0.8077               0.03749             -1.2
4             -2.19733               1.7747               0.09302              0.2

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept):1 -1.27326    0.15264  -8.342  < 2e-16 ***
(Intercept):2  0.93341    0.15406   6.059 1.37e-09 ***
(Intercept):3 -0.05933    0.11471  -0.517  0.60503
(Intercept):4 -0.66471    0.12614  -5.270 1.37e-07 ***
dose           0.06998    0.02263   3.092  0.00198 **
```

# Modèle à risques séquentiels

- Dans le cas d'une réponse ordinale, une autre approche usuelle consiste à considérer les rapports de cotes des modalités "séquentielles"

$$g\left(\frac{\pi_{i1}+\cdots+\pi_{ij}}{\pi_{ij+1}}\right)=\theta_i+\mathbf{x}_i^T\boldsymbol{\beta}$$

ou

$$g\left(\frac{\pi_{ij}}{\pi_{ij+1}+\cdots+\pi_{iJ}}\right)=\theta_i+\mathbf{x}_i^T\boldsymbol{\beta}$$

```
> fit.seq = vglm(cbind(y1,y2,y3,y4,y5)~dose,
+family = cratio( parallel=TRUE),data=trauma)
> summary(fit.seq)
Pearson residuals:
  logit(P[Y>1|Y>=1]) logit(P[Y>2|Y>=2]) logit(P[Y>3|Y>=3]) logit(P[Y>4|Y>=4])
1          -0.121816           -1.5414            0.8151            1.2524
2          -0.008446           -1.4041            0.1730            0.4764
3           0.564893            0.5629           -0.5312           -1.3353
4          -0.454814            2.6652           -0.4582           -0.1990
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept):1  0.83085    0.13441   6.181 6.35e-10 ***
(Intercept):2  1.82657    0.16939  10.783  < 2e-16 ***
(Intercept):3  0.27102    0.14206   1.908  0.05641 .
(Intercept):4 -0.81712    0.16030  -5.097 3.44e-07 ***
dose           0.12759    0.04418   2.888  0.00388 **
```

# Outline

# Outline

# Données de comptage

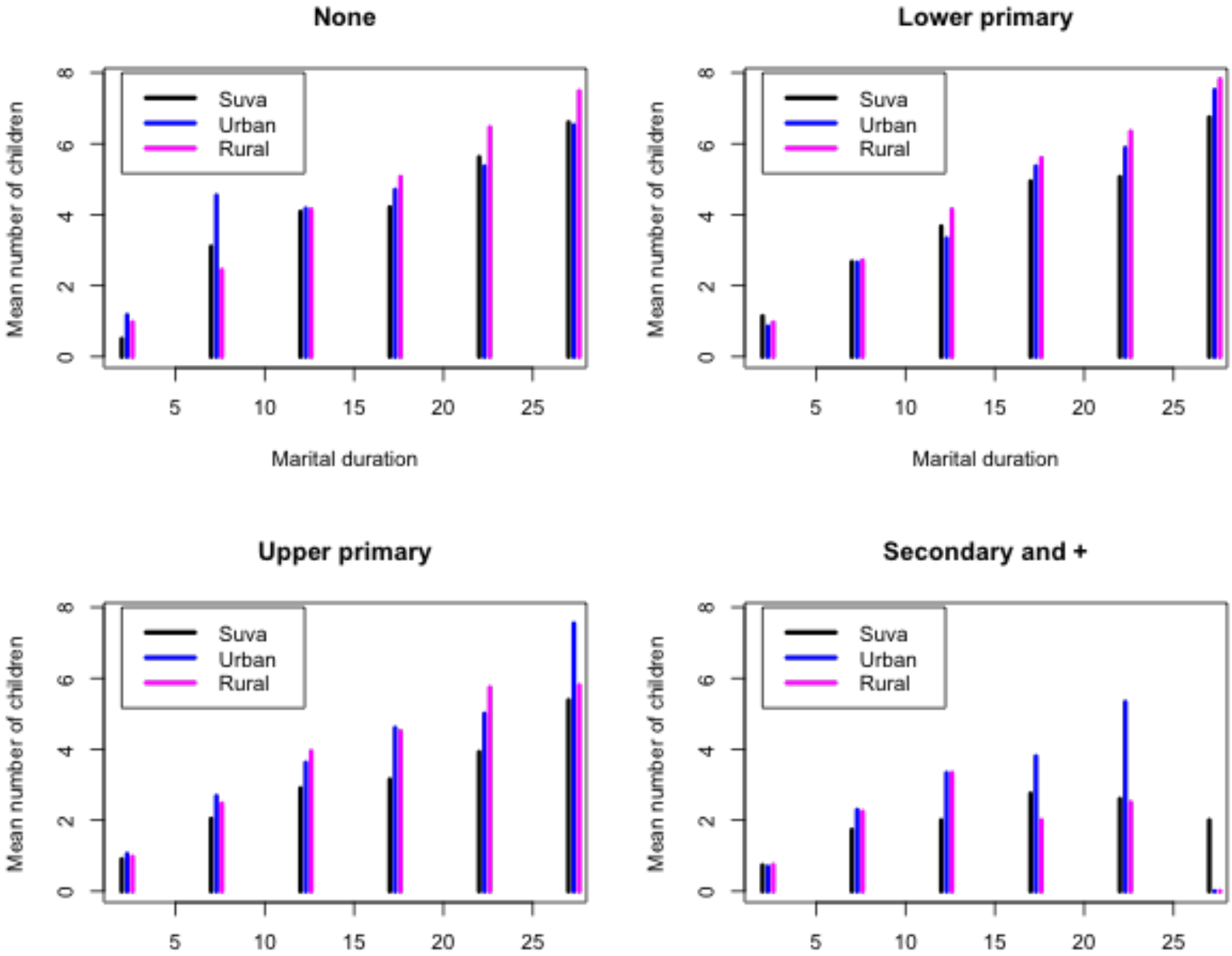
- Dans ce chapitre, nous allons parler des modèles log-linéaires pour les données de comptage sous l'hypothèse d'une erreur qui suit une loi de Poisson.
- Ces modèles ont de nombreuses applications pour analyser des nombres d'évènements (nombre de cas de cancers) mais aussi pour les tables de contingence et les données de survie.
- Comme exemple, nous considérons des données d'un sondage sur la fertilité dans les Fidji. On s'intéresse au nombre d'enfants nés de femmes mariées de race Indienne en fonction du temps écoulé depuis leur 1er mariage (6 catégories), le lieu de résidence (Suva, Urban, Rural) et le niveau d'éducation (none, lower primary, upper mrimary, secondary or higher).
- Dans chaque cellule du tableau ci dessous on donne la moyenne, la variance et l'effectif observés.

# Exemple : nombre de naissances

TABLE 4.1: Number of Children Ever Born to Women of Indian Race By Marital Duration, Type of Place of Residence and Educational Level (Each cell shows the mean, variance and sample size)

| Marr. | Suva | | | | Urban | | | | Rural | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Dur. | N | LP | UP | S+ | N | LP | UP | S+ | N | LP | UP | S+ |
| 0–4 | 0.50 | 1.14 | 0.90 | 0.73 | 1.17 | 0.85 | 1.05 | 0.69 | 0.97 | 0.96 | 0.97 | 0.74 |
| | 1.14 | 0.73 | 0.67 | 0.48 | 1.06 | 1.59 | 0.73 | 0.54 | 0.88 | 0.81 | 0.80 | 0.59 |
| | 8 | 21 | 42 | 51 | 12 | 27 | 39 | 51 | 62 | 102 | 107 | 47 |
| 5–9 | 3.10 | 2.67 | 2.04 | 1.73 | 4.54 | 2.65 | 2.68 | 2.29 | 2.44 | 2.71 | 2.47 | 2.24 |
| | 1.66 | 0.99 | 1.87 | 0.68 | 3.44 | 1.51 | 0.97 | 0.81 | 1.93 | 1.36 | 1.30 | 1.19 |
| | 10 | 30 | 24 | 22 | 13 | 37 | 44 | 21 | 70 | 117 | 81 | 21 |
| 10–14 | 4.08 | 3.67 | 2.90 | 2.00 | 4.17 | 3.33 | 3.62 | 3.33 | 4.14 | 4.14 | 3.94 | 3.33 |
| | 1.72 | 2.31 | 1.57 | 1.82 | 2.97 | 2.99 | 1.96 | 1.52 | 3.52 | 3.31 | 3.28 | 2.50 |
| | 12 | 27 | 20 | 12 | 18 | 43 | 29 | 15 | 88 | 132 | 50 | 9 |
| 15–19 | 4.21 | 4.94 | 3.15 | 2.75 | 4.70 | 5.36 | 4.60 | 3.80 | 5.06 | 5.59 | 4.50 | 2.00 |
| | 2.03 | 1.46 | 0.81 | 0.92 | 7.40 | 2.97 | 3.83 | 0.70 | 4.91 | 3.23 | 3.29 | – |
| | 14 | 31 | 13 | 4 | 23 | 42 | 20 | 5 | 114 | 86 | 30 | 1 |
| 20–24 | 5.62 | 5.06 | 3.92 | 2.60 | 5.36 | 5.88 | 5.00 | 5.33 | 6.46 | 6.34 | 5.74 | 2.50 |
| | 4.15 | 4.64 | 4.08 | 4.30 | 7.19 | 4.44 | 4.33 | 0.33 | 8.20 | 5.72 | 5.20 | 0.50 |
| | 21 | 18 | 12 | 5 | 22 | 25 | 13 | 3 | 117 | 68 | 23 | 2 |
| 25–29 | 6.60 | 6.74 | 5.38 | 2.00 | 6.52 | 7.51 | 7.54 | – | 7.48 | 7.81 | 5.80 | – |
| | 12.40 | 11.66 | 4.27 | – | 11.45 | 10.53 | 12.60 | – | 11.34 | 7.57 | 7.07 | – |
| | 47 | 27 | 8 | 1 | 46 | 45 | 13 | – | 195 | 59 | 10 | – |

# Exemple : nombre moyen de naissances



# Distribution de Poisson

- On dit qu'une variable $Y$ a une distribution de Poisson de paramètre $\mu > 0$ si elle prend des valeurs entières $y = 0, 1, 2, \cdots$ avec la probabilité

$$P(Y = y) = \frac{e^{-\mu}\mu^y}{y!}$$

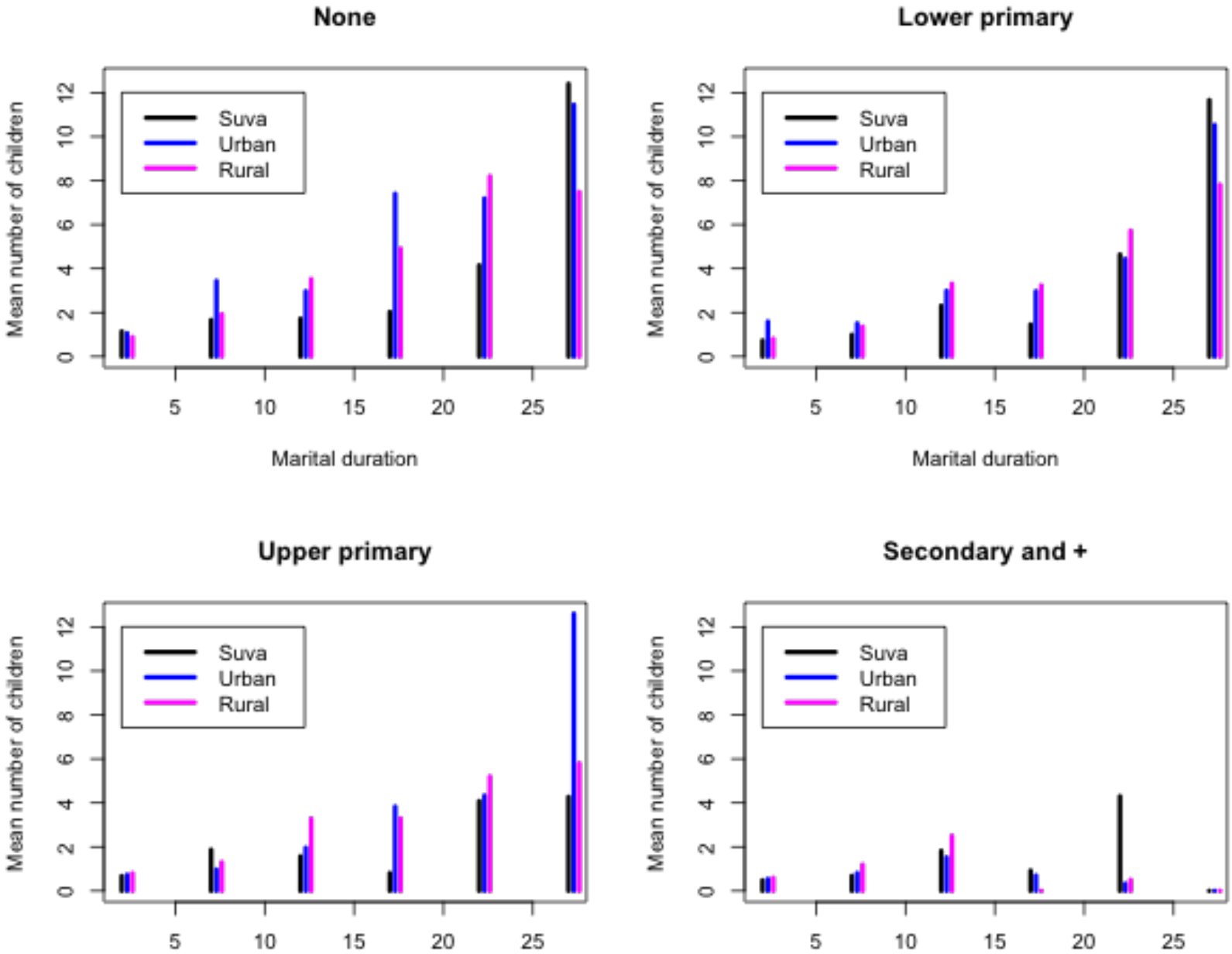
- La moyenne et la variance d'une loi de Poisson sont égales à $\mu$

$$E(Y) = var(Y) = \mu$$

Ainsi, dans la loi de Poisson, la variance augmente avec la moyenne. C'est une des caractéristiques essentielle de la loi de Poisson.

- Exemples connus de données qui suivent une loi de Poisson : nombre de bombes qui ont touché Londres pendant la 2nde guerre mondiale par unité de surface, désintégrations radioactives, échanges de chromosomes dans une cellule, nombre de bactéries dans différentes parties d'une boite de Petri.

# Exemple : variance nombre de naissances



# Propriétés de la distribution de Poisson

- La distribution de Poisson peut être obtenue comme la limite d'une distribution binomiale du nombre de succès parmi un très grand nombre d'essais de Bernoulli avec une faible probabilité de succès.
  Si $Y \sim B(n, \pi)$ alors la distribution de $Y$ quand $n$ tend vers l'infini et $\pi$ avec $\mu = n\pi$ tend vers une loi de Poisson de paramètre $\mu$.
  Autrement dit la loi de Poisson est une loi qui compte les évènements rares.

- La somme de variables aléatoires de loi de Poisson suit aussi une loi de Poisson.
  Si $Y_1$ et $Y_2$ suivent des lois de Poisson de paramètres $\mu_1$ et $\mu_2$ alors $Y_1 + Y_2$ suit une loi de Poisson de paramètre $\mu_1 + \mu_2$.
  Ce résultat se généralise à plus de 2 variables.

- Une conséquence de ce résultat est qu'il est équivalent d'analyser des données individuelles et des données groupées.
  Sous une hypothèse d'indépendance, si les données individuelles $Y_{ij} \sim Pois(\mu_i)$ pour $j = 1, \cdots, n_i$ alors le total du groupe $Y_i \sim Pois(n_i\mu_i)$.

# Outline

# Modèle log-linéaire

- Soient $n$ observations $y_1, \cdots, y_n$ considérées comme des réalisations indépendantes de variables aléatoires de Poisson telles que $Y_i \sim Pois(\mu_i)$.
- On suppose de plus que la moyenne $\mu_i$ dépend de covariables $\mathbf{x}_i$.
- On aurait envie de poser le modèle

$$\mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$$

Mais le terme de droite peut prendre des valeurs négatives... alors que la moyenne d'une loi de Poisson est forcément positive.

- On pose alors

$$\log(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$$

Dans ce modèle $\beta_j$ représente le changement attendu pour le log de la moyenne par unité de changement du prédicteur $x_j$.

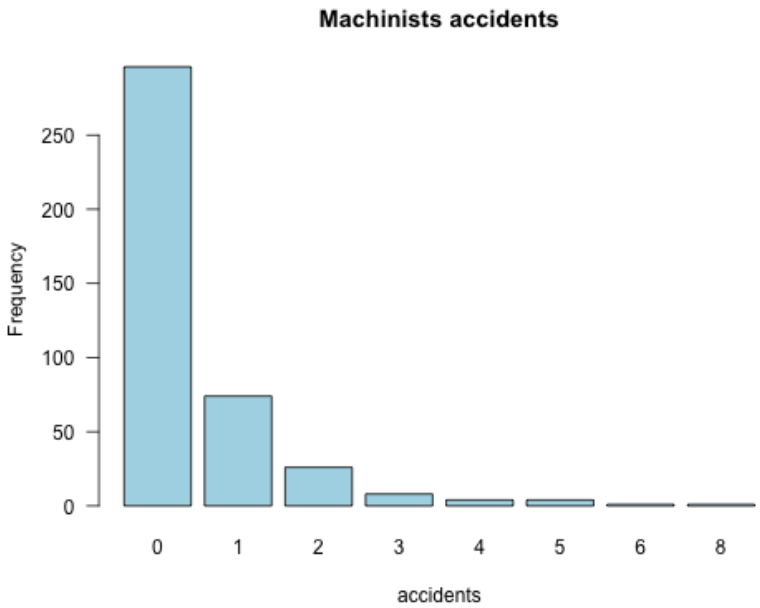
- En prenant l'exponentiel, on obtient un modèle multiplicatif pour la moyenne

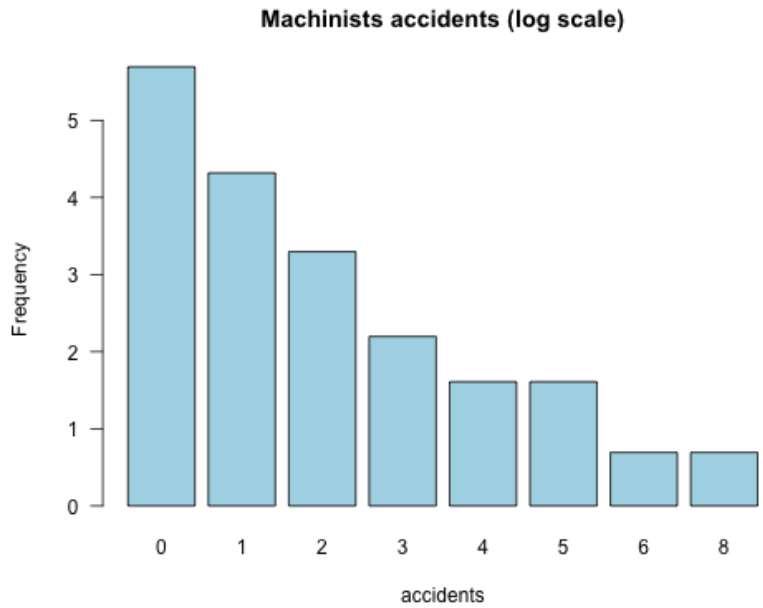
$$\mu_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta})$$

Quand $x_j$ augmente de 1 point, la moyenne est multipliée par $\exp(\beta_j)$.

# Effet du log

- Un des effets avantageux du log est qu'il ramène les données d'évènements rares à une échelle plus linéaire.

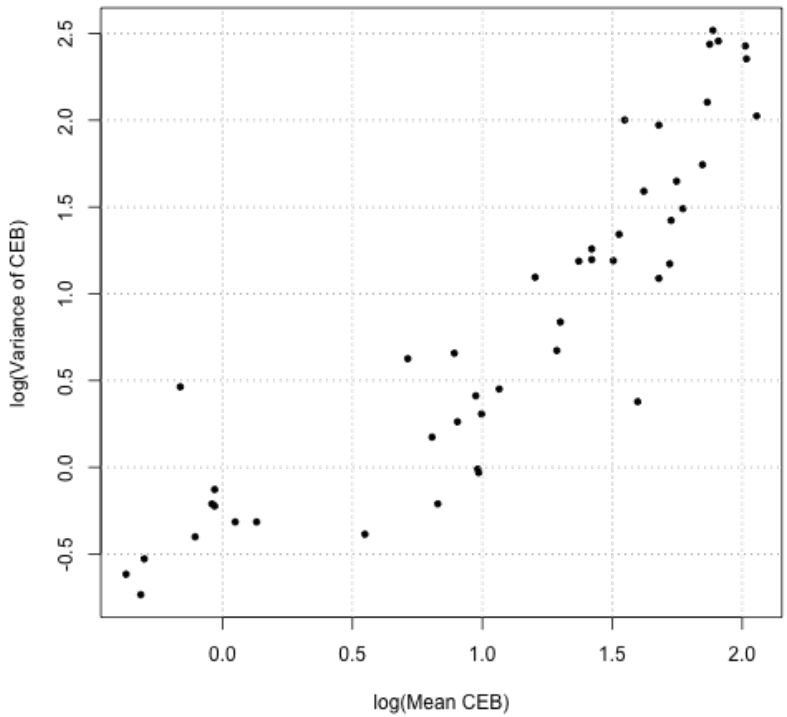


# Outline

# Données surdispersées

- Les données de nombre d'enfants étaient modélisées traditionnellement par un modèle linéaire gaussien.
- Ce choix peut être discuté car le nombre d'enfants varie de 0 à 6... il ne peut pas suivre une loi de Gauss.
- Cependant le point le plus critique, n'est pas la loi des erreurs mais le fait que la variance n'est pas constante.
- On trace la variance de chaque cellule du tableau en fonction de la moyenne en échelle log-log.




On voit que l'hypothèse de variance constante n'est pas du tout vérifiée.
La variance est proche de la moyenne
$\rightarrow$ plus cohérent avec le modèle de Poisson que le modèle Gaussien.

# Données groupées et "offset"

- A t-on besoin des données individuelles pour la régression de Poisson ?
- Notons $Y_{ijk\ell}$ le nombre d'enfants nés de la femme $\ell$ dans le groupe $ijk$ et $Y_{ijk} = \sum_\ell Y_{ijk\ell}$ le total du groupe.
- Si chaque observation de ce groupe est issue d'une variable aléatoire de Poisson de moyenne $\mu_{ijk}$ alors le total suit une loi de Poisson de moyenne $n_{ijk}\mu_{ijk}$
- Et dans ce cas

$$\begin{aligned} \log E(Y_{ijk}) &= \log(n_{ijk}\mu_{ijk}) + \mathbf{x}_{ijk}^T\boldsymbol{\beta} \\ &= \log(n_{ijk}) + \log(\mu_{ijk}) + \mathbf{x}_{ijk}^T\boldsymbol{\beta} \end{aligned}$$

  Ainsi, le total du groupe suit un modèle log-linéaire avec exactement le même coefficient $\boldsymbol{\beta}$ que le modèle pour les données individuelles, à l'exception de l'offset $\log(n_{ijk})$

- L'offset est fréquent dans les modèles log-linéaires. Il représente souvent le log d'une mesure d'exposition (ici le nombre de femmes).

# Exemple

- Les données du nombre d'enfants par femme ne donnent pas d'information individuelle. Mais on peut travailler avec les données groupées si on introduire un offset, qui est ici le log nombre de femme dans chaque groupe.

```
> require(foreign)
> ceb <- read.dta("http://data.princeton.edu/wws509/datasets/ceb.dta")
> head(ceb)
  i dur   res          educ mean  var  n
1 1 0-4  Suva          None 0.50 1.14  8
2 2 0-4  Suva Lower primary 1.14 0.73 21
3 3 0-4  Suva Upper primary 0.90 0.67 42
>ceb$y <- round(ceb$mean*ceb$n, 0)
>ceb$os = log(ceb$n)
>m0 <- glm( y ~ offset(os), data=ceb, family=poisson)
>summary(m0)
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 1.376346   0.009712   141.7   <2e-16 ***

Null deviance: 3731.9  on 69  degrees of freedom
Residual deviance: 3731.9  on 69  degrees of freedom
AIC: 4163.3
>exp(coef(m0))
(Intercept)
   3.960403
   > sum(ceb$y)/sum(ceb$n))
[1]  3.960403
```

# Outline

# Modèle log-linéaire

- Le modèle décrit ci-dessous est un modèle linéaire généralisé avec une erreur Poisson et un lien log.
- La log vraisemblance de $n$ observations de Poisson indépendantes s'écrit

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i \log \mu_i - \mu_i) + cte$$

où $\mu_i$ dépend des covariables $\mathbf{x}_i$ et du vecteur de paramètres $\boldsymbol{\beta}$ et $cte$ est une constante indépendante des paramètres $\boldsymbol{\beta}$

- On montre facilement qu'en dérivant la log-vraisemblance en fonction de $\boldsymbol{\beta}$ et en annulant ces dérivées, on obtient les équations d'estimation

$$\mathbf{X}^T \mathbf{y} = \mathbf{X}^T \widehat{\boldsymbol{\mu}}$$

où $\mathbf{X}$ est la matrice de design (incluant une colonne de 1), $\mathbf{y}$ est la réponse observée et $\widehat{\boldsymbol{\mu}}$ est la réponse prédite avec l'estimateur du maximum de vraisemblance.

- Les équations d'estimation ont cette forme dans tous les modèles linéaires généralisés avec la fonction de lien canonique.

# Équations d'estimation (1/2)

- Les équations d'estimation du modèle de Poisson s'écrivent

$$\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\widehat{\boldsymbol{\mu}}$$

- Considérons un exemple pour mieux comprendre : dans les données des naissances, on considère le modèle $Y_i \sim Pois(\mu_i)$ avec $\mu_i = \exp(\beta_0 + \beta_1 X_i^{(U)})$ où $X^{(U)}$ est une variable qui vaut 1 si la mère vit en ville et 0 sinon.
- Le modèle inclut un intercept, $\mathbf{X}$ contient une colonne de 1 : $\mathbf{1}$. Or

$$\mathbf{1}\mathbf{y} = \sum_{i=1}^{n} y_i$$

et la première équation s'écrit

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \exp(\beta_0 + \beta_1 x_i^{(U)}) = e^{\beta_0} \sum_{i=1}^{n} e^{\beta_1 x_i^{(U)}} = e^{\beta_0}(n_0 + n_1 e^{\beta_1})$$

avec $n_1$ le nombre de mères qui vivent en ville et $n_0$ le nombre des autres mères.

- La seconde colonne de $\mathbf{X}$ correspond à la variable dichotomique et l'équation associée s'écrit donc

$$\sum_{\{i|x_i^{(U)}=1\}} y_i = \sum_{\{i|x_i^{(U)}=1\}} \exp(\beta_0 + \beta_1 x_i^{(U)}) = e^{\beta_0} n_1 e^{\beta_1 x_U(i)}$$

# Équations d'estimation (2/2)

- On obtient ainsi un système d'équations à 2 équations et deux inconnues $e^{\beta_0}$ et $e^{\beta_1}$

$$\begin{cases} e^{\beta_0}(n_0 + n_1 e^{\beta_1}) = \sum_{i=1}^{n} y_i = N_{\text{total}} \\ e^{\beta_0} n_1 e^{\beta_1} = \sum_{\{i|x_U(i)=1\}} y_i = N_{\text{né en ville}} \end{cases}$$

- En faisant le rapport des deux équations, on simplifie le terme $e^{\beta_0}$ et on obtient

$$\frac{n_0 + n_1 e^{\beta_1}}{n_1 e^{\beta_1}} = \frac{\sum_{i=1}^{n} y_i}{\sum_{\{i|x_U(i)=1\}} y_i} = \frac{N_{\text{total}}}{N_{\text{né en ville}}}$$

d'où

$$e^{\beta_1} = \frac{n_0}{n_1} \frac{N_{\text{né en ville}}}{N_{\text{total}} - N_{\text{né en ville}}} = \frac{n_0}{n_1} \frac{N_1}{N_0}$$

- On obtient $e^{\beta_0}$ en utilisant la seconde équation

$$e^{\beta_0} = \frac{N_{\text{né en ville}}}{n_1 e^{\beta_1}} = \frac{N_{\text{total}} - N_{\text{né en ville}}}{n_0} = \frac{N_0}{n_0}$$

- Ce résultat se généralise à plus de variables et des variables à plus de modalités.

# Estimation en pratique

- Le résultat précédent se généralise à plus de variables et des variables à plus de modalités.
- Cependant, en général, l'estimation se fait en utilisant l'algorithme numérique Iterated Re-weighted Least Square (IRLS) comme pour la régression logistique avec ici

$$z_i = \eta_i + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}$$

et les poids

$$w_{ii} = \hat{\mu}_i$$

- Cet algorithme est initialisé en prenant le log des réponses $y_i$ et en le régressant sur les variables observées par moindres carrés. Si certaines réponses sont nulles, leur log n'est pas défini alors on leur ajoute une petite constante.

```
> m1r <- glm( y ~ res + offset(os), data=ceb, family=poisson)
> summary(m1r)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.20460    0.02499  48.199  < 2e-16 ***
resUrban     0.14429    0.03245   4.447 8.72e-06 ***
resRural     0.22806    0.02783   8.194 2.52e-16 ***

Null deviance: 3731.9  on 69  degrees of freedom
Residual deviance: 3659.3  on 67  degrees of freedom
```

# Goodness-of-fit

- La déviance permet de mesurer l'écart entre le modèle et l'observation.

$$D = 2\sum_{i=1}^{n}\left(y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right)$$

- Le premier terme est identique à la déviance binomiale.
- Le second terme est égal à zero puisque qu'une des propriétés de la régression de Poisson est de reproduire les totaux marginaux (voir exemple ci-dessus).
- Si $n$ est grand, la déviance suit approximativement une loi du chi2 à $n - p$ degrés de liberté avec $n$ le nombre d'observations et $p$ le nombre de degrés de liberté.

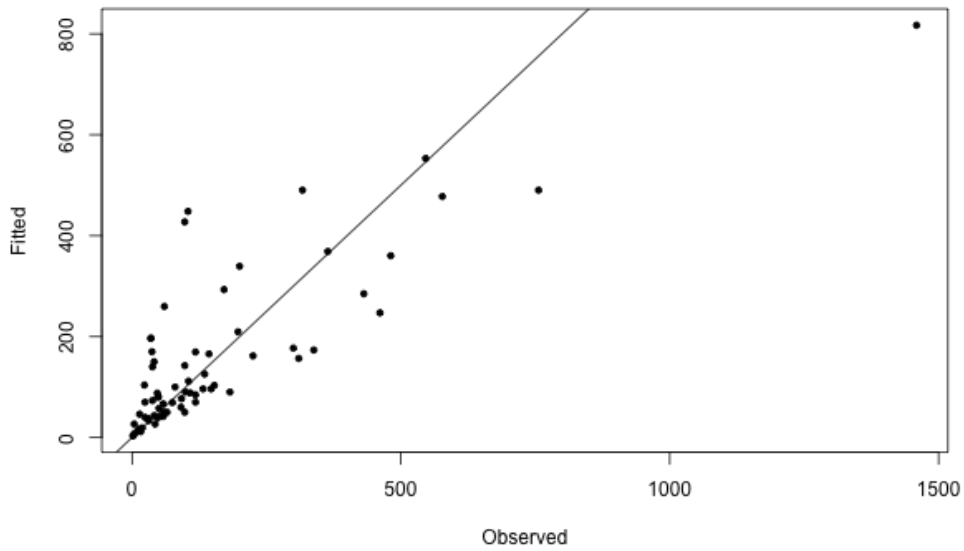
```
> m1r <- glm( y ~ res + offset(os), data=ceb, family=poisson)
> summary(m1r)
 Null deviance: 3731.9  on 69  degrees of freedom
Residual deviance: 3659.3  on 67  degrees of freedom
AIC: 4094.8
> qchisq(.95,df=67)
[1] 88.25016
> anova(m0,m1r)
Model 1: y ~ offset(os)
Model 2: y ~ res + offset(os)
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        69     3731.9
2        67     3659.3  2   72.572 < 2.2e-16 ***
```

# Goodness-of-fit, exemple

- Modèle

$$\text{Nb naissance} \sim \textit{Pois}(\textit{os} \exp(\beta_0 + \beta_1 \text{Résidence}))$$

- Le test d'ajustement montre que la variable "Résidence" ne suffit pas à bien prédire le nombre de naissances.

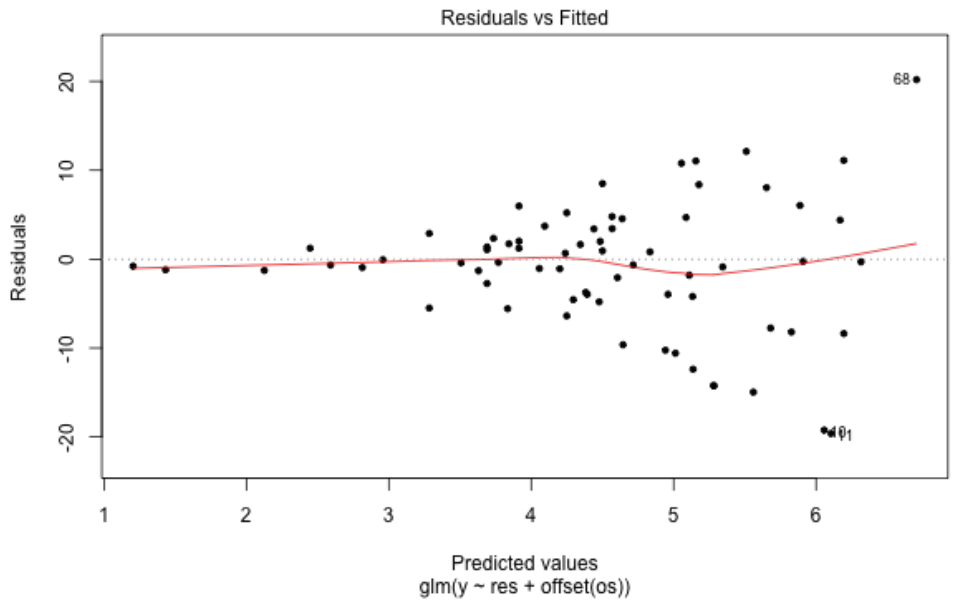


# Goodness-of-fit, exemple

- Modèle

$$\text{Nb naissance} \sim \textit{Pois}(\exp(\beta_0 + \beta_1 \text{Résidence}))$$

- Le graphe des résidus montre qu'on explique bien la moyenne mais pas la variance.



# Autres tests d'ajustement

- Pour tester la qualité d'ajustement,

  $H_0$ Le modèle permet de reproduire les observations | $H_1$ Le modèle ne permet pas de reproduire les observations

  on peut aussi utiliser la statistique de Pearson

$$\chi^2_p = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$

  qui suit approximativement une loi du chi2 à $n - p$ degrés de libertés si $n$ est grand.

```
> res.pearson = sum((ceb$y-m1r$fitted.values)^2/m1r$fitted.values)
[1] 3304.612
>
```

- On peut aussi construire un test de rapport de vraisemblance ou un test de Wald.

# Déviance

- On compare les différents modèles à l'aide des déviances.

| Modèle | Déviance | d.d.l. |
|---|---|---|
| Null | 3731.52 | 69 |
| *Modèles à 1 facteur* | | |
| Durée | 165.84 | 64 |
| Résidence | 3659.23 | 67 |
| Education | 2661.00 | 66 |
| *Modèles à 2 facteurs* | | |
| D+R | 120.68 | 62 |
| D+E | 100.01 | 61 |
| DR | 108.84 | 52 |
| DE | 84.46 | 46 |
| *Modèles à 3 facteurs* | | |
| D+R+E | 70.65 | 59 |
| D + RE | 59.89 | 53 |
| E + DR | 57.06 | 49 |
| R+DE | 54.91 | 44 |
| DR+RE | 44.27 | 43 |
| DE+RE | 44.60 | 38 |
| DR+DE | 42.72 | 34 |
| DR + DE + RE | 30.95 | 28 |

- On prédit le nombre d'enfants en fonction de la durée du mariage, la résidence et le niveau d'éducation.
- Le modèle nul a une déviance de 3731.52 pour 69 ddl ce qui revient à rejeter l'hypothèse selon laquelle le nombre moyen d'enfants est le même pour tous les groupes.
- L'introduction de la durée du mariage dans le modèle permet de faire décroitre significativement la déviance pour une faible décroissance du nombre de ddl. Ceci reflète le fait que le nombre moyen d'enfant est fortement dépendant du temps pendant lequel la femme est "exposée" à une maternité possible.
- Il est clair qu'il n'est pas raisonnable de considérer un modèle ne tenant pas compte de cette variable.

# Interprétation : effet de l'éducation (1/2)

| Modèle | Déviance | d.d.l. |
|---|---|---|
| Null | 3731.52 | 69 |
| *Modèles à 1 facteur* | | |
| Durée | 165.84 | 64 |
| Résidence | 3659.23 | 67 |
| Education | 2661.00 | 66 |
| *Modèles à 2 facteurs* | | |
| D+R | 120.68 | 62 |
| D+E | 100.01 | 61 |
| DR | 108.84 | 52 |
| DE | 84.46 | 46 |
| *Modèles à 3 facteurs* | | |
| D+R+E | 70.65 | 59 |
| D + RE | 59.89 | 53 |
| E + DR | 57.06 | 49 |
| R+DE | 54.91 | 44 |
| DR+RE | 44.27 | 43 |
| DE+RE | 44.60 | 38 |
| DR+DE | 42.72 | 34 |
| DR + DE + RE | 30.95 | 28 |

- Pour tester l'effet brut de l'éducation, on peut comparer la déviance du modèles incluent uniquement l'éducation avec celle du modèle nul

$$3731 - 2661 = 1071 / \text{pour } 3 \; ddl$$

C'est très significatif !

- Mais pour ce modèle ça n'a pas de sens d'exclure la durée du mariage, on préfère donc tester l'effet de l'éducation celui de la durée étant pris en compte ie comparer les modèles D+E et D.

$$165 - 100 = 65 \text{ pour } 3 \; ddl$$

soit

$$p_{value} = 2.5e - 14$$

# Interprétation : effet de l'éducation (2/2)

| Modèle | Déviance | d.d.l. |
|---|---|---|
| Null | 3731.52 | 69 |
| *Modèles à 1 facteur* | | |
| Durée | 165.84 | 64 |
| Résidence | 3659.23 | 67 |
| Education | 2661.00 | 66 |
| *Modèles à 2 facteurs* | | |
| D+R | 120.68 | 62 |
| D+E | 100.01 | 61 |
| DR | 108.84 | 52 |
| DE | 84.46 | 46 |
| *Modèles à 3 facteurs* | | |
| D+R+E | 70.65 | 59 |
| D + RE | 59.89 | 53 |
| E + DR | 57.06 | 49 |
| R+DE | 54.91 | 44 |
| DR+RE | 44.27 | 43 |
| DE+RE | 44.60 | 38 |
| DR+DE | 42.72 | 34 |
| DR + DE + RE | 30.95 | 28 |

- Les femmes ayant un haut niveau d'éducation ont tendance à être plus jeunes, l'effet de l'éducation est donc amplifié.
- En comparant, D+R et D+E+R, la différence de déviance est un peu plus faible : 50. Elle reste significative.
- Les femmes éduquées ont tendance à vivre en ville. E et R sont donc en partie redondantes.

# Interprétation : interactions

| Modèle | Déviance | d.d.l. |
|---|---|---|
| Null | 3731.52 | 69 |
| *Modèles à 1 facteur* | | |
| Durée | 165.84 | 64 |
| Résidence | 3659.23 | 67 |
| Education | 2661.00 | 66 |
| *Modèles à 2 facteurs* | | |
| D+R | 120.68 | 62 |
| D+E | 100.01 | 61 |
| DR | 108.84 | 52 |
| DE | 84.46 | 46 |
| *Modèles à 3 facteurs* | | |
| D+R+E | 70.65 | 59 |
| D + RE | 59.89 | 53 |
| E + DR | 57.06 | 49 |
| R+DE | 54.91 | 44 |
| DR+RE | 44.27 | 43 |
| DE+RE | 44.60 | 38 |
| DR+DE | 42.72 | 34 |
| DR + DE + RE | 30.95 | 28 |

- Le niveau d'éducation fait-il plus de différence à la campagne qu'en ville ?
  La décroissance de déviance entre D+R+E et D+RE est seulement de 10 environ pour 6 ddl ; ce n'est pas significatif (pvalue = 0.09).
- L'effet de du niveau d'éducation croit-il avec la durée du mariage ?
  La décroissance de déviance entre D+R+E et D+RE est seulement de 15 environ pour 15 ddl ; (pvalue = 0.54).
- Mêmes remarques pour les autres interactions $\rightarrow$ on garde le modèle D+R+E.

# Modèle additif D+R+E

| Param. | | Estim. | Std Err | z-ratio |
|---|---|---|---|---|
| Constant | | -0.1173 | 0.0549 | -2.14 |
| Durée | 0-4 | - | | |
| | 5-9 | 0.9977 | 0.0528 | 18.91 |
| | 10-14 | 1.3705 | 0.0511 | 26.83 |
| | 15-19 | 1.6142 | 0.0512 | 31.52 |
| | 20-24 | 1.7855 | 0.0512 | 34.86 |
| | 25-29 | 1.9768 | 0.0500 | 39.50 |
| Résidence | Suva | - | | |
| | Urbaine | 0.1123 | 0.0325 | 3.46 |
| | Rurale | 0.1512 | 0.0283 | 5.34 |
| Education | Aucune | - | | |
| | Faible | 0.0231 | 0.0227 | 1.02 |
| | Élevée | -0.1017 | 0.0310 | -3.28 |
| | Sec+ | -0.3096 | 0.0552 | -5.61 |

- La constante représente le log du nombre d'enfant pour les cellules de références. $e^{-0.1173} = 0.89$, les femmes mariées depuis moins de 5 ans vivant à Suva et n'ayant pas d'éducation ont donc en moyenne 0.89 enfants.
- On observe que la durée du mariage accroît le nombre moyen d'enfant par femme. En passant de "moins de 5 ans" de mariage à "enter 5 et 9 ans", le log de la moyenne croit de 1. $e^{0.9977} = 2.71$. Tout lieux de résidence et niveaux d'éducation confondus, les femmes mariées depuis plus de 10 ans ont en moyenne 7.22 plus d'enfants que les femmes mariées depuis moins de 5 ans.

# Modèle additif D+R+E

| Param. | | Estim. | Std Err | z-ratio |
|---|---|---|---|---|
| Constant | | -0.1173 | 0.0549 | -2.14 |
| Durée | 0-4 | - | | |
| | 5-9 | 0.9977 | 0.0528 | 18.91 |
| | 10-14 | 1.3705 | 0.0511 | 26.83 |
| | 15-19 | 1.6142 | 0.0512 | 31.52 |
| | 20-24 | 1.7855 | 0.0512 | 34.86 |
| | 25-29 | 1.9768 | 0.0500 | 39.50 |
| Résidence | Suva | - | | |
| | Urbaine | 0.1123 | 0.0325 | 3.46 |
| | Rurale | 0.1512 | 0.0283 | 5.34 |
| Education | Aucune | - | | |
| | Faible | 0.0231 | 0.0227 | 1.02 |
| | Élevée | -0.1017 | 0.0310 | -3.28 |
| | Sec+ | -0.3096 | 0.0552 | -5.61 |

- L'effet de la résidence montre que les femmes de Suva ont la fertilité la plus faible. Quelque soit la durée du mariage et le niveau d'éducation, les femmes vivant dans un autre zone urbaine ont 12% fois plus d'enfants (exp0.1123 = 1.12), et dans les zones rurales 16%.
- Plus le niveau d'éducation est élevé, plus le nombre moyen d'enfants par femme est faible, quelque soit la durée du mariage et le lieu résidence.
- On a vu plus haut que les interactions n'apportent pas beaucoup au modèle. Cependant, le modèle de Poisson est un modèle additif en échelle log. Dans l'échelle d'origine le modèle est multiplicatif [a].

a. $\beta_1 \log(x_1) + \beta_2 \log(x_2) = log(x_1^{\beta_1} x_2^{\beta_2})$

# Modèle de Poisson et effet multiplicatif

- Un modèle additif en échelle log traduit des interactions dans l'échelle d'origine.
- Prédiction du nombre moyen d'enfants

| Durée du mariage | 0-4 | 5-9 | 10-14 | 15-19 | 20-24 | 25-29 |
|---|---|---|---|---|---|---|
| Sans éducation | 0.89 | 2.41 | 3.50 | 4.47 | 5.30 | 6.42 |
| Sec + | 0.65 | 1.77 | 2 .57 | 3.28 | 3.89 | 4.71 |
| Différence | 0.24 | 0.64 | 0.93 | 1.19 | 1.41 | 1.71 |

- Les femmes qui ont un niveau d'éducation secondaire ou plus ont en moyenne 27% $(1 - e^{-0.3096})$ plus d'enfants que les femmes "sans éducation".
- Le tableau montre que la différence est croissante avec la durée du mariage. Ainsi l'effet de l'éducation mesuré dans l'échelle d'origine dépend de la durée du mariage.
- On rappelle par ailleurs que le modèle de Poisson permet à la variance d'évoluer avec la moyenne : on a plus de variabilité dans les cellules de moyenne élevée.

# Outline

# Sur-dispersion

- Une des caractéristiques clé de la distribution de Poisson est l'égalité de la moyenne et de la variance.

$$E(Y) = Var(Y) = \mu$$

- Cependant, les données réelles présentent parfois de la sur-dispersion ie une variance plus importante que la moyenne.
- Il existe plusieurs modèles/solutions permettant de prendre ne compte la sur-dispersion
  - modèle quasi-Poisson,
  - modèle de Poisson avec estimation robuste de la variance,
  - modèle à réponse binomiale négative.
- Supposons que la variance est proportionnelle à la moyenne

$$Var(Y) = \phi E(Y) = \phi\mu$$

Si $\phi > 1$ on a une sur-dispersion et si $\phi < 1$ un sous-dispersion (mais ce second cas est rare en pratique).

# Inférence (1/2)

- Dans le cas où la variance est proportionnelle à la moyenne $Var(Y) = \phi E(Y) = \phi\mu$, il est facile d'adapter l'algorithme IRLS.
- Il suffit en effet de remplacer les poids $w_{ii}$ par

$$w_{ii}^* = \frac{\hat{\mu}}{\phi}$$

- Ce sont les poids du modèle de Poisson divisés par $\phi$. La constante $\phi$ disparait quand on calcule

$$\left(\mathbf{X}^T W \mathbf{X}\right)^{-1} \mathbf{X}^T W \mathbf{z}$$

de telle sorte que l'estimateur du maximum de vraisemblance du modèle de Poisson est un estimateur du maximum de la quasi-vraisemblance pour le modèle avec variance proportionnelle à la moyenne.

- La variance de $\widehat{\beta}$ est alors

$$Var(\widehat{\beta}) = \phi \left(\mathbf{X}^T W \mathbf{X}\right)^{-1}$$

avec $W = diag(\mu_1, \cdots, \mu_n)$

- Ainsi les écart-types d'estimation du modèle de Poisson sont conservatifs en cas de sur-dispersion.

# Inférence (2/2)

- On a cependant besoin d'estimer la constante $\phi$.
- L'approche classique s'appuie sur la statistique du chi2 de de Pearson

$$\chi^2_p = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{var(y_i)} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{\phi \mu_i}$$

  Si le modèle est correct l'espérance de cette statistique est n-p.
- Par la méthode des moments, on a donc

$$\hat{\phi} = \frac{\chi^2_p}{n - p}$$

- Remarque - D'habitude une valeur importante de la statistique de Pearson traduit un mauvais ajustement. Donc, quand on utilise la statistique il faut être assez sûr que l'erreur d'ajustement vient bien de la sur-dispersion.

# Régression binomiale négative (1/3)

- Quand la sur-dispersion est due à une hétérogénéité de la moyenne, une alternative consiste à ajouter un effet multiplicatif aléatoire pour décrire l'hétérogénéité dans le modèle de Poisson. Ceci conduit au modèle de régression binomiale négative.
- On suppose que la loi conditionnelle de $Y$ sachant une variable non observée $\theta$ suit une loi de Poisson de moyenne et de variance $\mu\theta$ :

$$Y \sim \mathcal{P}(\mu\theta)$$

$\theta$ représente un facteur latent qui a tendance à faire croître $Y$ par rapport à ce qu'on attend sachant les covariables observées si $\theta > 1$ et décroitre sinon.

- En pratique on fixe $E(\theta) = 1$. Si on suppose de plus que $\theta$ suit une loi Gamma de paramètres $\alpha$ et $\gamma$, alors $Y$ suit une loi binomiale négative

$$P(Y = y) = \frac{\Gamma(\alpha + y)}{y!\Gamma(\alpha)} \frac{\gamma^\alpha \mu^y}{(\mu + \beta)^{\alpha+y}}$$

C'est la loi qui modélise habituellement le nombre d'échec avant le $k$ième succès dans une série de tirage de Bernoulli indépendants de paramètre $\pi$.

$$\alpha = k, \quad \pi = \gamma/(\mu + \gamma)$$

# Régression binomiale négative (2/3)

- La distribution binomial négative avec $\alpha = \gamma = 1/\sigma^2$ est telle que

$$E(Y) \text{ et } Var(Y) = \mu(1 + \sigma^2\mu)$$

SI $\sigma = 0$ on retrouve la distribution de Poisson. Si $\sigma > 0$ la variance est plus grande que la variance dans une distribution de Poisson. La loi binomiale négative est donc sur-dispersée par rapport à celle de Poisson.

- Avec $\theta$ telle que $E(Y|\theta) = var(Y|\theta) = \theta\mu$ , $E(\theta) = 1$ et $Var(\theta) = \sigma^2$, on a

$$E(Y) = E_\theta\left(E_{Y|\theta}(Y|\theta)\right) = E_\theta\left(\theta\mu\right) = \mu$$

et

$$\begin{aligned} Var(Y) &= E_\theta\left(Var_{Y|\theta}(Y|\theta)\right) + Var_\theta\left(E_\theta\left(\theta\mu\right)\right) \\ &= E_\theta\left(\theta\mu\right) + Var_\theta\left(\theta\mu\right) \\ &= \mu + \mu^2\sigma^2 = \mu(1 + \sigma^2\mu) \end{aligned}$$

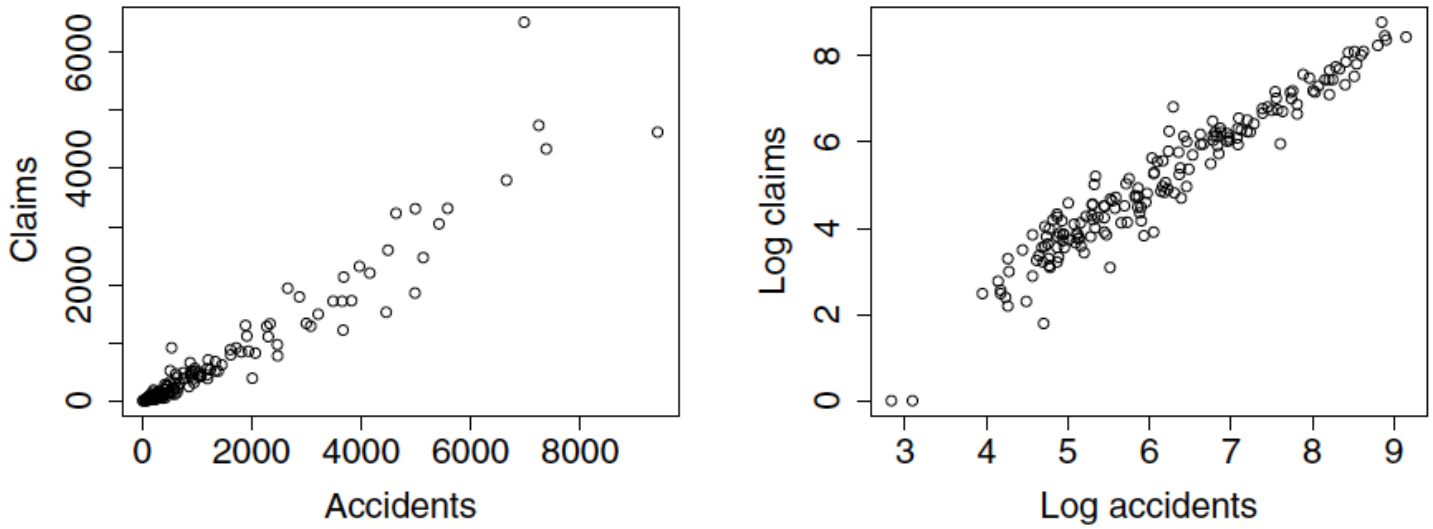
- On retrouve les relations précédentes sans faire d'hypothèse sur la loi de $\theta$ et on peu en déduire des estimateurs des paramètres.

# Régression binomiale négative (3/3)

- La distribution de Poisson est un cas particulier de la distribution binomiale négative avec $\sigma^2 = 0$. On peut donc construire des tests de rapport de vraisemblance pour comparer les deux modèles.
- Mais comme le modèle de Poisson (ie l'hypothèse nulle) est sur la frontière de l'espace des paramètres, la statistique de test ne converge pas vers une loi du chi2 à 1 ddl comme pourrait s'y attendre.
- On peut approcher la loi de la statistique de test par simulation ou utiliser le fait que si on considère la statistique de test comme un chi2 à 1 ddl le test obtenu est conservatif.

# Régression binomiale négative - Exemple (1/3)

- *Assurance au tiers, Australie* - On répertories le nombre de plaintes sur une période de 12 mois entre 1984 et 1986 dans 176 aires géographiques en New South Wales. Les autre variables mesurées sont le nombre d'accidents, le nombre de personnes tuées ou accidentées et la population.




- Le log du nombre de plaintes varie linéairement avec le log du nombre d'accidents.

# Régression binomiale négative - Exemple (2/3)

- Pour se faire une idée de la sur-dispersion, on peut estimer la moyenne et la variance du nombre plaintes en fonction de cinq groupes du nombre d'accidents correspond à cinq quantiles.

| | Nombre d'accidents | | | | | |
|---|---|---|---|---|---|---|
| | 0-138 | 139-267 | 268-596 | 597-1810 | $\geq$1811 | Overall |
| Mean | 30 | 68 | 178 | 497 | 2176 | 587 |
| Variance | 397 | 1206 | 25 622 | 31 751 | $1.8\times10^6$ | $1.0\times10^6$ |
| Variance/Mean | 13 | 18 | 144 | 64 | 847 | 1 751 |
| Coefft of var | 0.66 | 0.51 | 0.90 | 0.36 | 0.62 | 1.73 |
| n | 36 | 35 | 35 | 35 | 35 | 176 |

- La variance est bien plus importante que la moyenne, un modèle de régression de Poisson n'est donc pas approprié.
- Le modèle de régression binomiale négative avec lien log s'écrit

$$Y \sim BN(k, \mu), \quad \log \mu = \log(n) + \mathbf{x}^T \boldsymbol{\beta}$$

- Dans le cas des plaintes de tiers, on propose

$$Y \sim BN(k, \mu), \quad \log \mu = \log(n) + \beta_1 + \beta_2 \log(z)$$

avec $z$ le nombre d'accidents et $n$ la population d'une aire géographique.

# Régression binomiale négative - Exemple (3/3)

- Pour le modèle

$$Y \sim BN(\kappa, \mu), \quad \log \mu = \log(n) + \beta_1 + \beta_2 \log(z),$$

on obtient les résultats suivants

| Paramètre | ddl | $\beta$ | *se* | chi2 | pvalue |
|---|---|---|---|---|---|
| Intercept | 1 | -6.954 | 0.162 | 1836.69 | < 0.0001 |
| log accidents | 1 | 0.254 | 0.025 | 100.04 | < 0.0001 |
| $\kappa$ | | 0.172 | 0.020 | | |

- On a donc que le nombre moyen de plaintes varie avec le nombre d'accidents selon l'équation

$$\hat{\mu} = ne^{-6.954+0.254 \log(z)}$$

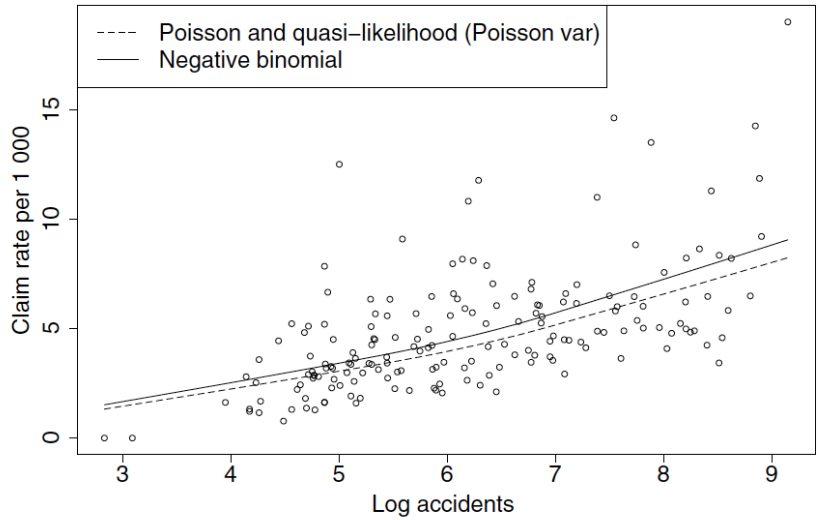
- Ainsi, si le nombre d'accident augmente d'un facteur $a$ alors le taux $\mu/n$ augmente d'un facteur $e^{0.254 \log(a)} = a^{0.254}$.
  Par exemple, si le nombre d'accidents augmente de 10%, $a = 1.1$, on estime un effet sur le nombre moyen de plaintes de $1.1^{0.254} = 1.02$ ie 2%.
- Si on fixe $\kappa = 0$, on retrouve le modèle de Poisson. Ce modèle a une déviance de 15836.7 pour 174 ddl, ce qui traduit un très mauvais ajustement.
  En comparaison, le modèle de binomial négatif a une déviance de 192.3 pour 174 ddl.
- La statistique de test du rapport de vraisemblances des deux modèles est égale à 15027. On rejette donc très largement l'hypothèse $\kappa = 0$.

# Régression binomiale négative ou quasi-vraisemblance ?

- Dans le cas des plaintes de tiers, on ajuste le modèle basé sur la quasi-vraisemblance.
- On peut alors comparer les différents modèles

| Modèle | Variance | $\phi$ | $\hat{\beta}_1$ (se) | $\hat{\beta}_2$ (se) |
|---|---|---|---|---|
| Poisson | $\mu$ | $\phi = 1$ | -7.094 (0.027) | 0.259 (0.003) |
| Quasi-vrais. | $\phi\mu$ | $\hat{\phi}$= 91.02 | -7.094 (0.258) | 0.259 (0.032) |
| Bin. nég. | $\mu(1+\kappa\mu)$ | - | -6.954 (0.162) | 0.254 (0.025) |

- Le modèle basé sur la quasi-vraisemblance conduit aux mêmes estimateurs que le modèle de Poisson mais avec un écart-type plus grand qui traduit l'inflation de la variance.
  Le modèle binomial négatif conduit aussi à des estimations très proches.
- Le modèle binomial négatif est intuitivement plus satisfaisant car il modélise vraiment le mécanisme de sur-dispersion. Cependant le modèle basé sur la quasi-vraisemblance conduit à des résultats très similaires.



# Outline

# Validation, sélection de modèles

- Validation : Est-ce que le modèle sélectionné est "bon" ? En statistique cette question peut être abordée de différentes façons :
  - Est-ce que la qualité d'ajustement globale est satisfaisante : le modèle décrit-il bien les valeurs observées ? Ce type de question fait l'objet des tests d'ajustement ou d'adéquation (goodness of fit) $\rightarrow$ tests basés sur la déviance).
  - Est-ce que le modèle est un bon prédicteur ? $\rightarrow$ performances en prédiction.

- Sélection : étant donnés $M$ modèles $\mathcal{M}_1, \cdots, \mathcal{M}_M$, comment choisir le "meilleur" à partir de l'échantillon dont on dispose.

# Performance en prédiction

- Il convient de définir au préalable une règle de prévision (on se restreint dans un premier temps au modèle logistique).
- Un modèle $M_\beta$ fournit une estimation $\hat{p}_\beta(\mathbf{x}) = p_{\hat{\beta}}(\mathbf{x})$, il est naturel de définir une règle de prévision $\hat{\delta}_\beta$ à partir de cette estimation :

$$\hat{\delta}_\beta(x) = \begin{cases} 1 & \text{si} \quad \hat{p}_\beta(\mathbf{x}) \geq s \\ 0 & \text{sinon} \end{cases}$$

où $s \in [0, 1]$ est un seuil fixé par l'utilisateur.
Il existe plusieurs façons de choisir ce seuil, les logiciels statistiques prennent généralement par défaut la valeur 0.5.

- Etant donné $\hat{\delta} : \mathbb{R}^{p+1} \to \{0, 1\}$ une règle de prévision construite à partir d'un échantillon $\mathcal{D}_n = \{(X_1, Y_1), \cdots, (X_n, Y_n)\}$, on définit la probabilité d'erreur de $\hat{\delta}$ par

$$L(\hat{\delta}) = P\left(\hat{\delta}(X) \neq Y | \mathcal{D}_n\right)$$

- Pour comparer $K$ règles (ou modèles), l'approche consiste à estimer les probabilités d'erreur de toutes les règles candidates à l'aide de l'échantillon puis à choisir la règle qui possède la plus petite estimation.

# Outline

# Estimation de l'erreur de prédiction (ou de classement)

- La difficulté est de trouver un "bon" estimateur de l'erreur de prédiction $L(\hat{\delta})$.
- Une première idée est d'utiliser

$$\hat{L}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\hat{\delta}(X_i)\neq Y_i}$$

  Comme le modèle saturé ajuste de manière parfait les données, cette procédure revient à choisir de manière quasi systématique le modèle saturé.
- Exemple caricatural : prédiction par le plus proche voisin.
  Considérons le modèle suivant : pour prédire la classe d'un individu $x$, on cherche le plus proche voisin de $x$ parmi $X_1, \cdots, X_n$ (notons le $X_{i^*}$) et on pose $\hat{\delta}(x) = Y_{i^*}$.
  Dans l'approche précédente pour estimer l'erreur de prédiction, on cherche pour chaque $X_i$ son plus proche voisin parmi $X_1, \cdots, X_n$, on trouve bien sûr $X_i$ et l'erreur de prédiction est donc nulle.
- La faiblesse de cette approche vient du fait qu'on utilise le même échantillon pour construire le modèle et pour estimer la probabilité d'erreur. Ceci introduit un biais dans l'estimation de la probabilité d'erreur.

# Estimation de l'erreur de prédiction par apprentissage-validation

- La procédure apprentissage-validation s'affranchit de ce problème en séparant de manière aléatoire les données $\mathcal{D}_n = \{(X_1, Y_1), \cdots, (X_n, Y_n)\}$ en deux parties distinctes :
  - $\mathcal{D}_\ell = \{(X_i, Y_i), i \in \mathcal{I}_\ell\}$, un échantillon d'apprentissage de taille $\ell$ qui est utilisé pour ajuster les modèles ;
  - $\mathcal{D}_m = \{(X_i, Y_i), i \in \mathcal{I}_m\}$, un échantillon de validation de taille $m = n - \ell$ qui est utilisé pour estimer les erreurs de prédiction

$$L_n(\hat{\delta}) = \frac{1}{m} \sum_{i \in \mathcal{I}_m} \mathbb{I}_{\hat{\delta}(X_i) \neq Y_i} \; ;$$

  avec $\mathcal{I}_\ell \cup \mathcal{I}_m = \{1, \cdots, n\}$ et $\mathcal{I}_\ell \cap \mathcal{I}_m = \emptyset$.
- $L_n(\hat{\delta})$ est un estimateur sans biais de $L(\hat{\delta})$.
- Exemple : on montre que l'erreur de prédiction est sous estimée par l'estimation en resubstitution.

# Estimation de l'erreur de prédiction par apprentissage-validation

- Exemple

```
> file = 'http://www.indiana.edu/~statmath/stat/all/cdvm/gss_cdvm.csv'
> df=read.table(file, sep=',', header=T)
> df$www = as.factor(df$www) ; df$male = as.factor(df$male) ; df$trust = as.fa
> head(df)
  trust belief educate income age male www
1     0      2      15    9.0  48    0   0
2     1      3      14   27.5  39    0   1
3     0      0      14   27.5  25    0   1
> appri = sample(nrow(df),nrow(df)*2/3) ; testi = setdiff(1:nrow(df),appri)
> blm<-glm(trust~educate+income+age+male+www,data=df, family=binomial)
> predr = predict(blm,df[testi,],type="response")
> (c.r=table(predr>.5,df$trust[testi]))
         0   1
  FALSE 168  94
  TRUE   50  80
> (err.r = 1-sum(diag(c.r))/length(testi))
[1] 0.3673469
> blm<-glm(trust~educate+income+age+male+www,data=df,
      family=binomial,subset=appri)
> predcv = predict(blm,df[testi,],type="response")
> (c.cv = table(predcv>.5,df$trust[testi]))
         0   1
  FALSE 174 102
  TRUE   44  72
> (err.cv = 1-sum(diag(c.cv))/length(testi))
[1] 0.372449
```

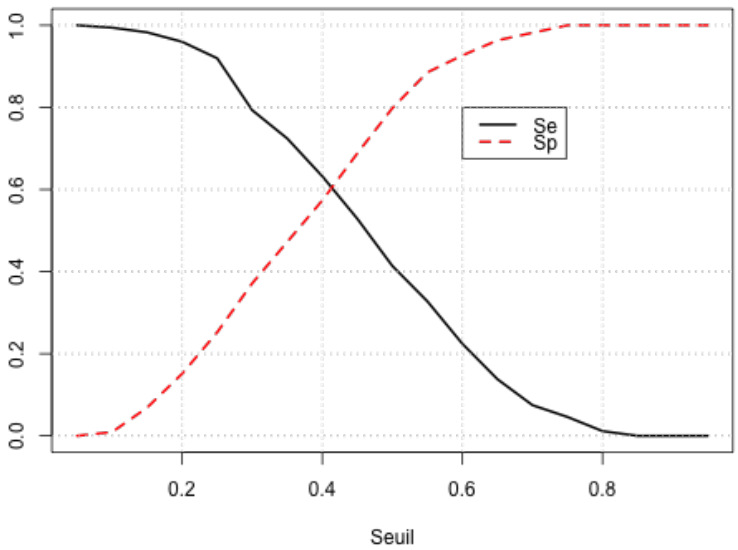
# Matrice de confusion, sensibilité, spécificité

- Pour certaines applications, on évalue d'autres indices que l'erreur de classement.
- Pour un seuil $s$ donné, on construit par exemple une matrice de confusion

| Prévision | Observation | | Total |
|---|---|---|---|
| | Y=1 | Y=0 | |
| $\hat{y}_i = 1$ | $n_{11}(s)$ | $n_{01}(s)$ | $n_{1+}(s)$ |
| $\hat{y}_i = 0$ | $n_{01}(s)$ | $n_{00}(s)$ | $n_{0+}(s)$ |
| Total | $n_{+1}$ | $n_{+0}$ | $n$ |

- sensibilité (ou taux de vrais positifs) : $n_{11}(s)/n_{+1}(s)$
- spécificité (ou taux de vrais négatifs) : $n_{00}(s)/n_{+0}(s)$

- Il est possible de choisir un seuil $s$ qui permette d'atteindre une sensibilité fixé ou une spécificité fixée.



# Liste d'indices

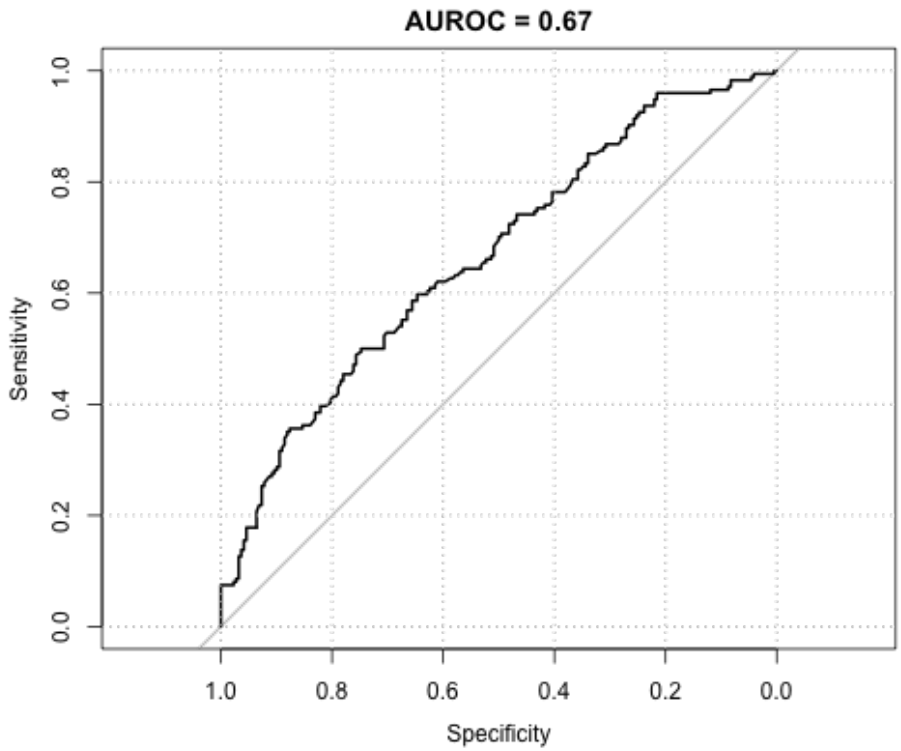
| | Observation | |
|---|---|---|
| Prévision | event | no event |
| event | A | B |
| no event | C | D |

- Sensitivity = A/(A+C)
- Specificity = D/(B+D)
- Prevalence = (A+C)/(A+B+C+D)
- PPV = (sensitivity * Prevalence)/((sensitivity*Prevalence) + ((1-specificity)*(1-Prevalence)))
- NPV = (specificity * (1-Prevalence))/(((1-sensitivity)*Prevalence) + ((specificity)*(1-Prevalence)))
- Detection Rate = A/(A+B+C+D)
- Detection Prevalence = (A+B)/(A+B+C+D)
- Balanced Accuracy = (Sensitivity+Specificity)/2

# Courbe ROC

- Tous les indices listés requièrent le choix d'un seuil.
- Il existe des indicateurs plus flexibles (toujours basés sur la prévision) qui n'imposent pas de fixer le seuil.
- La courbe ROC fait partie de ces critères.
- C'est une courbe paramétrée par le seuil

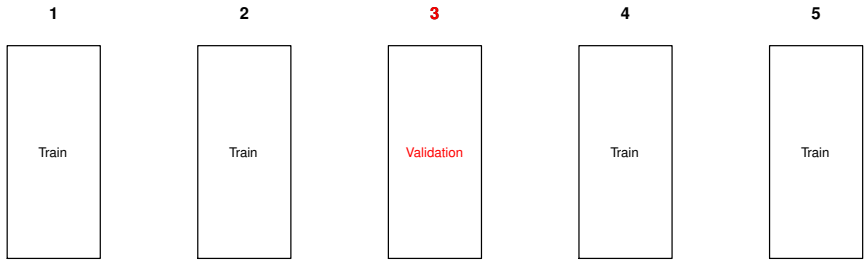
$$\begin{cases} x(s) = 1 - sp(s) \\ y(s) = se(s) \end{cases}$$



# Outline

# K fold cross validation

- On a vu que a validation croisée est une bonne méthode pour estimer une erreur de prédiction.
- Le plus souvent, on combine plusieurs estimations en apprentissage-validation.

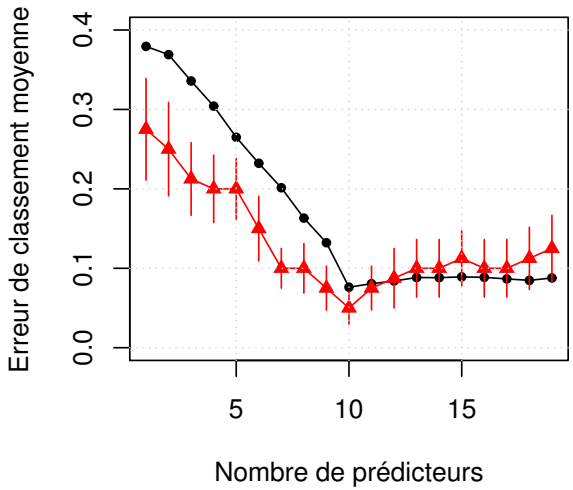



$$CV(\hat{f}) = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{[n/K]}\sum_{i\in\mathcal{I}_k}\mathbb{I}_{\hat{\delta}(X_i)\neq Y_i}$$

- Quelle valeur de $K$ choisir ?
  - Leave-one-out : $K = N \rightarrow$ estimation de $Err_{\mathcal{A}}$ ; dans ce cas l'estimateur est approximativement sans biais, mais peut avoir une grande variance.
  - $K = 5$ ou $10 \rightarrow$ estimation de $Err$ ; les ensembles d'apprentissages sont assez différents de l'ensemble d'origine. La variance est faible, mais le biais peut-être un problème si chaque ensemble est "petit".

# Validation croisée

- Exemple : $X$ suit une loi uniforme sur $[0,1]^{\times 20}$ et $Y = 1$ si $\sum_{p=1}^{10} X > 5$ et $0$ sinon.
- L'ensemble d'apprentissage est composé de 80 individus.
- On estime l'erreur de prédiction (en noir) sur un ensemble de test très grand et on la compare à l'erreur estimée par validation croisée 10 folds (en rouge). Les bâtons représentent un intervalle de plus ou moins un écart-type de l'erreur moyenne individuelle de chaque sous ensemble.




- L'estimation de l'erreur par validation croisée 10-folds sous-estime l'erreur de prédiction pour les petites valeurs de $d$ mais elle identifie bien le nombre de composantes.

# Estimation de l'erreur de prédiction, K-fold

- Exemple

```
> library(boot)
> blm<-glm(trust~educate+income+age+male+www, data=df, family=binomial(link="
> predr = predict(blm,df[testi,],type="response")
> c.r=table(predr>.5,df$trust[testi])
> (err.r = 1-sum(diag(c.r))/length(testi))
[1] 0.3673469
> mean(abs((as.numeric(trust)-1)-prev_prob)>.5)
[1] 0.3673469
> cout = function(trust,prev_prob){
c = mean(abs(trust-prev_prob)>.5)
return(c)
}
> cv.glm(df,blm,cout)$delta # leave-one-out
[1] 0.3492334 0.3498624
> cv.glm(df,blm,cout,K=5)$delta
[1] 0.3509370 0.3483838
```

# Courbe ROC, K-fold

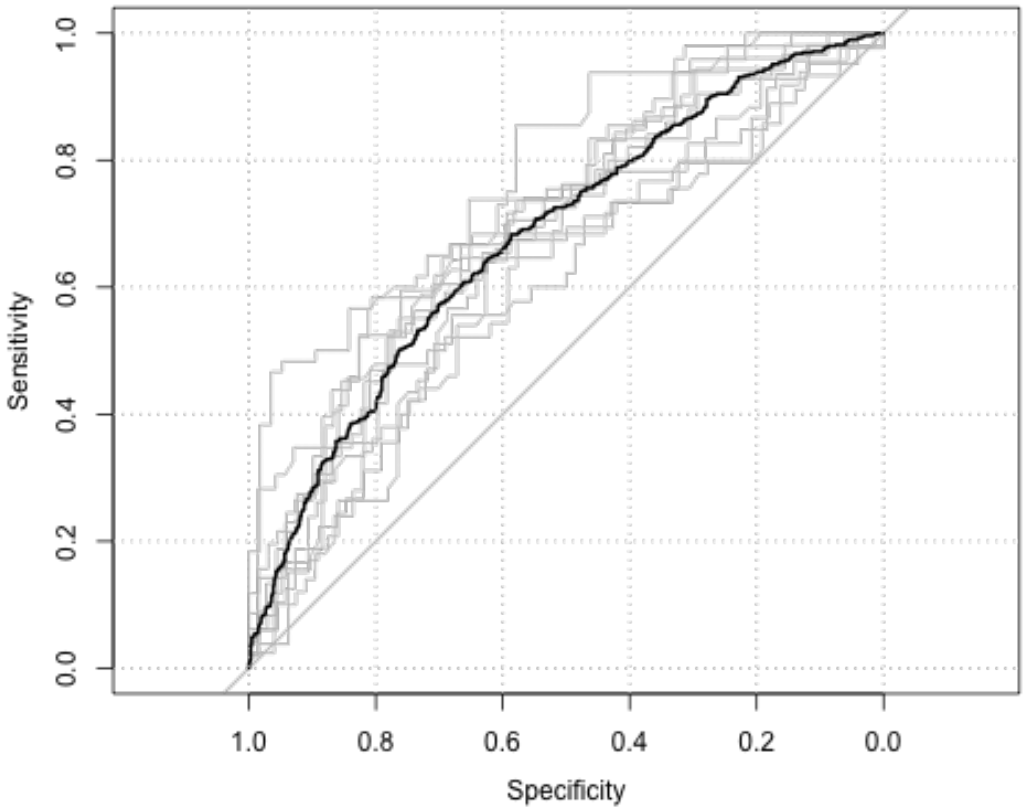
- Exemple

```
K = 10
nk = floor(nrow(df)/K)
pred = NULL
Yobs = NULL
rk=list()
ii = sample(nrow(df),nrow(df)) # permutation aléatoire des indices
for (k in 1:K){
testi = ii[((k-1)*nk+1):(k*nk)]
appri = setdiff(1:nrow(df),testi)
blm<-glm(trust~educate+income+age+male+www, data=df, family=binomial(link="lo
pr = predict(blm,df[testi,],type="response")

pred = c(pred,pr)
Yobs = c(Yobs,df$trust[testi])
    rk[[k]]=roc(df$trust[testi],pr)
}
add = FALSE
for (k in 1:K){
if (k>1) {add=TRUE}
plot.roc(rk[[k]],col="gray",add=add,lwd=1)
}
r=roc(Yobs,pred)
plot(r,main=paste("AUROC =",round(r$auc*100)/100),add=TRUE)
grid()
```

# Courbe ROC, K-fold

- On peut, par exemple, tracer les $K$ courbe ROC et leur moyenne.
- On a ainsi une idée de la variabilité des résultats.



# Bien utiliser la validation croisée

- Un processus de sélection/validation de modèle comporte généralement plusieurs étapes :
  1. Sélection d'un sous ensemble de "bons" prédicteurs ;
  2. Construction d'un modèle basé sur ce sous ensemble et estimation des paramètres ;
  3. Estimation de l'erreur de prédiction.
- La validation croisée doit être appliquée à l'ensemble du processus ie qu'on tire aléatoirement les sous échantillons avant de réaliser les étapes 1 à 3.

# Outline

# Méthodes bootstrap

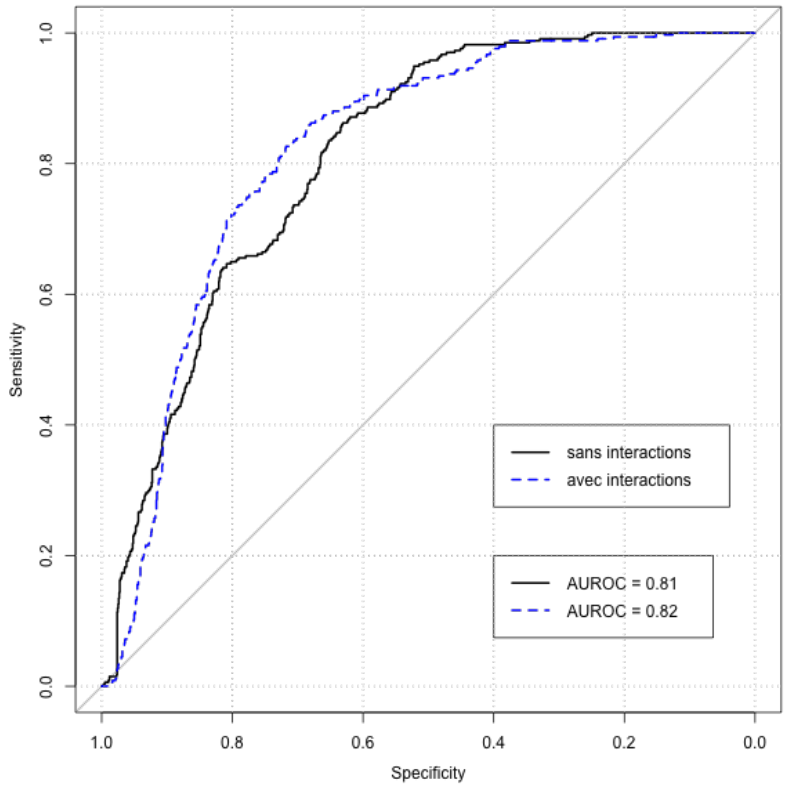
- L'objectif du bootstrap est d'estimer $L_n(\hat{\delta})$ mais il estime bien l'erreur de prédiction attendue $L(\hat{\delta})$.
- Algorithme

  Pour b = 1,$\cdots$, B
  Tirer aléatoirement avec remise un ensemble $\mathcal{I}_b$ de taille $n_{app}$
  dans l'ensemble des données
  Sélectionner les prédicteurs
  Ajuster le modèle
  Estimer l'erreur de prédiction avec les individus qui ne sont pas dans $\mathcal{I}_b$.
  fin
  Calculer la moyenne des erreurs de prédiction
- En pratique on choisi souvent B = 100.
- On peut aussi choisir de faire un tirage sans remise d'un ensemble de taille prédéfinie.
- Si la taille $n$ de l'échantillon est assez grande, on peut le décomposer aléatoirement en 3 sous ensembles
  - Ensemble d'apprentissage [50%] $\rightarrow$ estimation des paramètres
  - Ensemble de validation [25%] $\rightarrow$ sélection de variables
  - Ensemble de test [25%] $\rightarrow$ estimation de $L_n(\hat{\delta})$

# Comparaison de modèles

```
library(rpart)
data(kyphosis)
summary(kyphosis)
n = nrow(kyphosis)
ntest = round(n/5)
p.add <- p.mul <- Yobs <- NULL
B = 100
for (b in 1:B){
testi = sample(1:n,ntest)
appri = setdiff(1:n,testi)
mod = glm(Kyphosis~.,data=kyphosis,family=binomial,subset=appri)
pmod = predict(mod,kyphosis[testi,])
mod2 = glm(Kyphosis~.^2,data=kyphosis,family=binomial,subset=appri)
pmod2 = predict(mod2,kyphosis[testi,])
Yobs = c(Yobs,kyphosis$Kyphosis[testi])
p.add = c(p.add,pmod)
p.mul = c(p.mul,pmod2)
}
r.add=roc(Yobs,p.add)
plot(r.add,main=paste("AUROC =",round(r.add$auc*100)/100))
r.mul=roc(Yobs,p.mul)
plot(r.mul,add=TRUE,col="blue",lty=2)
legend(.4,.4,legend=c("sans interactions","avec interactions"),lty=1:2,
grid()
```

# Comparaison de modèles




L'ajout des interactions permet d'améliorer légèrement le modèle selon la courbe ROC.

# Remarques

- Les estimations du risque empirique considérées (validation croisée, bootstrap) sont asymptotiquement équivalentes et il n'est pas possible de savoir laquelle sera la meilleure à $n$ fini.
- La validation par bootstrap est de plus en plus utilisée et remplace petit à petit la validation croisée car elle permet de mieux tenir compte de la variabilité des données.
- L'estimation d'une erreur de prévision est une opération délicate aux conséquences importantes. Il est donc nécessaire d'utiliser le même estimateur pour comparer l'efficacité de deux méthodes ou modèle.

# Outline

# Problème de la sélection de variables

- On a vu au travers des différentes partie du cours qu'une des questions importantes en modélisation est de choisir les variables explicatives.
- Cette question devient même cruciale quand on travaille avec des données comportant un grand nombre de variables (spectrométrie, données omics, ...).
- Solutions proposées plus haut
  - tests sur des modèles imbriqués (ex : déviance, AIC, ... )
  - comparaison de modèles basée sur l'erreur de classification ou la courbe ROC.
- Méthodes pas à pas
  Méthode descendante (ou backward) : on ajuste le modèle introduisant toutes les variables disponibles, puis à chaque étape on retire la variable ayant le moins fort pouvoir explicatif (choix fait d'après un test de Fisher). On s'arrête lorsque toutes les variables restantes sont significatives (ou que le fait de retirer la variable dégrade trop le modèle).
  Méthode ascendante (ou forward) : on ajuste le modèle introduisant uniquement une constante, puis à chaque étape on ajoute la variable ayant le plus fort pouvoir explicatif (choix fait d'après un test de Fisher). On s'arrête lorsque l'ajout d'une nouvelle variable n'améliore pas le modèle.
  Méthode combinée : on peut combiner les méthodes descendantes et ascendantes.

# Méthode descendante

```
> df<-read.table('http://www.indiana.edu/~statmath/stat/all/cdvm/gss_cdvm.csv', se
> df$www = as.factor(df$www)
> df$male = as.factor(df$male)
> df$trust = as.factor(df$trust)
> blm<-glm(trust~(educate+income+age+male+www)^2, data=df, family=binomial)
> summary(blm)
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)    -5.115e+00  2.283e+00  -2.240   0.0251 *
educate         1.329e-01  1.525e-01   0.871   0.3836
income          9.801e-03  8.048e-02   0.122   0.9031
age             5.025e-02  3.205e-02   1.568   0.1170
male1           7.423e-01  9.783e-01   0.759   0.4480
www1            3.456e-01  1.261e+00   0.274   0.7841
educate:income  1.088e-03  5.010e-03   0.217   0.8280
educate:age    -9.966e-04  2.019e-03  -0.494   0.6215
educate:male1   4.877e-02  5.300e-02   0.920   0.3575
educate:www1    9.999e-03  7.495e-02   0.133   0.8939
income:age     -4.522e-05  8.101e-04  -0.056   0.9555
income:male1    3.502e-04  2.456e-02   0.014   0.9886
income:www1     1.097e-02  2.722e-02   0.403   0.6870
age:male1      -1.732e-02  1.007e-02  -1.721   0.0853 .
age:www1        1.787e-03  1.241e-02   0.144   0.8855
male1:www1     -5.715e-01  3.375e-01  -1.693   0.0904 .
```

# Méthode descendante

```
> blm.st = step(blm,direction="backward")
> anova(blm.st)
Analysis of Deviance Table
Terms added sequentially (first to last)
         Df Deviance Resid. Df Resid. Dev
NULL                      1173     1596.6
educate   1   64.810      1172     1531.8
income    1   17.994      1171     1513.8
age       1   29.989      1170     1483.8
male      1    4.381      1169     1479.5
www       1   11.506      1168     1467.9
age:male  1    2.484      1167     1465.5
male:www  1    2.170      1166     1463.3
>summary(blm.st)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.515585   0.552140  -9.989  < 2e-16 ***
educate      0.150519   0.026237   5.737 9.65e-09 ***
income       0.030862   0.011609   2.658 0.007852 **
age          0.036382   0.006986   5.207 1.91e-07 ***
male1        1.348318   0.525191   2.567 0.010250 *
www1         0.777063   0.231423   3.358 0.000786 ***
age:male1   -0.016686   0.009651  -1.729 0.083802 .
male1:www1  -0.479713   0.326321  -1.470 0.141544
    Null deviance: 1596.6  on 1173  degrees of freedom
Residual deviance: 1463.3  on 1166  degrees of freedom
AIC: 1479.3
```

# Régression Lasso

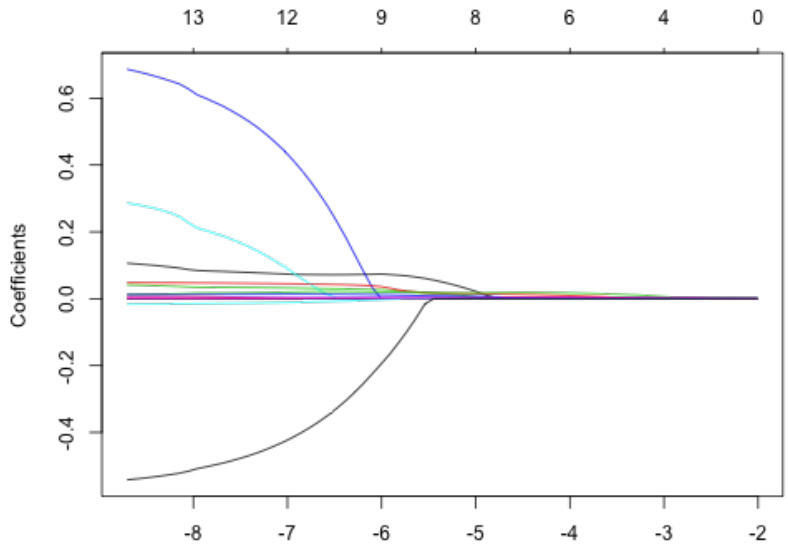
- Les méthodes pas à pas sont optimales si les variables sont orthogonales (aucune dépendance entre elles).
- Une alternative est la régression pénalisée comme la régression Lasso. L'idée est de pénaliser la vraisemblance par un terme qui devient grand si on met trop de variables (inutiles) dans le modèle.
- En régression Lasso, on minimise alors le critère pénalisé suivant

$$-\log \mathcal{L}(\beta) + \lambda \sum_{j=1}^{p} |\beta_j|$$

- La constante $\lambda$ joue le rôle d'une pondération et donne plus ou moins d'importance à chacun des deux termes.
- En pratique, la pénalisation va forcer les paramètre $\beta_j$ à rester nul si la variable $X_j$ n'apporte pas d'information utile.
- Un des avantages de cette approche est qu'on peut l'utiliser même si le nombre de variables est plus grand que le nombre d'observation (ex cas des données OMIC).

# Lasso

```
df<-read.table('http://www.indiana.edu/~statmath/stat/all/cdvm/gss_cdvm.csv', sep=
df1 = df[,-2]
df1$EducInc = df$educate*df$income
df1$EducAge = df$educate*df$age
df1$EducMale = df$educate*df$male
df1$EducWww = df$educate*df$www
df1$IncAge = df$income*df$age
df1$IncMale = df$income*df$male
df1$IncWww = df$income*df$www
df1$AgeMale = df$age*df$male
df1$AgeWww = df$age*df$www
df1$MaleWww = df$male*df$www
library(glmnet)
blm.lasso = glmnet(as.matrix(df1[,-1]),df1[,1],family="binomial")
plot(blm.lasso,xvar="lambda") # Valeurs des coefficients en fonction de lambda
```



# Lasso

- Un des points délicats de la régression Lasso est de bien choisir la valeur de $\lambda$.
- On peut le faire par validation croisée.

```
blm.cv = cv.glmnet(as.matrix(df1[,-1]),df1[,1],family="binomial")
> blm.lasso = glmnet(as.matrix(df1[,-1]),df1[,1],family="binomial",lambda=blm
> blm.lasso$beta
15 x 1 sparse Matrix of class "dgCMatrix"
                    s0
educate  0.025559246
income   .
age      0.006306436
male     .
www      .
EducInc  0.001687337
EducAge  0.001201131
EducMale 0.014927644
EducWww  0.015936374
IncAge   .
IncMale  .
IncWww   0.006594953
AgeMale  .
AgeWww   0.002718073
MaleWww  .
```

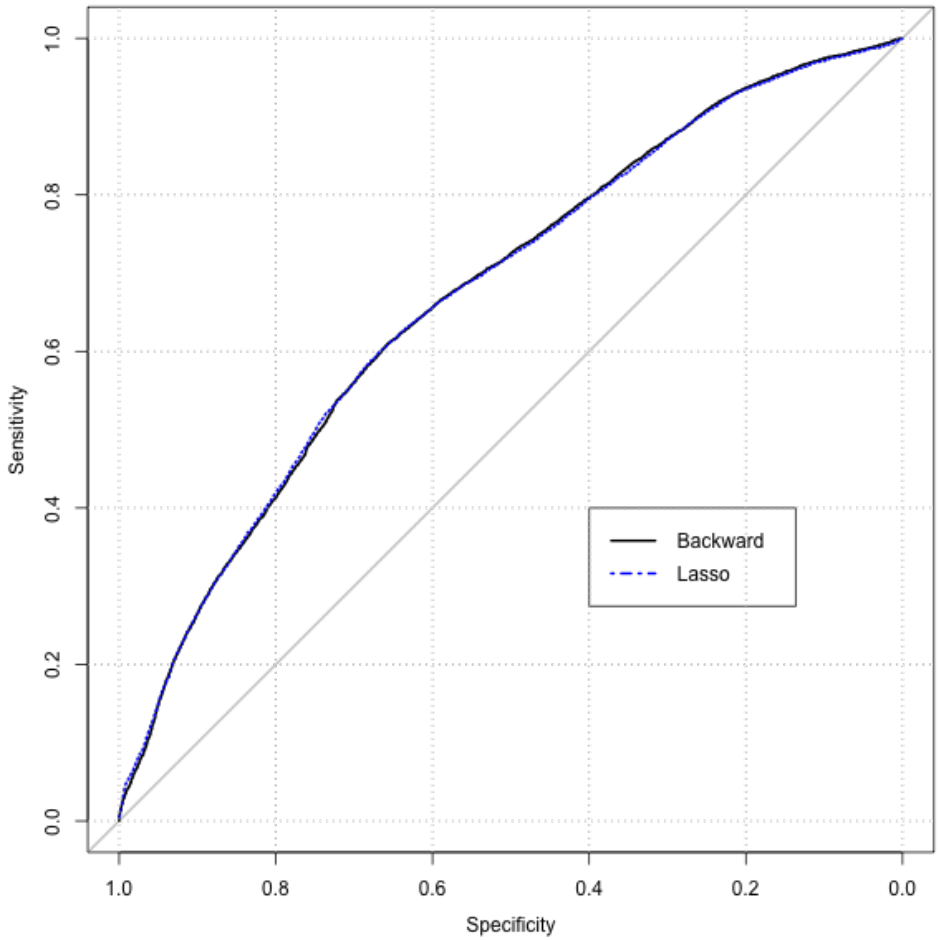
# Validation croisée et sélection de variables

- Si on utilise la validation croisée (Kfold ou bootstrap) pour comparer des méthodes/modèles l'étape de la sélection de variables doit être incluse dans la boucle de validation (ie réalisée pour chaque échantillon d'apprentissage).

```
n = nrow(df)
ntest = round(n/5)
p.st <- p.lasso <- Yobs <- p.st <- NULL
B = 100
for (b in 1:B){
testi = sample(1:n,ntest)
appri = setdiff(1:n,testi)
blm = glm(trust~(educate+income+age+male+www)^2, data=df, family=binomial,sub
blm.st = step(blm,direction="backward")
pmod = predict(blm.st,df[testi,])
blm.cv = cv.glmnet(as.matrix(df1[appri,-1]),df1[appri,1],family="binomial")
blm.lasso = glmnet(as.matrix(df1[appri,-1]),df1[appri,1],family="binomial",la
pmod2 = predict(blm.lasso,as.matrix(df1[testi,-1]),type="response")
Yobs = c(Yobs,df$trust[testi])
p.st = c(p.st,pmod)
p.lasso = c(p.lasso,pmod2)
}
r.st=roc(Yobs,p.st)
plot(r.st)
r.lasso=roc(Yobs,p.lasso)
plot(r.lasso,add=TRUE,col="blue",lty=2)
```

# Validation croisée et sélection de variables

- Dans cette exemple, les deux méthodes conduisent au même résultat, avec une aire sous la courbe ROC d'environ 0.62.



# Conclusion

# Calcul Matriciel

- Transposée

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \quad \mathbf{X}^T = \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & & & \vdots \\ x_{1p} & x_{2p} & \cdots & x_{np} \end{pmatrix}$$

- Produit matrice vecteur

$$\mathbf{X}^T\beta = \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & & & \vdots \\ x_{1p} & x_{2p} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} = \begin{pmatrix} \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} \\ \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} \\ \vdots \\ \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np} \end{pmatrix}$$

Back

# Vraisemblance pour le modèle logistique

- On suppose qu'on observe $n$ couples $(y_i, \mathbf{x}_i)$ avec $y_i \in \{0, 1\}$ et $\mathbf{x}_i \in \mathbb{R}^p$
- Modèle logistique avec lien binomial

$$P(Y_i = y_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i} \text{ et } \pi_i = \pi(x_i) = \frac{e^{\mathbf{x}_i^T \beta}}{1 + e^{\mathbf{x}_i^T \beta}}$$

- Log vraisemblance

$$\begin{aligned} \log \mathcal{L}(\beta) &= \sum_{i=1}^{n} \log P_\beta(Y_i = y_i) \\ &= \sum_{i=1}^{n} \log \left( \pi_i^{y_i}(1 - \pi_i)^{1-y_i} \right) \\ &= \sum_{i=1}^{n} y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i) \end{aligned}$$

Back

# Intervalle de confiance (1/2)

- Soit $\hat{\theta}_n$ un estimateur du paramètre $\theta$ tel que $\hat{\theta}_n$ vérifie un théorème central limit ie que quand $n$ tend vers l'infini,

$$\frac{\hat{\theta}_n - \theta}{\sqrt{Var(\hat{\theta}_n)}} \to Z$$

avec $Z$ une variable aléatoire de loi de Gauss centrée et réduite.

- En général, les estimateurs du maximum de vraisemblance vérifient le théorème central limit.

- On définit un intervalle de confiance au risque $\alpha$ pour $\hat{\theta}_n$ à partir des bornes $-z_{1-\alpha/2}$ et $-z_{1-\alpha/2}$ telles que

$$P\left(-z_{1-\alpha/2} < \frac{\hat{\theta}_n - \theta}{\sqrt{Var(\hat{\theta}_n)}} < z_{1-\alpha/2}\right) = 1 - \alpha$$

# Intervalle de confiance (2/2)

- Si $n$ est assez grand, on peut supposer que $\frac{\hat{\theta}_n - \theta}{\sqrt{Var(\hat{\theta}_n)}}$ suit approximativement une loi de Gauss et on a donc

$$P\left(-z_{1-\alpha/2} < \frac{\hat{\theta}_n - \theta}{\sqrt{Var(\hat{\theta}_n)}} < z_{1-\alpha/2}\right) = \Phi(z_{1-\alpha/2}) - \Phi(-z_{1-\alpha/2})$$
$$= 2\Phi(z_{1-\alpha/2}) - 1$$

avec $\Phi$ la fonction de répartition de la loi de Gauss centrée réduite. On a donc

$$2\Phi(z_{1-\alpha/2}) - 1 = 1 - \alpha$$

soit

$$z_{1-\alpha/2} = \Phi^{-1}(1 - \alpha/2)$$

- Les bornes de l'intervalle de confiance pour $\hat{\theta}_n$ s'écrivent alors

$$\mu_- = \hat{\theta}_n - \Phi^{-1}(1 - \alpha/2)\sqrt{Var(\hat{\theta}_n)}$$

$$\mu_+ = \hat{\theta}_n + \Phi^{-1}(1 - \alpha/2)\sqrt{Var(\hat{\theta}_n)}$$

Back