La régression logistique

Celine Helbert

Plan

- La variable à expliquer est qualitative,
- n observations réparties en k classes, décrites par p variables explicatives.

Exemples :

- $Y =$ Crédit accordé,
  $X_1 =$ Age , $X_2 =$ Type emploi, $X_3 =$ Niveau revenus
- $Y =$ Maladie cardiovasculaire,
  $X_1 =$ Age , $X_2 =$ Sexe, $X_3 =$ Tabac, $X_4 =$ Alcool.

- La variable à expliquer est qualitative,
- n observations réparties en k classes, décrites par p variables explicatives.

Exemples :

- $Y=$ Crédit accordé,
  $X_1 =$ Age , $X_2 =$ Type emploi, $X_3 =$ Niveau revenus
- $Y=$ Maladie cardiovasculaire,
  $X_1 =$ Age , $X_2 =$ Sexe, $X_3 =$ Tabac, $X_4 =$ Alcool.

Rappel du modèle pour la régression linéaire (ANOVA, Plans d'expériences) :

$\mathbf{y_i} \rightsquigarrow \mathcal{N}(x_i\beta, \sigma^2)$ où $\mathbf{y_1}, ..., \mathbf{y_n}$ sont indépendantes.

Ici le modèle gaussien ne convient plus, la variable $\mathbf{y_i}$ est discrête.

Quelle loi peut-on proposer (cas à 2 classes $\{0, 1\}$) ?
La loi binomiale $B(p_i)$, où $p_i = P(\mathbf{y}_i = 1|X = x_i)$

Rappel du modèle pour la régression linéaire (ANOVA, Plans d'expériences) :

$\mathbf{y_i} \rightsquigarrow \mathcal{N}(x_i\beta, \sigma^2)$ où $\mathbf{y_1}, ..., \mathbf{y_n}$ sont indépendantes.

Ici le modèle gaussien ne convient plus, la variable $\mathbf{y_i}$ est discrête.

Quelle loi peut-on proposer (cas à 2 classes $\{0, 1\}$) ?
La loi binomiale $B(p_i)$, où $p_i = P(\mathbf{y}_i = 1|X = x_i)$

Rappel du modèle pour la régression linéaire (ANOVA, Plans d'expériences) :

$\mathbf{y_i} \rightsquigarrow \mathcal{N}(x_i\beta, \sigma^2)$ où $\mathbf{y_1}, ..., \mathbf{y_n}$ sont indépendantes.

Ici le modèle gaussien ne convient plus, la variable $\mathbf{y_i}$ est discrête.

Quelle loi peut-on proposer (cas à 2 classes $\{0, 1\}$) ?
La loi binomiale $B(p_i)$, où $p_i = P(\mathbf{y}_i = 1|X = x_i)$

Rappel du modèle pour la régression linéaire (ANOVA, Plans d'expériences) :

$\mathbf{y_i} \rightsquigarrow \mathcal{N}(x_i\beta, \sigma^2)$ où $\mathbf{y_1}, ..., \mathbf{y_n}$ sont indépendantes.

Ici le modèle gaussien ne convient plus, la variable $\mathbf{y_i}$ est discrête.

Quelle loi peut-on proposer (cas à 2 classes $\{0, 1\}$) ?
La loi binomiale $B(p_i)$, où $p_i = P(\mathbf{y}_i = 1|X = x_i)$

La régression logistique binaire, k=2, propose le modèle suivant. On supposera que l'on cherche à prédire **y** qui prend deux valeurs 0 ou 1. En fait on cherche à quantifier $P(\mathbf{y} = 1|X = x)$.

La règle de décision sera alors la suivante : si $P(\mathbf{y} = 1|X = x) > 0.5$ alors $Y = 1$ sinon $Y = 0$.

La régression logistique binaire, k=2, propose le modèle suivant. On supposera que l'on cherche à prédire **y** qui prend deux valeurs 0 ou 1. En fait on cherche à quantifier $P(\mathbf{y} = 1|X = x)$.
La règle de décision sera alors la suivante : si
$P(\mathbf{y} = 1|X = x) > 0.5$ alors $Y = 1$ sinon $Y = 0$.

Une première possibilité de modélisation pourrait consister à écrire :

$$P(\mathbf{y} = 1|X = x) = \pi(x) = \beta_0 + \beta'\mathbf{x}$$

Dans ce cas, $\pi(x)$ appartient à $\mathbb{R}$, ce qui n'est pas le cas.

Une première possibilité de modélisation pourrait consister à écrire :

$$P(\mathbf{y} = 1|X = x) = \pi(x) = \beta_0 + \beta'\mathbf{x}$$

Dans ce cas, $\pi(x)$ appartient à $\mathbb{R}$, ce qui n'est pas le cas.

Définition

La régression logistique consiste à écrire :

$$\log\left(\frac{P(\mathbf{y}=1|X=x)}{P(\mathbf{y}=0|X=x)}\right) = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta'\mathbf{x}$$

$$\Leftrightarrow \pi(x) = \frac{e^{\beta_0+\beta'\mathbf{x}}}{1+e^{\beta_0+\beta'\mathbf{x}}}$$

La fonction LOGIT est $p \to \log\left(\frac{p}{1-p}\right)$.

Une modélisation alternative (Régression PROBIT) également utilisée en pratique consiste à écrire

$$\pi(x) = \Phi(\beta_0 + \beta' \mathbf{x})$$

où Φ est la fonction de répartition de la loi normale.

$$\Leftrightarrow \Phi^{-1}(\pi(x)) = \beta_0 + \beta' \mathbf{x}$$

Plan

La probabilité d'être malade si l'on fume est

$$P(\mathbf{y} = 1|X = 1) = \frac{e^{\beta_0+\beta_1}}{1 + e^{\beta_0+\beta_1}}$$

que l'on compare à la probabilité de ne pas être malade si l'on fume $P(\mathbf{y} = 0/X = 1) = 1 - P(\mathbf{y} = 1/X = 1)$.
L'odds est le rapport de ces deux probabilités. On fait de même pour les non-fumeurs et on définit l'odds-ratio comme le rapport des deux odds

$$O.R = \frac{P(\mathbf{y} = 1|X = 1)/(\mathbf{y} = 0|X = 1)}{P(\mathbf{y} = 1|X = 0)/(\mathbf{y} = 0|X = 0)} = e^{\beta_1}$$

C'est le facteur par lequel la cote (ou odds) est multiplié si l'on passe de $X = 0$ à $X = 1$.

Dans notre exemple, un $O.R = e^{\beta_1} > 1$ indique que le fait de fumer est un facteur aggravant. En effet le risque d'être malade est e^{β_1} fois supérieur pour les fumeurs que pour les non-fumeurs.Par la suite, on regarde le signe de β_1. L'effet de X est aggravant si $\beta_1 > 0$

Remarques : si X n'a pas d'effet sur $\mathbf{y}$, $O.R. = 1$ et $\beta_1 = 0$

C'est le facteur par lequel la cote (ou odds) est multiplié si l'on passe de $X = 0$ à $X = 1$.
Dans notre exemple, un $O.R = e^{\beta_1} > 1$ indique que le fait de fumer est un facteur aggravant. En effet le risque d'être malade est e^{β_1} fois supérieur pour les fumeurs que pour les non-fumeurs. Par la suite, on regarde le signe de β_1. L'effet de X est aggravant si $\beta_1 > 0$
Remarques : si X n'a pas d'effet sur $\mathbf{y}$, $O.R. = 1$ et $\beta_1 = 0$

C'est le facteur par lequel la cote (ou odds) est multiplié si l'on passe de $X = 0$ à $X = 1$.
Dans notre exemple, un $O.R = e^{\beta_1} > 1$ indique que le fait de fumer est un facteur aggravant. En effet le risque d'être malade est e^{β_1} fois supérieur pour les fumeurs que pour les non-fumeurs.Par la suite, on regarde le signe de β_1. L'effet de X est aggravant si $\beta_1 > 0$

Remarques : si X n'a pas d'effet sur $\mathbf{y}$, $O.R. = 1$ et $\beta_1 = 0$

C'est le facteur par lequel la cote (ou odds) est multiplié si l'on passe de $X = 0$ à $X = 1$.
Dans notre exemple, un $O.R = e^{\beta_1} > 1$ indique que le fait de fumer est un facteur aggravant. En effet le risque d'être malade est e^{β_1} fois supérieur pour les fumeurs que pour les non-fumeurs.Par la suite, on regarde le signe de β_1. L'effet de X est aggravant si $\beta_1 > 0$
Remarques : si X n'a pas d'effet sur $\mathbf{y}$, $O.R. = 1$ et $\beta_1 = 0$

On peut sans difficulté utiliser des prédicteurs qualitatifs à m modalités. Chaque variable est remplacée par (m-1) indicatrices après élimination d'une des modalités dite de référence, qui aura un coefficient nul. Les comparaisons de coefficients se font alors par rapport à cette modalité : une valeur proche de 1 signifie que la modalité est proche de la modalité de référence.
Sous R la modalité de référence est la première modalité rencontrée.

On peut sans difficulté utiliser des prédicteurs qualitatifs à m modalités. Chaque variable est remplacée par (m-1) indicatrices après élimination d'une des modalités dite de référence, qui aura un coefficient nul. Les comparaisons de coefficients se font alors par rapport à cette modalité : une valeur proche de 1 signifie que la modalité est proche de la modalité de référence.
Sous R la modalité de référence est la première modalité rencontrée.

Plan

Elle s'effectue par la méthode du maximum de vraisemblance à partir d'un échantillon i.i.d de n observations (y_i). On fait le calcul dans le cas logit.

$$\begin{aligned} L(\beta_0, \beta) &= P(\mathbf{y}_1 = y_1, \ldots, \mathbf{y}_n = y_n | X_1 = x_1, \ldots, X_n = x_n) \\ &= \prod_{i=1}^{n} P(\mathbf{y}_i = y_i | X_i = x_i) \\ &= \prod_{i=1}^{n} \left(\frac{e^{\beta_0 + \beta' \mathbf{x_i}}}{1 + e^{\beta_0 + \beta' \mathbf{x_i}}} \right)^{y_i} * \left(1 - \frac{e^{\beta_0 + \beta' \mathbf{x_i}}}{1 + e^{\beta_0 + \beta' \mathbf{x_i}}} \right)^{(1-y_i)} \\ &= \prod_{i=1}^{n} \pi(x_i)^{y_i} (1 - \pi(x_i))^{(1-y_i)} \end{aligned}$$

Elle s'effectue par la méthode du maximum de vraisemblance à partir d'un échantillon i.i.d de n observations (y_i). On fait le calcul dans le cas logit.

$$\begin{aligned} L(\beta_0, \beta) &= P(\mathbf{y}_1 = y_1, \ldots, \mathbf{y}_n = y_n | X_1 = x_1, \ldots, X_n = x_n) \\ &= \prod_{i=1}^{n} P(\mathbf{y}_i = y_i | X_i = x_i) \\ &= \prod_{i=1}^{n} \left(\frac{e^{\beta_0 + \beta' \mathbf{x_i}}}{1 + e^{\beta_0 + \beta' \mathbf{x_i}}} \right)^{y_i} * \left(1 - \frac{e^{\beta_0 + \beta' \mathbf{x_i}}}{1 + e^{\beta_0 + \beta' \mathbf{x_i}}} \right)^{(1-y_i)} \\ &= \prod_{i=1}^{n} \pi(x_i)^{y_i} (1 - \pi(x_i))^{(1-y_i)} \end{aligned}$$

Elle s'effectue par la méthode du maximum de vraisemblance à partir d'un échantillon i.i.d de n observations (y_i). On fait le calcul dans le cas logit.

$$\begin{aligned} L(\beta_0, \beta) &= P(\mathbf{y}_1 = y_1, \ldots, \mathbf{y}_n = y_n | X_1 = x_1, \ldots, X_n = x_n) \\ &= \prod_{i=1}^{n} P(\mathbf{y}_i = y_i | X_i = x_i) \\ &= \prod_{i=1}^{n} \left(\frac{e^{\beta_0 + \beta' \mathbf{x_i}}}{1 + e^{\beta_0 + \beta' \mathbf{x_i}}} \right)^{y_i} * \left(1 - \frac{e^{\beta_0 + \beta' \mathbf{x_i}}}{1 + e^{\beta_0 + \beta' \mathbf{x_i}}} \right)^{(1-y_i)} \\ &= \prod_{i=1}^{n} \pi(x_i)^{y_i} (1 - \pi(x_i))^{(1-y_i)} \end{aligned}$$

En annulant les dérivées de la log-vraisemblance, on obtient le système

$$\begin{cases} \frac{\partial l(\beta)}{\partial \beta_0} & = & \sum (y_i - \pi(x_i)) = 0 \\ \frac{\partial l(\beta)}{\partial \beta_j} & = & \sum x_i^j (y_i - \pi(x_i)) = 0 \text{ j=1,..,p} \end{cases}$$

qui n'a pas de solution analytique et se résout par des algorithmes numériques. Remarque : dans le cas Probit, il suffit de remplacer l'expression de $\pi(x_i)$.

En annulant les dérivées de la log-vraisemblance, on obtient le système

$$\begin{cases} \frac{\partial l(\beta)}{\partial \beta_0} & = & \sum (y_i - \pi(x_i)) = 0 \\ \frac{\partial l(\beta)}{\partial \beta_j} & = & \sum x_i^j (y_i - \pi(x_i)) = 0 \text{ j=1,..,p} \end{cases}$$

qui n'a pas de solution analytique et se résout par des algorithmes numériques. Remarque : dans le cas Probit, il suffit de remplacer l'expression de $\pi(x_i)$.

La théorie de l'estimation par maximum de vraisemblance donne le théorème suivant :

$$(J(\hat{\beta}))^{1/2}(\hat{\beta} - \beta) \rightsquigarrow \mathcal{N}(0_n, Id_n)$$

où J la matrice d'information de fisher définie par
$J(\hat{\beta}) = \left[-\frac{\partial^2 l(\beta)}{\partial \beta^2}\right]_{\beta=\hat{\beta}}$.
On obtient la matrice de variance-covariance asymptotique des estimateurs comme inverse de la matrice J

$$\hat{V}(\hat{\beta}) = J(\hat{\beta})^{-1} = (\mathbf{X}'\hat{\Sigma}\mathbf{X})^{-1}$$

$$\text{où } \hat{\Sigma} = \begin{bmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & & 0 \\ & \ldots & \\ 0 & & \hat{\pi}_n(1-\hat{\pi}_n) \end{bmatrix}$$

Plan

Remarques :

- Dans les deux modèles, Probit et Logit, on ne peut pas estimer séparément **b** et la variance des erreurs.
- Les deux fonction Probit et Logit ne conduisent pas aux mêmes valeurs des paramètres. La seule information réellement utilisable est le signe des paramètres, indiquant si la variable associée influence la probabilité à la hausse ou à la baisse.

Plan

Comme en régression linéraire, l'objectif est de sélectioner les variables réellement influentes sur la réponse. On distingue trois types d'hypothèses :

1. Evaluer la contribution individuelle d'une variable

$$H_0 : \beta_j = 0$$

2. Evaluer la contribution d'un ensemble de "q" variables

$$H_0 : \beta_j = ... = \beta_{j+q} = 0$$

3. Evaluer la qualité du modèle dans son ensemble

$$H_0 : \beta_1 = ... = \beta_p = 0$$

Deux approches pour tester ces hypothèses :

- Tests basés sur le rapport de vraisemblance. C'est un test qui mesure l'accroissement de vraisemblance entre le modèle "réduit" et le modèle "complet". Ce test est puissant et cohérent avec l'estimation des paramètres, mais couteux car il nécessite l'évaluation de la vraisemblance pour chaque sous-modèle.
- Tests de Wald. C'est un test basé sur la normalité asymptotique des estimateurs par maximum de vraisemblance. Plus simple à mettre en oeuvre mais moins puissant.

Deux approches pour tester ces hypothèses :

- Tests basés sur le rapport de vraisemblance. C'est un test qui mesure l'accroissement de vraisemblance entre le modèle "réduit" et le modèle "complet". Ce test est puissant et cohérent avec l'estimation des paramètres, mais couteux car il nécessite l'évaluation de la vraisemblance pour chaque sous-modèle.
- Tests de Wald. C'est un test basé sur la normalité asymptotique des estimateurs par maximum de vraisemblance. Plus simple à mettre en oeuvre mais moins puissant.

Deux approches pour tester ces hypothèses :

- Tests basés sur le rapport de vraisemblance. C'est un test qui mesure l'accroissement de vraisemblance entre le modèle "réduit" et le modèle "complet". Ce test est puissant et cohérent avec l'estimation des paramètres, mais couteux car il nécessite l'évaluation de la vraisemblance pour chaque sous-modèle.
- Tests de Wald. C'est un test basé sur la normalité asymptotique des estimateurs par maximum de vraisemblance. Plus simple à mettre en oeuvre mais moins puissant.

$H_0 : \beta_j = 0$

On note M le modèle complet et M_{-j} le modèle sans la variable X_j. On note

$$LR = -2ln\left(\frac{L(M_{-j})}{L(M)}\right) = D_{-j} - D$$

où L est la vraisemblance et D la déviance.

Sous l'hypothèse H_0, LR suit une loi du χ^2 à 1 degré de liberté.

Remarques :

- $LR \geq 0$ car plus il y a de termes dans le modèle et plus la déviance est faible.
- Pour $H_0 : \beta_j = ... = \beta_{j+q} = 0$, $LR \rightsquigarrow \chi^2(q)$

$H_0 : \beta_j = 0$

On note M le modèle complet et M_{-j} le modèle sans la variable X_j. On note

$$LR = -2ln\left(\frac{L(M_{-j})}{L(M)}\right) = D_{-j} - D$$

où L est la vraisemblance et D la déviance.
Sous l'hypothèse H_0, LR suit une loi du χ^2 à 1 degré de liberté.
Remarques :

- $LR \geq 0$ car plus il y a de termes dans le modèle et plus la déviance est faible.
- Pour $H_0 : \beta_j = ... = \beta_{j+q} = 0$, $LR \rightsquigarrow \chi^2(q)$

Asymptotiquement $\hat{\beta}$ suit une loi normale multidimensionnelle de matrice de variances-covariances :

$$V(\hat{\beta}) = \left(X'\hat{\Sigma}X\right)^{-1} \text{ avec}$$

$$\hat{\Sigma} = \begin{bmatrix} \hat{\pi}(x_1)(1-\hat{\pi}(x_1)) & & 0 \\ & \ldots & \\ 0 & & \hat{\pi}(x_n)(1-\hat{\pi}(x_n)) \end{bmatrix}$$

- Sous l'hypothèse $H_0 : \beta_j = 0$, la statistique de Wald $\frac{\hat{\beta}_j^2}{s_{\beta_j}^2}$ suit une loi du $\chi^2(1)$.
- Sous l'hypothèse $H_0 : \beta_1 = ... = \beta_q = 0$, la statistique de Wald $\hat{\beta}'_{(q)} V(\hat{\beta})^{-1}_{(q)} \hat{\beta}_{(q)}$ suit une loi du $\chi^2(q)$.

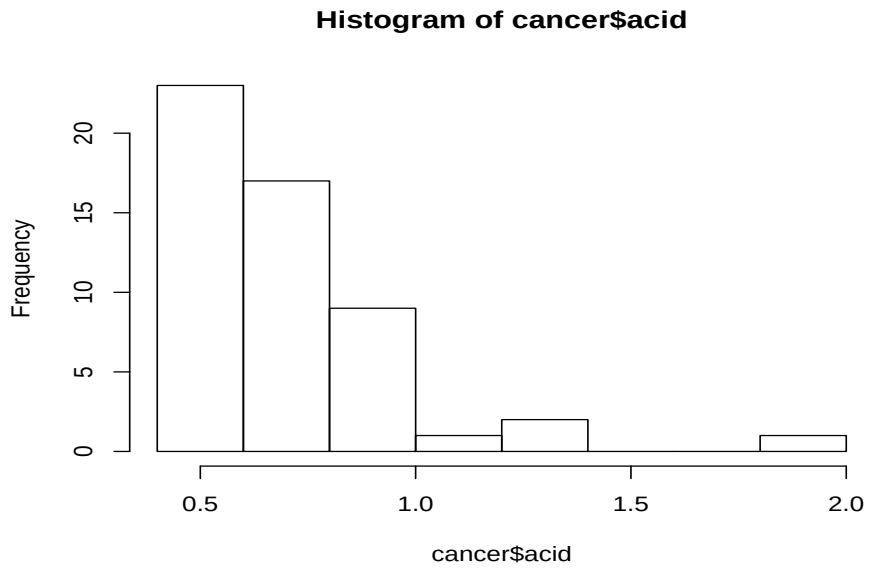
Exemple

L'objectif est de prévoir l'implication ou non du système lymphatique dans le cas de cancer observé et donc de construire un meilleur modèle de prédiction. Les variables considérées sont les suivantes :

- age : âge du patient
- acid : niveau de serum acid phosphatase,
- radio : résultat d'une analyse radiographique (0 : négatif, 1 : positif),
- taille : taille de la tumeur (0 : petite, 1 : grande),
- gravite : résultat de la biopsie (0 : moins sérieux, 1 : sérieux).
- lymph : La sixième variable indique l'implication (1) ou non (0) du système lymphatique.

Etude préalable

Une étude au préalable des variables met en évidence des transformations à réaliser.



On introduit dans le modèle plutôt le log de la variable acid

Régression logistique

- Déclarer avec as.factor les variables qualitatives ou discrètes (radio, taille, gravité).
- Utiliser la fonction régression généralisée (glm avec l'option family=bibomial)

```
> l.acid=log(cancer$acid)
> cancer$radio=as.factor(cancer$radio)
> cancer$taille=as.factor(cancer$taille)
> cancer$gravite=as.factor(cancer$gravite)
> mod_logit=glm(lymph~(age+l.acid+radio+taille+gravite)^2,
              family=binomial(link="logit"),data=cancer)
> mod_probit=glm(lymph~(age+l.acid+radio+taille+gravite)^2,
              family=binomial(link="probit"),data=cancer)
```

Résultat de la régression logistique-link=Logit

```
> Anova(mod_logit,test.statistic="LR",type = "III")
Analysis of Deviance Table (Type III tests)

Response: lymph
               LR Chisq Df Pr(>Chisq)
age              1.6659  1   0.196807
l.acid           1.2479  1   0.263960
radio            0.0177  1   0.894210
taille           0.3407  1   0.559436
gravite          3.8290  1   0.050374 .
age:l.acid       1.0119  1   0.314453
age:radio        0.0748  1   0.784418
age:taille       0.5296  1   0.466783
age:gravite      1.0212  1   0.312235
l.acid:radio     0.0017  1   0.966764
l.acid:taille    1.4986  1   0.220889
l.acid:gravite   1.4646  1   0.226202
radio:taille     0.0859  1   0.769505
radio:gravite    0.0461  1   0.829915
taille:gravite   8.0089  1   0.004655 **
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1
```

Résultat de la régression logistique-link=Probit

```
> Anova(mod_probit,test.statistic="LR",type = "III")
Analysis of Deviance Table (Type III tests)

Response: lymph
               LR Chisq Df Pr(>Chisq)
age              1.7318  1   0.188179
l.acid           1.3198  1   0.250620
radio            0.0206  1   0.885767
taille           0.3031  1   0.581922
gravite          3.7590  1   0.052525 .
age:l.acid       1.0871  1   0.297109
age:radio        0.0828  1   0.773505
age:taille       0.4768  1   0.489876
age:gravite      1.0727  1   0.300344
l.acid:radio     0.0149  1   0.902698
l.acid:taille    1.4720  1   0.225029
l.acid:gravite   1.5986  1   0.206100
radio:taille     0.1549  1   0.693942
radio:gravite    0.0473  1   0.827767
taille:gravite   8.2624  1   0.004048 **
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1
```

On peut faire une sélection backward sur le test du χ^2 en enlevant la variable la moins influente.

```
> Anova(etape9,test.statistic="LR",type = "III")
Analysis of Deviance Table (Type III tests)

Response: lymph
              LR Chisq Df Pr(>Chisq)
l.acid          1.4349  1  0.2309665
radio           5.6945  1  0.0170182 *
taille         10.0862  1  0.0014938 **
gravite        11.9409  1  0.0005492 ***
l.acid:gravite  4.1672  1  0.0412149 *
taille:gravite  6.8699  1  0.0087659 **
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1
```

On peut faire une sélection backward sur le test du χ^2 en enlevant la variable la moins influente.

```
> Anova(etape9,test.statistic="LR",type = "III")
Analysis of Deviance Table (Type III tests)

Response: lymph
               LR Chisq Df Pr(>Chisq)
l.acid           1.4349  1  0.2309665
radio            5.6945  1  0.0170182 *
taille          10.0862  1  0.0014938 **
gravite         11.9409  1  0.0005492 ***
l.acid:gravite   4.1672  1  0.0412149 *
taille:gravite   6.8699  1  0.0087659 **
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1
```

```
> summary(etape9)
Call:
glm(formula = lymph ~ (age + l.acid + radio + taille + gravite)^2 -
    l.acid:taille - age - l.acid:radio - radio:taille - radio:gravite -
    radio:age - age:taille - age:l.acid - age:gravite, family = binomial(link =
    data = cancer)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)        -2.553      1.040  -2.455  0.01408 *
l.acid              1.709      1.420   1.203  0.22902
radio1              2.340      1.084   2.158  0.03092 *
taille1             3.138      1.171   2.681  0.00735 **
gravite1            9.961      4.666   2.135  0.03278 *
l.acid:gravite1    10.426      6.640   1.570  0.11635
taille1:gravite1   -5.648      2.434  -2.320  0.02034 *
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 70.252  on 52  degrees of freedom
```

```
> Anova(etape9,test.statistic="Wald",type = "III")
Analysis of Deviance Table (Type III tests)

Response: lymph
               Df  Chisq Pr(>Chisq)
(Intercept)     1 6.0286   0.014076 *
l.acid          1 1.4469   0.229025
radio           1 4.6576   0.030916 *
taille          1 7.1855   0.007349 **
gravite         1 4.5573   0.032779 *
l.acid:gravite  1 2.4658   0.116347
taille:gravite  1 5.3823   0.020342 *
Residuals      46
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1
```

```
> exp(etape9$coef)
     (Intercept)           l.acid           radio1           taille1
    7.787017e-02     5.520827e+00     1.038589e+01     2.305661e+01
        gravite1  l.acid:gravite1 taille1:gravite1
    2.118736e+04     3.372472e+04     3.525473e-03
> etape9$fitted.values
          1           2           3           4           5           6
0.021737930 0.028103690 0.023271633 0.024844422 0.023271633 0.022499839
          7           8           9          10          11          12
0.176682822 0.263277665 0.592094827 0.225537028 0.033264001 0.041572044
         13          14          15          16          17          18
0.035960303 0.992529501 0.147620836 0.022499839 0.268439092 0.048465885
         19          20          21          22          23          24
0.053602775 0.069964865 0.024844422 0.045467449 0.999315743 0.184932630
         25          26          27          28          29          30
0.577637283 0.052561844 0.001985131 0.354567705 0.028963024 0.272838623
         31          32          33          34          35          36
0.086605715 0.181843489 0.841799859 0.282599930 0.346708766 0.017850718
         37          38          39          40          41          42
0.836516162 0.650007793 0.529056521 0.621896125 0.464206740 0.994079603
         43          44          45          46          47          48
0.990825496 0.980332990 0.638897574 0.985573396 0.493963073 0.475273469
```

```
> predict(etape9,newdata =cancer[,-6] )
          1           2           3           4           5           6
-3.80671894 -3.54334823 -3.73697343 -3.66996376 -3.73697343 -3.77149031
          7           8           9          10          11          12
-1.53898536 -1.02900183  0.37263208 -1.23368550 -3.36944969 -3.13786646
         13          14          15          16          17          18
-3.28871683  4.88929493 -1.75338439 -3.77149031 -1.00255644 -2.97721540
         19          20          21          22          23          24
-2.87106154 -2.58722918 -3.66996376 -3.04422508  7.28649176 -1.48327917
         25          26          27          28          29          30
 0.31308167 -2.89177122 -6.22008309 -0.59902100 -3.51234455 -0.98026793
         31          32          33          34          35          36
-2.35580184 -1.50390728  1.67168134 -0.93160144 -0.63353789 -4.00769961
         37          38          39          40          41          42
 1.63253171  0.61907346  0.11635719  0.49760412 -0.14341836  5.12341383
         43          44          45          46          47          48
 4.68211015  3.90894971  0.57058250  4.22414961 -0.02414888 -0.09898687
         49          50          51          52          53
 0.24618121  0.03905897  2.36442969  2.72658673 10.04353585
```

```
> exp(predict(etape9,newdata =cancer[,-6]))/(1+exp(predict(etape9,newdata =cance
          1           2           3           4           5           6
0.021737930 0.028103690 0.023271633 0.024844422 0.023271633 0.022499839
          7           8           9          10          11          12
0.176682822 0.263277665 0.592094827 0.225537028 0.033264001 0.041572044
         13          14          15          16          17          18
0.035960303 0.992529501 0.147620836 0.022499839 0.268439092 0.048465885
         19          20          21          22          23          24
0.053602775 0.069964865 0.024844422 0.045467449 0.999315743 0.184932630
         25          26          27          28          29          30
0.577637283 0.052561844 0.001985131 0.354567705 0.028963024 0.272838623
         31          32          33          34          35          36
0.086605715 0.181843489 0.841799859 0.282599930 0.346708766 0.017850718
         37          38          39          40          41          42
0.836516162 0.650007793 0.529056521 0.621896125 0.464206740 0.994079603
         43          44          45          46          47          48
0.990825496 0.980332990 0.638897574 0.985573396 0.493963073 0.475273469
         49          50          51          52          53
0.561236344 0.509763501 0.914074363 0.938577357 0.999956536
```

Sélection automatique avec le critère AIC

Méthode Backward

```
> backward <- step(mod_logit,method = "backward",k= 2)
> backward$anova
> Anova(backward,test.statistic="LR",type = "III")
```

Méthode forward

```
> constant = glm(lymph~1,family=binomial,data=cancer)
> forward <- step(constant, direction="forward", scope=list(lower=~1,
+ upper=~(age+l.acid+radio+taille+gravite)*(age+l.acid+radio+taille+gravite)),
+ trace = TRUE)
> forward$anova
> Anova(forward,test.statistic="LR",type = 'III')
```

Résultat méthode backward

```
               Step Df    Deviance Resid. Df Resid. Dev      AIC
1                   NA          NA        37   30.17407 62.17407
2    - l.acid:radio  1 0.001736144        38   30.17580 60.17580
3  - radio:gravite   1 0.058491134        39   30.23430 58.23430
4   - radio:taille   1 0.042399111        40   30.27670 56.27670
5      - age:radio   1 0.083178600        41   30.35987 54.35987
6     - age:taille   1 0.799502943        42   31.15938 53.15938
7     - age:l.acid   1 1.176430869        43   32.33581 52.33581
8    - age:gravite   1 0.710915607        44   33.04672 51.04672
9             - age  1 1.320311701        45   34.36703 50.36703
10 - l.acid:taille  1 1.920111975        46   36.28715 50.28715

Analysis of Deviance Table (Type III tests)

Response: lymph
               LR Chisq Df Pr(>Chisq)
l.acid           1.4349  1  0.2309665
radio            5.6945  1  0.0170182 *
taille          10.0862  1  0.0014938 **
gravite         11.9409  1  0.0005492 ***
l.acid:gravite   4.1672  1  0.0412149 *
taille:gravite   6.8699  1  0.0087659 **
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1
```

Résultat méthode forward

```
      Step Df  Deviance Resid. Df Resid. Dev      AIC
1           NA        NA        52   70.25215 72.25215
2  + radio -1 11.251350        51   59.00080 63.00080
3 + taille -1  5.647384        50   53.35342 59.35342
4 + l.acid -1  4.367406        49   48.98601 56.98601
Analysis of Deviance Table (Type III tests)

Response: lymph
        LR Chisq Df Pr(>Chisq)
radio     7.4983  1   0.006176 **
taille    6.2856  1   0.012172 *
l.acid    4.3674  1   0.036633 *
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1
```

Plan

Pour prédire la variable $\hat{y}$, la règle d'affectation est la suivante : On affecte un nouvel individu $\mathbf{x}$ à la classe telle que la probabilité à postériori est maximale, i.e

$$\hat{y} = argmax_{1 \leq i \leq k} \hat{\pi}(x)$$

Si la variable Y est binaire, cela revient à prédire $\hat{y} = 1$ si $\hat{\pi}(x) > 0.5$. On appelle probabilité de coupure la valeur 0.5. On peut éventuellement modifier cette probabilité. On dresse ensuite la matrice de confusion (ou cross-over) qui croisent les valeurs de Y et les valeurs prédites $\hat{y}$.

Pour prédire la variable $\hat{y}$, la règle d'affectation est la suivante : On affecte un nouvel individu **x** à la classe telle que la probabilité à postériori est maximale, i.e

$$\hat{y} = argmax_{1 \leq i \leq k} \hat{\pi}(x)$$

Si la variable Y est binaire, cela revient à prédire $\hat{y} = 1$ si $\hat{\pi}(x) > 0.5$. On appelle probabilité de coupure la valeur 0.5. On peut éventuellement modifier cette probabilité. On dresse ensuite la matrice de confusion (ou cross-over) qui croisent les valeurs de Y et les valeurs prédites $\hat{y}$.

Pour prédire la variable $\hat{y}$, la règle d'affectation est la suivante : On affecte un nouvel individu **x** à la classe telle que la probabilité à postériori est maximale, i.e

$$\hat{y} = argmax_{1 \leq i \leq k} \hat{\pi}(x)$$

Si la variable Y est binaire, cela revient à prédire $\hat{y} = 1$ si $\hat{\pi}(x) > 0.5$. On appelle probabilité de coupure la valeur 0.5. On peut éventuellement modifier cette probabilité. On dresse ensuite la matrice de confusion (ou cross-over) qui croisent les valeurs de Y et les valeurs prédites $\hat{y}$.

Pour prédire la variable $\hat{y}$, la règle d'affectation est la suivante : On affecte un nouvel individu **x** à la classe telle que la probabilité à postériori est maximale, i.e

$$\hat{y} = argmax_{1 \leq i \leq k} \hat{\pi}(x)$$

Si la variable Y est binaire, cela revient à prédire $\hat{y} = 1$ si $\hat{\pi}(x) > 0.5$. On appelle probabilité de coupure la valeur 0.5. On peut éventuellement modifier cette probabilité. On dresse ensuite la matrice de confusion (ou cross-over) qui croisent les valeurs de Y et les valeurs prédites $\hat{y}$.

Pour prédire la variable $\hat{y}$, la règle d'affectation est la suivante : On affecte un nouvel individu **x** à la classe telle que la probabilité à postériori est maximale, i.e

$$\hat{y} = argmax_{1 \leq i \leq k} \hat{\pi}(x)$$

Si la variable Y est binaire, cela revient à prédire $\hat{y} = 1$ si $\hat{\pi}(x) > 0.5$. On appelle probabilité de coupure la valeur 0.5. On peut éventuellement modifier cette probabilité. On dresse ensuite la matrice de confusion (ou cross-over) qui croisent les valeurs de Y et les valeurs prédites $\hat{y}$.

Exemple

Sur la prévision du cancer lymphatique, on dresse les matrices de confusions pour les deux modèles obtenus

Pour le modèle obtenu par sélection pas à pas et sélection backward par le critère d'AIC.

```
> table(backward$fitted.values>0.5,cancer$lymph)
       0  1
FALSE 29  5
TRUE   4 15
```

Pour le modèle obtenu par sélection forward en utilisant le critère d'AIC, on obtient

```
> table(forward$fitted.values>0.5,cancer$lymph)

        0  1
  FALSE 29  9
  TRUE   4 11
```

Le meilleur modèle est le premier modèle.

La comparaison de performance ne va pas toujours de soi, quand les modèles n'ont pas le même nombre de paramètres ou ne sont pas du même type. Le modèle le plus complexe sera plus performant sur les données ayant servi à l'estimation. Pour palier ce biais, deux méthodes sont proposées.

La comparaison de performance ne va pas toujours de soi, quand les modèles n'ont pas le même nombre de paramètres ou ne sont pas du même type. Le modèle le plus complexe sera plus performant sur les données ayant servi à l'estimation. Pour palier ce biais, deux méthodes sont proposées.

La comparaison de performance ne va pas toujours de soi, quand les modèles n'ont pas le même nombre de paramètres ou ne sont pas du même type. Le modèle le plus complexe sera plus performant sur les données ayant servi à l'estimation. Pour palier ce biais, deux méthodes sont proposées.

- On partage l'échantillon du départ en deux sous-échantillons : le premier dit d'apprentissage sert à estimer le modèle, le second de validation sert à construire la matrice de confusion. Cette méthode nécessite un grand nombre d'observations.
- On réalise une validation croisée en partageant en 10 échantillons les données : les 9 premiers servant à l'apprentissage et le dixième à la validation, puis on réalise une permutation circulaire sur le rôle des échantillons. Le taux d'erreur est obtenu comme une moyenne sur des 10 taux d'erreurs obtenus.

- On partage l'échantillon du départ en deux sous-échantillons : le premier dit d'apprentissage sert à estimer le modèle, le second de validation sert à construire la matrice de confusion. Cette méthode nécessite un grand nombre d'observations.
- On réalise une validation croisée en partageant en 10 échantillons les données : les 9 premiers servant à l'apprentissage et le dixième à la validation, puis on réalise une permutation circulaire sur le rôle des échantillons. Le taux d'erreur est obtenu comme une moyenne sur des 10 taux d'erreurs obtenus.

- Partition si possible du fichier en apprentissage et validation
- Régression complète
- Sélection de variables
  - Soit par une méthode de sélection pas à pas.
  - Soit en réalisant le test du rapport de vraisemblance pour enlever plusieurs variables a la fois.
  - Sélection automatique avec le critère d'Akaike.
- Comparaison des différents modèles avec les matrices de confusions.
- Choix du meilleur modèle s'il existe.