Quelques Rappels sur la régression logistique

1 Objectif et écriture du modèle

On explique une variable Y_i binaire (évènement/non-évènement) en modélisant la probabilité de survenue de l'évènement. Y_i est distribué selon une loi de Bernoulli :

$$Y_i \sim Bernoulli(p_i) \text{ avec } E(Y_i) = p_i$$

La fonction de lien canonique en régression logistique est la fonction *logit*, d'où l'écriture du modèle :

$$logit(p_i) = \ln(\frac{p_i}{1-pi}) = \beta_0 + \beta_1 x_i^1 + ... + \beta_p x_i^p$$

On estime β par $\widehat{\beta}$ et on en déduit les probabilités prédites $\widehat{p_i}$:

$$\widehat{p_i} = \frac{1}{1 + e^{-\widehat{\beta}_0 - \widehat{\beta}_1 x_i^1 - ... - \widehat{\beta}_p x_i^p}}$$

2 Interprétations des paramètres estimés

En régression logistique, pour interpréter les paramètres estimés, on utilise souvent $e^{\widehat{\beta_j}}$ qui représente l'odds ratio associé à X^j ; on estime ainsi que la probabilité que l'évènement soit présent plutôt qu'absent est multipliée par $e^{\widehat{\beta_j}}$ quand x^j augmente d'une unité.
Un paramètre $\widehat{\beta_j}$ nul équivaut à l'absence d'effet de x^j sur la variable réponse.
Si $\widehat{\beta_j}$ est positif, alors $e^{\widehat{\beta_j}} > 1$, la probabilité de présence de l'évènement augmente avec x^j.
Si $\widehat{\beta_j}$ est négatif, alors $e^{\widehat{\beta_j}} < 1$, la probabilité de présence de l'évènement diminue avec x^j.
$\frac{1}{1 + e^{-\widehat{\beta}_0}}$ représente la probabilité de présence de l'évènement quand toutes les variables explicatives sont nulles.

3 Tests d'hypothèses

Comme dans toute démarche de modélisation, l'objectif est d'obtenir le meilleur modèle en terme d'ajustement aux données et de variables explicatives significatives (selon le principe de parcimonie). Pour cela, on met en place des tests d'hypothèses, permettant de tester l'ajustement global du modèle aux données et l'absence d'effet de variable(s) explicative(s).

Comparer deux modèles emboîtés (un modèle de référence noté M_1 à k paramètres et un modèle réduit à $k - q$ paramètres, noté M_0) revient à tester la nullité des q paramètres (présents dans le modèle complet, mais pas dans le modèle réduit), formulée par l'hypothèse nulle H_0 :

$$H_0 : \forall j = 1, ...q \;\; \beta_j = 0$$

q représente le nombre de contraintes posées sous H_0, correspondant à la différence entre le nombre de paramètres estimés dans les deux modèles. Ce test permet de conclure sur l'effet des q variables associées aux q paramètres. Dans le cadre du modèle linéaire généralisé, on dispose de trois tests statistiques.

On dispose de trois statistiques de tests possibles, comme pour le modèle de Cox :

- le test du rapport des vraisemblances basée sur la statistique $LR = -2(\ln L_0 - \ln L_1)$ où L_0 est la vraisemblance maximisée du modèle sous H_0, et L_1 la vraisemblance maximisée pour le modèle de référence.
- le test du Score
- le test de Wald

Pour ces 3 statistiques, on les compare au quantile $\chi^2_{q,1-\alpha}$. On rejette H_0 si la statistique est supérieure à $\chi^2_{q,1-\alpha}$.

3.1 Test d'ajustement global du modèle

On teste l'hypothèse d'absence d'effet de toutes les variables explicatives :

$$H_0 : \forall j = 1, ...p, \;\; \beta_j = 0 \text{ versus } H_1 : \exists j / \beta_j \neq 0$$

Cela revient à comparer le modèle estimé au modèle blanc (ne comprenant aucune variable explicative). On pose alors p contraintes.

On utilise les trois tests à notre disposition : Rapport des vraisemblances, Score et Wald. On compare les trois statistiques de test calculées au quantile d'ordre 95% d'une loi du χ^2 à p ddl. Quand les 3 p-valeurs associées à ces statistiques de test sont largement inférieures à 5%, on rejette H_0 et on conclut à la présence d'au moins une variable explicative significative dans le modèle.

3.2 Test de significativité d'une variable explicative

Pour tester la nullité d'un seul paramètre du modèle, c'est-à-dire "$H_0 : \beta_j = 0$", on peut mettre en œuvre le test de Wald (correspondant au test de Student). Le test de Wald est basé sur la statistique :

$$W_j = \frac{\widehat{\beta}_j(y)^2}{(\widehat{s.e.}(\widehat{\beta}_j))^2}$$

où $\widehat{s.e.}(\widehat{\beta}_j)$ est l'erreur standard de $\widehat{\beta}_j$ ou bien $(\widehat{s.e.}(\widehat{\beta}_j))^2$ est le j^e élément diagonal de la matrice de variance-covariance de $\widehat{\beta}$. Sous H_0, W_j est distribué selon une loi du χ^2 à 1 ddl.

Si la p-valeur associée à cette statistique de test est inférieure à 5%, on rejette H_0 et on conclut à un effet signifcatif de la variable explicative testée sur la variable réponse.

3.3 Intervalle de confiance d'un paramètre β_j

Pour construire un IC pour β_j, on utilise le test de Wald et la propriété asymptotique sur laquelle il s'appuie, et on obtient :

$$IC_{1-\alpha}(\beta_j) = [\widehat{\beta}_j \pm z_{(1-\alpha/2)} \times \widehat{s.e.}(\widehat{\beta}_j)]$$

4 Prédictions

Le modèle permet d'obtenir les probabilités prédites de présence de l'évènement pour chaque observation i. On peut prédire y_i par $\widehat{y_i}$ en choisissant un seuil (compris entre 0 et 1) tel que :

- Si $\widehat{p_i} > seuil$ alors $\widehat{y_i} = 1$;
- Si $\widehat{p_i} < seuil$ alors $\widehat{y_i} = 0$.

On peut ainsi comparer y_i et $\widehat{y_i}$, et en déduire le taux de mal-classés.

Le D de Somer est un critère mesurant la qualité prédictive du modèle selon l'échelle suivante :

- D nul => les prédictions ne valent rien ;
- D inférieur à 0.4 => les prédictions ne sont pas très bonnes ;
- D compris entre 0.4 et 0.6 => la qualité des prédictions est correcte ;
- D supérieur à 0.6 => le modèle donne d'excellentes prédictions.

5 Critères de qualité

Pour juger de la qualité d'un modèle estimé, on dispose de trois critères principaux.

Vraisemblance et log-vraisemblance
En régression logistique, les vraisemblance et log-vraisemblance sont de la forme :

$$L(y,p) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}$$

$$\ln L(y,p) = \sum_{i=1}^{n} y_i \ln(p_i) + (1-y_i)\ln(1-p_i)$$

Le modèle saturé de régression logistique a une log-vraisemblance nulle. La déviance d'un modèle de régression logistique est donc égale à $-2\ln L$.

Déviance
La déviance mesure l'écart de qualité d'ajustement entre le modèle estimé et le modèle saturé (dont l'ajustement est parfait : $\forall i, \widehat{y_i} = y_i$). Elle est basée sur les log-vraisemblances de ces deux modèles (notées $\ln L_S$ pour le modèle saturé et $\ln L$ pour le modèle estimé) :

$$Deviance = 2(\ln L_S - \ln L)$$

On utilise en général $\dfrac{deviance}{ddl}$ où $ddl = n - k$, avec n le nombre d'observations et k le nombre de paramètres estimés dans le modèle estimé. Plus petits sont ces deux critères, meilleur est l'ajustement du modèle.

Critères AIC et BIC
Pour choisir entre deux modèles, on sélectionnera le modèle ayant les plus faibles valeurs de AIC et BIC :

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

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

6 Mise en œuvre sous SAS

Sous SAS, on utilise la procédure LOGISTIC pour mettre en œuvre une régression logistique à l'aide de la syntaxe suivante, `rep` représentant la variable réponse binaire et `var1`, `var2`, ... les variables explicatives.

```
proc logistic data=... ;
class var2 ;
model rep = var1 var2 ... ;
run ;
```

Dans cette procédure, SAS modélise la probabilité associée à la plus petite valeur de la variable réponse (par exemple, si une variable est codée 0 ou 1, SAS modélise la probabilité que la variable soit égale à 0). Si on veut modéliser la probabilité que la variable soit égale à 1, on utilise alors l'option `descending` à spécifier sur la ligne `proc...` .
La variable `var2` mentionnée dans l'instruction `class` est prise en compte comme variable explicative qualitative.

La procédure `LOGISTIC` vous fournit en sortie :

- des statistiques d'ajustement vus au paragraphe 5 (critères de déviance, AIC et BIC) pour le modèle estimé (avec covariables) et pour le modèle blanc (sans covariables) pour vérifier que les variables explicatives introduites dans le modèle améliorent la qualité d'ajustement du modèle, et donc sont intéressantes.
- Le test global d'ajustement du modèle (vus au paragraphes 3) : 3 tests pour tester la nullité des paramètres associés aux variables explicatives introduites dans le modèle.
- Les estimations des paramètres associés à chaque variable explicative pour tester l'effet de chaque variable par le test de Wald (vu au paragraphe 3.2). Vous sont également donnés les odds-ratios ou rapports de côtes associés à chaque variable (vus au paragraphe 2). Ceux sont ces éléments qu'il faut commenter pour interpréter l'effet significatif d'un facteur ou d'une covariable sur la probabilité modélisée.
- Des critères de qualité associés aux prédictions tels que le D de Somer (vu au paragraphe 5).

Il est également possible de sélectionner automatiquement les variables explicatives selon une démarche descendante (la plus adéquate dans la plupart des modélisations) :

```
proc logistic data=... ;
class var2 ;
model rep = var1 var2 ... / selection=backward ;
run ;
```

7 Mise en œuvre sous R

Sous R, la régression logistique est mise en œuvre par la fonction `glm` où l'on spécifie le modèle (la variable réponse et la (ou les) variable(s) explicative(s)) et le type de la variable réponse, à savoir binomial.

Pour obtenir les résultats de la régression logistique mise en œuvre, il faut utiliser les fonctions `anova` et `summary` :

— La fonction `anova` donne les résultats des tests d'absence d'effet de chaque variable explicative.
— La fonction `summary` donne les estimations des paramètres et les tests associés (sur la nullité du paramètre), ainsi que les critères de qualité de la régression (déviance et AIC).

Dans le cas de variables explicatives quantitatives ou qualitatives à deux classes, les résultats des tests fournis par les deux fonctions sont les mêmes (car tester l'absence d'effet d'une variable explicative quantitative, c'est tester la nullité du paramètre associé). En revanche pour les variables qualitatives à plus de deux classes, les résultats fournis par la fonction `anova` sont indispensables pour juger si l'effet de la variable explicative est significatif ou non.

```
> reg.log = glm(rep ~ var1 + var2 + var3 ... , family="binomial")
> anova(reg.log, test="Chisq")
> summary(reg.log)
```

7.1 Probabilités prédites

On peut calculer les probabilités prédites par le modèle pour chaque individu :

```
> predict(reg.log,type="response")
```

Pour les variables explicatives quantitatives, on peut tracer les probabilités prédites en fonction de la variable explicative :

```
> plot(var1,rep,ylab="probabilité prédite")
> xp=seq(min(var1),max(var1))
> yp=predict(reg.log,data.frame(var1=xp),type="response")
> lines(xp,yp,col="red")
```

On peut aussi représenter les probabilités prédites sous forme de deux histogrammes (l'un pour les décès et l'autre pour les non-décédés) :

```
> prob.pred = predict(reg.log, type="response")
> par(mfrow=c(1,2))
> hist(prob.pred[rep==1], probability=T, col='light blue')
> lines(density(prob.pred[rep==1]),col='red',lwd=3)
> hist(prob.pred[rep==0], probability=T, col='light blue')
> lines(density(prob.pred[rep==0]),col='red',lwd=3)
```

7.2 Sélection des variables explicatives

La fonction `step` permet de sélectionner automatiquement les variables à partir d'un modèle dit "complet" selon une démarche descendante présentée ici (elle permet aussi une sélection ascendante ou mixte, non présentée ici) :

```
> reg.log.backward = step(reg.log, direction = "backward")
```

```
> anova(reg.log.backward, test="Chisq")
> summary(reg.log.backward)
```

L'objet `reg.log.backward` contient les éléments du modèle sélectionné.

7.3 Interprétation des effets

Pour interpréter les effets des covariables, on affiche les exponentielles des coefficients estimés :

```
> exp(cbind(OR = coef(m3.backward), confint(m3.backward)))
```

7.4 La courbe ROC

A partir des probabilités prédites, on peut classer les individus comme "décédé" ou "non décédé" en fixant un seuil t : si la probabilité prédite de décès est supérieur à t, on classe l'individu en "décédé" ; sinon, on le classe en "non-décédé".
On peut ainsi comparer ce classement avec l'observation, et calculer les taux de bien classés ou de mal classés. Si on change le seuil t, le classement est modifié et les taux également.
Pour juger de la qualité d'une régression logistique, on utilise la **courbe ROC**. Elle représente l'évolution de la sensibilité (taux de vrais positifs) en fonction de 1 - spécificité (taux de faux positifs) quand on fait varier le seuil t.
C'est une courbe croissante entre le point (0,0) et le point (1, 1) et en principe au-dessus de la première bissectrice. Plus la courbe est au-dessus la première bissectrice, meilleure est la prédiction. Une prédiction idéale est l'horizontale y=1 sur $]0,1]$ et le point (0,0).
L'aire sous la courbe ROC (AUC, Area Under the Curve) donne un indicateur de la qualité de la prédiction (1 pour une prédiction idéale, 0.5 pour une prédiction random).

Le package `ROCR` permet de tracer la courbe ROC et d'obtenir la valeur de AUC.

```
> library(ROCR)
> pred = prediction(prob.pred, dcd)
> perf = performance(pred, "tpr", "fpr")
> plot(perf)
> AUC=performance(pred, "auc"@y.values[[1]])
> AUC
```

7.5 D'autres critères de qualité du modèle par la fonction `lrm` du package `HH`

La fonction `lrm` fournit des sorties complémentaires à la fonction `glm`, en particulier des indices de qualité de la régression logistique, tels que le D de Somers par exemple.

```
> library(HH)
> reg.log.lrm=lrm(rep   var1+var2+var3...)
> reg.log.lrm
```