Régression logistique ou modèle binomial

Résumé

Rappels sur la régression logistique *ou modèle binomial du modèle linéaire général, Les lois des observations sont discrètes et associées à des dénombrements : binomiale, multinomiale. Définition de la notion de rapport de cotes utile dans l'interprétation du rôle des paramètres ; modèle pour une variable binomiale ou une variable binaire* $(0, 1)$ *de Bernoulli ; estimation, propriétés et difficultés spécifiques à ce modèle ; extension à la modélisation d'une variable ordinale. Choix de modèle en régression logistique et exemples.*

Retour à l'introduction.

Tous les tutoriels sont disponibles sur le dépôt :
`github.com/wikistat`

1 Introduction

Historiquement, la régression logistique ou régression binomiale fut la première méthode utilisée, notamment en épidémiologie et en marketing (*scoring*), pour aborder la modélisation d'une variable binaire binomiale (nombre de succès pour n_i essais) ou de Bernoulli (avec $n_i = 1$) : décès ou survie d'un patient, absence ou présence d'une pathologie, possession on non d'un produit, bon ou mauvais client...

Bien connue dans ces types d'application et largement répandue, la régression logistique conduit à des interprétations pouvant être complexes mais rentrées dans les usages pour quantifier, par exemple, des facteurs de risque liés à une pathologie, une faillite... Cette méthode reste donc celle la plus utilisée car interprétable même si, en terme de qualité prévisionnelle, d'autres approches sont susceptibles, en fonction des données étudiées, de conduire à de meilleures prévisions. Enfin, robuste, cette méthode passe à l'échelle des données massives. Il est donc important de bien maîtriser les différents aspects de la régression logistiques dont l'interprétation des paramètres, la sélection de modèle par sélection de variables ou par régularisation (Lasso).

Cas particulier de modèle linéaire général, la régression logistique reprend la plupart des usages des méthodes de cette famille : estimation par maximisation de la vraisemblance, statistiques de test suivant asymptotiquement des lois du chi-deux, calcul des résidus, observations influentes, critère pénalisé (AIC) d'Akaïke[1] pour la sélection de modèle.

2 Cotes et rapports de cote

Une première section définit quelques notions relatives à l'étude de la liaison entre variables qualitatives. Elles sont couramment utilisées dans l'interprétation des modèles de régression logistique.

Une variable

Soit Y une variable qualitative à J modalités. On désigne la chance, cote ou *odd*[1] de voir se réaliser la j-ème modalité plutôt que la $k$ème par le rapport

$$\Omega_{jk} = \frac{\pi_j}{\pi_k}$$

où π_j est la probabilité d'apparition de la j-ème modalité. Cette quantité est estimée par le rapport n_j/n_k des effectifs observés sur un échantillon. Lorsque la variable est binaire et suit une loi de Bernouilli de paramètre π, l'*odd* ou la cote est le rapport $\pi/(1-\pi)$ qui exprime une chance de gain.

Par exemple, si la probabilité d'un succès est 0.8, celle d'un échec est 0.2. La cote du succès est 0.8/0.2=4 tandis que la cote de l'échec est 0.2/0.8=0.25. On dit encore que la chance de succès est de 4 contre 1 tandis que celle d'échec est de 1 contre 4.

2.1 Table de contingence

On considère maintenant une table de contingence 2×2 croisant deux variables qualitatives binaires X^1 et X^2. les paramètres de la loi conjointe se

1. Il n'existe pas, même en Québécois, de traduction très consensuelle de *odd* dan scet usage.

mettent dans une matrice :

$$\begin{bmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{bmatrix}$$

où $\pi_{ij} = P[\{X^1 = i\} \text{ et } \{X^2 = j\}]$ est la probabilité d'occurrence de chaque combinaison.

- Dans la ligne 1, la cote que la colonne 1 soit prise plutôt que la colonne 2 est :
$$\Omega_1 = \frac{\pi_{11}}{\pi_{12}}.$$
- Dans la ligne 2, la cote que la colonne 1 soit prise plutôt que la colonne 2 est :
$$\Omega_2 = \frac{\pi_{21}}{\pi_{22}}.$$

On appelle *odds ratio* (rapport de cotes) le rapport

$$\Theta = \frac{\Omega_1}{\Omega_2} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}.$$

Ce rapport prend la valeur 1 si les variables sont indépendantes, il est supérieur à 1 si les sujets de la ligne 1 ont plus de chances de prendre la première colonne que les sujets de la ligne 2 et inférieur à 1 sinon.

Exemple : supposons qu'à l'entrée dans une école d'ingénieurs, 7 garçons sur 10 sont reçus tandis que seulement 4 filles sur 10 le sont. La cote des garçons est alors de 0.7/0.3=2.33 tandis que celle des filles est de 0.4/0.6=0.67. Les raport de cote est de 2.33/0.67=3.5. La chance d'être reçu est 3.5 plus grande pour les garçons que pour les filles.

Le rapport de cotes est également défini pour deux lignes (a, b) et deux colonnes (c, d) quelconques d'une table de contingence croisant deux variables à J et K modalités. C'est le rapport :

$$\Theta_{abcd} = \frac{\Omega_a}{\Omega_b} = \frac{\pi_{ac}\pi_{bd}}{\pi_{ad}\pi_{bc}} \quad \text{estimé par le rapport de cotes empirique} \quad \widehat{\Theta}_{abcd} = \frac{n_{ac}n_{bd}}{n_{ad}n_{bc}}.$$

3 Régression logistique

3.1 Type de données

Cette section décrit la modélisation d'une variable qualitative Z à 2 modalités : 1 ou 0, succès ou échec, présence ou absence de maladie, panne d'un équipement, faillite d'une entreprise, bon ou mauvais client.... Les modèles de régression précédents adaptés à l'explication d'une variable quantitative ne s'appliquent plus directement car le régresseur linéaire usuel $\mathbf{X}\boldsymbol{\beta}$ ne prend pas des valeurs simplement binaires. L'objectif est adapté à cette situation en cherchant à expliquer les probabilités

$$\pi = P(Z = 1) \quad \text{ou} \quad 1 - \pi = P(Z = 0),$$

ou plutôt une transformation de celles-ci, par l'observation conjointe des variables explicatives. L'idée est en effet de faire intervenir une fonction réelle monotone g opérant de $[0, 1]$ dans $\mathbb{R}$ et donc de chercher un modèle linéaire de la forme :

$$g(\pi_i) = \mathbf{x}_i'\beta.$$

Il existe de nombreuses fonctions, dont le graphe présente une forme sigmoïdale et qui sont candidates pour remplir ce rôle, trois sont pratiquement disponibles dans les logiciels :

probit : g est alors la fonction inverse de la fonction de répartition d'une loi normale, mais son expression n'est pas explicite.

log-log avec g définie par

$$g(\pi) = \ln[-\ln(1 - \pi)]$$

mais cette fonction est dissymétrique.

logit est définie par

$$g(\pi) = \mathit{logit}(\pi) = \ln\frac{\pi}{1-\pi} \quad \text{avec} \quad g^{-1}(x) = \frac{e^x}{1+e^x}.$$

Plusieurs raisons, tant théoriques que pratiques, font préférer cette dernière solution. Le rapport $\pi/(1 - \pi)$, qui exprime une cote, est l'*odd*. La *régression logistique* s'interprète donc comme la recherche d'une modélisation linéaire du *log-odd* tandis que les coefficients de certains modèles expriment des rapports de cotes (*odds ratio*) c'est-à-dire l'influence d'un facteur qualitatif sur le risque (ou la chance) d'un échec (d'un succès) de Z.

Cette section se limite à la description de l'usage élémentaire de la régression logistique. Des compléments concernant l'intervention de variables explicatives avec effet aléatoire, l'utilisation de mesures répétées donc dépendantes, sont à rechercher dans la bibliographie.

3.2 Modèle binomial

Pour $i = 1, \ldots, I$, différentes valeurs *fixées* $x_i^1, \ldots, x_i^q$ des variables explicatives $X^1, \ldots, X^q$ sont observées. Ces dernières pouvant être des variables quantitatives ou qualitatives.

Pour chaque groupe, c'est-à-dire pour chacune des combinaisons de valeurs ou facteurs, sont réalisées n_i observations ($n = \sum_{i=1}^{I} n_i$) de la variable Z qui se mettent sous la forme $y_1/n_1, \ldots, y_I/n_I$ où y_i désigne le nombre de "succès" observés lors des n_i essais. Toutes les observations sont supposées indépendantes et, à l'intérieur d'un même groupe, la probabilité π_i de succès est supposée constante. Alors, la variable Y_i sachant n_i et d'espérance $E(Y_i) = n_i\pi_i$ suit une loi *binomiale* $\mathcal{B}(n_i, \pi_i)$ dont la fonction de densité s'écrit :

$$P(Y = y_i) = \binom{n_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{(n_i - y_i)}.$$

Le modèle suppose que le vecteur des fonctions *logit* des probabilités π_i appartient au sous-espace vect$\{X^1, \ldots, X^q\}$ engendré par les variables explicatives :

$$\mathit{logit}(\pi_i) = \mathbf{x}_i'\boldsymbol{\beta} \quad i = 1, \ldots, I$$

ce qui s'écrit encore

$$\pi_i = \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i'\boldsymbol{\beta}}} \quad i = 1, \ldots, I.$$

Le vecteur des paramètres est estimé par maximisation de la log-vraisemblance. Il n'y a pas de solution analytique, celle-ci est obtenue par des méthodes numériques itératives (par exemple Newton Raphson) dont certaines (Fisher) reviennent à itérer des estimations de modèles de régression par moindres carrés généralisés avec des poids et des métriques adaptés à chaque itération.

L'optimisation fournit une estimation $\mathbf{b}$ de $\boldsymbol{\beta}$, il est alors facile d'en déduire les estimations ou prévisions des probabilités π_i :

$$\widehat{\pi}_i = \frac{e^{\mathbf{x}_i'\mathbf{b}}}{1 + e^{\mathbf{x}_i'\mathbf{b}}}$$

et ainsi celles des effectifs

$$\widehat{y}_i = n_i\widehat{\pi}_i.$$

Remarques

1. La matrice $\mathbf{X}$ issue de la planification expérimentale est construite avec les mêmes règles que celles utilisées dans le cadre de l'analyse de covariance mixant variables explicatives quantitatives et qualitatives. Ainsi, les logiciels gèrent avec plus ou moins de clarté le choix des variables indicatrices et donc des paramètres estimables ou contrastes associés.
2. **Attention**, La situation décrite précédemment correspond à l'observation de données *groupées*. Dans de nombreuses situations concrètes et souvent dès qu'il y a des variables explicatives quantitatives, les observations $\mathbf{x}_i$ sont toutes distinctes. Ceci revient donc à fixer $n_i = 1; i = 1, \ldots, I$ dans les expressions précédentes et la loi de Bernouilli remplace la loi binomiale. Certaines méthodes ne sont alors plus applicables et les comportements asymptotiques des distributions des statistiques de test ne sont plus valides car le nombre de paramètres tend vers l'infini avec n.
3. Dans le cas d'une variable explicative X dichotomique, un logiciel comme SAS fournit, en plus de l'estimation d'un paramètre b, celle des rapports de cotes (*odds ratios*) ; b est alors le *log odds ratio* ou encore, e^b est le rapport de cotes. Ceci s'interprète en disant que Y a e^b fois plus de chance de succès (ou de maladie comme par un exemple un cancer du poumon) quand $X = 1$ (par exemple pour un fumeur).
4. **Attention**, les différences de paramétrisation $(-1, 1)$ ou $(0, 1)$ des indicatrices explique les différences observées dans l'estimation des paramètre d'un logiciel à l'autre mais les modèles sont identiques. Mêmes exprimés dans des bases différentes, les espaces engendrés par les vecteurs des indicatrices sélectionnées sont les mêmes.

3.3 Régressions logistiques polytomiques et ordinales

3.3.1 Généralisation

La régression logistique adaptée à la modélisation d'une variable dichotomique se généralise au cas d'une variable Y à plusieurs modalités. La généralisation la plus rudimentaire adoptée dans la librairie `Scikit-learn` de Python consiste à considérer autant de modèles dichotomiques que de modalités : une contre les autres.

Si ces modalités sont ordonnés, la variable est dite qualitative ordinale et la régression polytomique. Ce type de modélisation est très souvent utilisé en épidémiologie et permettent d'évaluer ou comparer des risques par exemples sanitaires. Des estimations d'*odds ratio* ou rapports de cotes sont ainsi utilisés pour évaluer et interpréter les facteurs de risques associés à différents types ou seuils de gravité d'une maladie ou, en marketing, cela s'applique à l'explication, par exemple, d'un niveau de satisfaction d'un client. Il s'agit de comparer entre elles des estimations de fonctions logit.

Dans une situation de *data mining* ou fouille de données, ce type d'approche se trouve lourdement pénalisé lorsque, à l'intérieur d'un même modèle polytomique ou ordinal, plusieurs types de modèles sont en concurrence pour chaque fonction logit associée à différentes modalités. Différents choix de variables, différents niveaux d'interaction rendent trop complexe et inefficace cette approche. Elle est à privilégier uniquement dans le cas d'un nombre restreint de variables explicatives.

Logits cumulatifs

À titre illustratif, explicitons le cas simple d'une variable Y à k modalités ordonnées expliquée par une seule variable dichotomique X. Notons $\pi_j(X) = P(Y = j|X)$ avec $\sum_{j=1}^{k} \pi_j(X) = 1$. Pour une variable Y à k modalités, il faut, en toute rigueur, estimer $k-1$ prédicteurs linéaires :

$$g_j(X) = \alpha_j + \beta_j X \quad \text{pour} j = 1, \ldots, k-1$$

et, dans le cas d'une variable ordinale, la fonction lien logit utilisée doit tenir compte de cette situation particulière.

Dans la littérature, trois types de fonction sont considérées dépendant de l'échelle des rapports de cotes adoptée :

- échelle basée sur la comparaison des catégories adjacentes deux à deux,
- sur la comparaison des catégories adjacentes supérieures cumulées,
- et enfin sur la comparaison des catégories adjacentes cumulées.

Pour $k = 2$, les trois situations sont identiques. C'est le dernier cas qui est le plus souvent adopté ; il conduit à définir les fonctions des logits cumulatifs de la forme :

$$\log \frac{\pi_{j+1} + \cdots + \pi_k}{\pi_1 + \cdots + \pi_j} \quad \text{pour} j = 1, \ldots, k-1.$$

Pour un seuil donné sur Y, les catégories inférieures à ce seuil, cumulées, sont comparées aux catégories supérieures cumulées. Les fonctions logit définies sur cette échelle dépendent chacune de tous les effectifs, ce qui peut conduire à une plus grande stabilité des mesures qui en découlent.

Proportionnalité des rapports de cote

Si les variables indépendantes sont nombreuses dans le modèle ou si la variable réponse Y comporte un nombre élevé de niveaux, la description des fonctions logit devient fastidieuse. La pratique consiste plutôt à déterminer un coefficient global b (mesure d'effet) qui soit la somme pondérée des coefficients b_j. Ceci revient à faire l'hypothèse que les coefficients sont homogènes (idéalement tous égaux), c'est-à-dire à supposer que les rapports de cotes sont proportionnels. C'est ce que calcule implicitement la procédure LOGISTIC de SAS appliquée à une variable réponse Y ordinale en estimant un seul paramètre b mais $k-1$ termes constants correspondant à des translations de la fonctions logit.

La procédure LOGISTIC fournit le résultat du test du score sur l'hypothèse H_0 de l'homogénéité des coefficients β_j.

Le coefficient b mesure donc l'association du facteur X avec la gravité de la maladie et peut s'interpréter comme suit : pour tout seuil de gravité choisi sur Y, la cote des risques d'avoir une gravité supérieure à ce seuil est e^b fois plus grande chez les exposés ($X = 1$) que chez les non exposés ($X = 0$).

3.4 Choix de modèle

Les algorithmes sont identiques à ceux décrits pour l'analyse de covariance mais, *attention*, du fait de l'utilisation d'une transformation non linéaire (logit), même si des facteurs sont orthogonaux, aucune propriété d'orthogonalité ne peut être prise en compte pour l'étude des hypothèses. Ceci impose l'élimination des termes un par un et la ré-estimation du modèle. D'autre part, un terme principal ne peut être supprimé que s'il n'intervient plus dans des termes d'interaction.

Les critères à optimiser sont les mêmes avec pour choix par défaut l'AIC en R. Dans `Scikit-learn` de Python, les mêmes versions de sélection par pénalisation : *ridge*, Lasso, elastic-net, sont proposées.

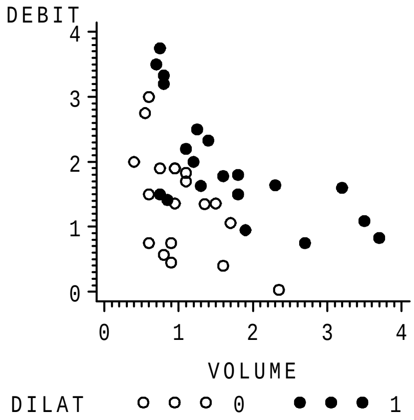

FIGURE 1 – Dilatation : Nuage des modalités de Y dans les coordonnées des variables explicatives.

4 Exemples

4.1 Exemple élémentaire avec SAS

Les données

L'influence du débit et du volume d'air inspiré impacte-t-elle l'occurrence (codée 1) de la dilatation des vaisseaux sanguins superficiels des membres inférieurs ? Un graphique élémentaire représentant les modalités de Y dans les coordonnées de $X^1 \times X^2$ est toujours instructif. Il montre une séparation raisonnable et de bon augure des deux nuages de points. Dans le cas de nombreuses variables explicatives quantitatives, une analyse en composantes principales s'impose. Les formes des nuages représentés, ainsi que l'allure des distributions (étudiées préalablement), incitent dans ce cas à considérer par la suite les logarithmes des variables. Une variable `un` ne contenant que des 1s dénombrant le nombre d'essais est nécessaire dans la syntaxe de `genmod`. Les données sont en effet non groupées.

```
proc logistic data=sasuser.debvol;
model dilat=l_debit l_volume;
run;
```

```
proc genmod data=sasuser.debvol;
model dilat/un=l_debit l_volume/d=bin;
run;
```

```
The LOGISTIC Procedure
                                          Intercept
                         Intercept           and
 Criterion                 Only          Covariates    Chi-Square for Covariates
 AIC                      56.040           35.216              .
 SC                       57.703           40.206              .
 -2 LOG L                 54.040           29.216(1)     24.824 with 2 DF (p=0.0001)
 Score                      .                 .          16.635 with 2 DF (p=0.0002)

            Parameter(2) Standard    Wald(3)    Pr >    Standardized  Odds
Variable DF  Estimate     Error    Chi-Square Chi-Square  Estimate    Ratio
INTERCPT 1     2.8782     1.3214     4.7443     0.0294        .         .
L_DEBIT  1    -4.5649     1.8384     6.1653     0.0130    -2.085068   0.010
L_VOLUME 1    -5.1796     1.8653     7.7105     0.0055    -1.535372   0.006
```

Cette procédure fournit des critères de choix de modèle dont la déviance (1), le vecteur **b** des paramètres (2) et les statistiques des tests (3) comparant le modèle excluant un terme par rapport au modèle complet tel qu'il est décrit dans la commande.

```
      Criteria For Assessing Goodness Of Fit
Criterion                DF          Value       Value/DF
Deviance                 36        29.2156         0.8115 (1)
Scaled Deviance          36        29.2156         0.8115 (2)
Pearson Chi-Square       36        34.2516         0.9514 (3)
Scaled Pearson X2        36        34.2516         0.9514
Log Likelihood            .       -14.6078          .

          Analysis Of Parameter Estimates
Parameter    DF    Estimate (4)   Std Err   ChiSquare (5)  Pr>Chi
INTERCEPT     1     -2.8782       1.3214      4.7443       0.0294
L_DEBIT       1      4.5649       1.8384      6.1653       0.0130
L_VOLUME      1      5.1796       1.8653      7.7105       0.0055
SCALE (6)     0      1.0000       0.0000        .            .
```

(1)	Déviance du modèle par rapport au modèle saturé.
(2)	Déviance pondérée si le paramètre d'échelle est différent de 1 en cas de sur-dispersion.
(3)	Statistique de Pearson, voisine de la déviance, comparant le modèle au modèle saturé .
(4)	Paramètres du modèle.
(5)	Statistique des tests comparant le modèle excluant un terme par rapport au modèle complet.
(6)	Estimation du paramètre d'échelle si la quasi-vraisemblance est utilisée.

4.2 Régression logistique ordinale

Les résultats d'une étude préalable à la législation sur le port de la ceinture de sécurité a été conduite dans la province de l'Alberta à Edmonton au Canada (Jobson, 1991). Un échantillon de 86 769 rapports d'accidents de voitures ont été compulsés afin d'extraire une table croisant :

1. Etat du conducteur : Normal ou Alcoolisé

2. Sexe du conducteur
3. Port de la ceinture : Oui Non
4. Gravité des blessures : 0 : rien à 3 : fatales

Les modalités de la variable à expliquer concernant la gravité de l'accident sont ordonnées. Mais dans cet exemple, l'hypothèse H_0 de proportionnalité des rapports de cotes est rejetée. Le problème est alors simplifié en regroupant les conséquences de l'accident en seulement 2 modalités avec ou sans séquelles.

```
                                  Standard          Wald
Parameter          DF    Estimate      Error    Chi-Square    Pr > ChiSq
Intercept Gr0       1      1.8699     0.0236     6264.9373        <.0001
Intercept Gr1       1      2.8080     0.0269    10914.3437        <.0001
Intercept Gr2       1      5.1222     0.0576     7917.0908        <.0001
sexe       Sfem     1     -0.3118     0.0121      664.3353        <.0001
alcool     A_bu     1     -0.5017     0.0190      697.0173        <.0001
ceinture   Cnon     1     -0.1110     0.0174       40.6681        <.0001

Test de score pour l'hypothèse des cotes proportionnelles
Khi-2   DDL   Pr > Khi-2
33.3161  6     <.0001

Modèle élémentaire GrO vs. GrN

Estimations des rapports de cotes
Effet              Valeur estimée  IC de Wald à 95 %
sexe Sfem vs Shom      1.873          1.786 1.964
alcool A_bu vs Ajeu    2.707          2.512 2.918
ceinture Cnon vs Coui  1.244          1.162 1.332
```

4.3 Cancer du sein

Les données (`Wisconsin BreastCancer Database`) sont disponibles dans la librairie `mlbench` du logiciel R. Elles servent très souvent de base de référence à des comparaisons de techniques d'apprentissage. Les variables considérées sont :

Cl.thickness Clump Thickness

Cell.size Uniformity of Cell Size

Cell.shape Uniformity of Cell Shape

Marg.adhesion Marginal Adhesion

Epith.c.size Single Epithelial Cell Size

Bare.nuclei Bare Nuclei

Bl.cromatin Bland Chromatin

Normal.nucleoli Normal Nucleoli

Mitoses Mitoses

Class "benign" et "malignant".

La dernière variable est celle à prédire, les variables explicatives sont ordinales ou nominales à 10 classes. Il reste 683 observations après la suppression de 16 présentant des valeurs manquantes.

Ce jeu de données est assez particulier car plutôt facile à ajuster. Une estimation utilisant toutes les variables conduit à des messages critiques indiquant un défaut de convergence et des probabilités exactement ajustées. En fait le modèle s'ajuste exactement aux données en utilisant toutes les variables aussi l'erreur de prévision nécessite une estimation plus soignée. Une séparation entre un échantillon d'apprentissage et un échantillon test ou une validation croisée permet une telle estimation.

Un modèle parcimonieux et obtenu par une démarche descendante minimisant le critère AIC. Ce modèle conduit à des erreurs de prévision plus faibles sur un échantillon test indépendant qu'un modèle ajustant exactement les données. La qualité de l'ajustement du modèle se résume sous la forme d'une matrice de *confusion* évaluant les taux de bien et mal classés sur l'échantillon d'apprentissage tandis que l'erreur de prévision est estimée à partir de l'échantillon test.

```
# erreur d'ajustement
        benign malignant
  FALSE    345         6
  TRUE      13       182

# erreur de prévision
         benign malignant
  FALSE      84         5
  TRUE        2        46
```

Le taux d'erreur apparent estimé sur l'échantillon d'apprentissage est de 3,5% (0% avec le modèle complet) tandis que le taux d'erreur estimé sans biais sur l'échantillon test est de 5,1% (5,8 avec le modèle complet).

4.4 Pic d'ozone

Plutôt que de prévoir la concentration de l'ozone puis un dépassement éventuel d'un seuil, il pourrait être plus efficace de prévoir directement ce dépassement en modélisant la variable binaire associée. Attention toutefois, ces dépassements étant relativement peu nombreux (17%), il serait nécessaire d'en

accentuer l'importance par l'introduction d'une fonction coût ou une pondération spécifique. Ceci est un problème général losqu'il s'agit de prévoir des phénomènes rares : un modèle trivial ne les prévoyant jamais ne commettrait finalement qu'une erreur relative faible. Ceci revient à demander au spécialiste de quantifier le risque de prévoir un dépassement du seuil à tort par rapport à celui de ne pas prévoir ce dépassement à tort. Le premier a des conséquences économiques et sur le confort des usagers par des limitations de trafic tandis que le 2ème a des conséquences sur l'environnement et la santé de certaines populations. Ce n'est plus un problème statistique mais politique.

La recherche descendante d'un meilleur modèle au sens du critère d'Akaïke conduit au résultat ci-dessous.

```
         Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                       831     744.34
O3_pr     1   132.89        830     611.46 9.576e-31
vmodule   1     2.42        829     609.04      0.12
s_rmh2o   1    33.71        828     575.33 6.386e-09
station   4    16.59        824     558.74 2.324e-03
TEMPE     1   129.39        823     429.35 5.580e-30
```

Un modèle de régression logistique faisant intervenir les interactions d'ordre 2 et optimisé par algorithme descendant aboutit à une erreur de 10,6% tandis que le modèle quantitatif de régression quadratique du chapitre précédent conduit à une erreur de 10,1% avec le même protocole et les mêmes échantillons d'apprentissage et de test.

```
Matrices de confusion de l'échantillon test pour différents modèles :
          0   1                 0   1               0   1               0   1
  FALSE 163  19          FALSE 162  18       FALSE 163  17       FALSE 160  13
  TRUE    5  21          TRUE    6  22       TRUE    5  23       TRUE    8  27
logistique sans vmodule   avec vmodule      avec interactions     quantitatif
```

Les matrices de confusion ne sont pas symétriques et affectées du même biais : tous ces modèles oublient systématiquement plus de dépassements de seuils qu'ils n'en prévoient à tort. Une analyse plus poussée de l'estimation de l'erreur de prévision est évidemment nécessaire. À ce niveau de l'étude, ce qui est le plus utile au météorologue, c'est l'analyse des coefficients les plus significativement présents dans la régression quadratique, c'est-à-dire avec les interactions. Ils fournissent des indications précieuses sur les faiblesses ou insuffisances du modèle physique.

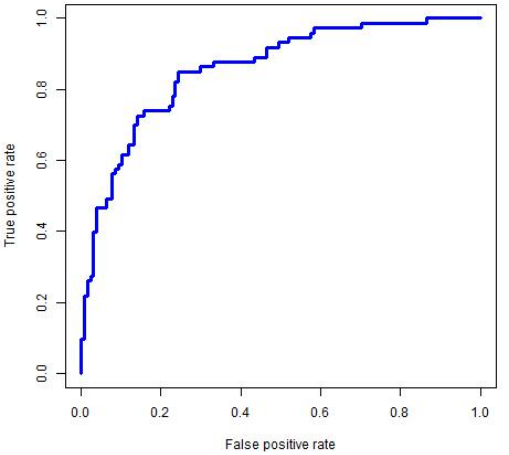

FIGURE 2 – Données bancaires : estimation sur l'échantillon test de la courbe ROC associée à la régression logistique.

4.5 Données bancaires

Il s'agit de modéliser une variable binaire représentant la possession ou non de la carte visa premier en fonction du comportement bancaire d'un client. Cet exemple est typique de la construction d'un score d'appétence en marketing quantitatif. Comme dans l'exemple précédent, la possession de ce type de produit est rare aussi, un échantillon spécifique, non représentatif, a été construit en sur-représentant cette possession.

Plusieurs stratégies peuvent être mises en œuvre sur ces données selon les transformations et codages réalisés sur les variables qualitatives. Elles sont explorées en R et en Python dans deux calepins ; avec minimisation de l'AIC en R, pénalisation Lasso en Python.

Dans ce type d'application, il est très classique d'estimer la courbe ROC sur l'échantillon test afin de calibrer le seuil en fontion des objectifs du service marketing plutôt que de le laisser par défaut à $0,5$.

Références

[1] H. Akaïke, *A new look at the statistical model identification*, IEEE Transactions on Automatic Control **19** (1974).