1 Introduction

Régression linéaire : $Y = X\beta + \epsilon = f(X,\beta) + \epsilon$
avec f fonction linéaire.

Régression non linéaire : $Y = f(X,\beta) + \epsilon$
avec f fonction non-linéaire.

Régression logistique : $Y = \Lambda(X.\beta) + \epsilon$
avec Λ fonction logistique.

La régression logistique généralise la régression linéaire et est un cas particulier de régression non-linéaire (on parle parfois de modèle linéaire généralisé (GLM)).

Dans une régression, on cherche à expliquer une variable aléatoire Y à partir de variables aléatoires (explicatives) X_i, qualitatives ou quantitatives.

Il peut s'agir de régression simple (une seule variable explicative quantitative), multiple (plusieurs variables explicatives quantitatives), d'analyse de la variance (une ou plusieurs variables explicatives, quantitatives) ou d'analyse de la covariance (variables explicatives quantitatives ou qualitatives).

Lorsque Y est une variable qualitative, le modèle linéaire n'est pas adapté. Par exemple, pour une variable binaire (on supposera qu'elle peut prendre comme valeurs 0 et 1) il est difficile d'approcher par une droite deux nuages de points ne prenant comme ordonnées que 0 ou 1. Pour expliquer une variable qualitative, on peut alors effectuer une régression logistique, une analyse discriminante, ou bien encore des techniques d'apprentissage supervisé type k-plus proches voisins.

2 Rappels sur les modèles de régression linéaire

2.1 Hypothèses du modèle

Dans le cas d'une régression simple, on doit supposer que le lien entre Y et X est donné par une relation linéaire, perturbée par un bruit aléatoire :

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{1}$$

ϵ est un bruit aléatoire, β_0 et β_1 les coefficients (inconnus) de la régression. Pour les déterminer, on les estime à partir d'un échantillon de n mesures observées (x_i, y_i), déterministes. Les n équations obtenues :

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \; ; i = 1,..,n \tag{2}$$

permettent le calcul de β_0 et β_1 qui sont solutions de l'équation des moindres carrés

$$(\widehat{\beta}_0, \widehat{\beta}_1) = \min_{\beta_0,\beta_1} \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right)^2 \tag{3}$$

La droite de régression est alors donnée par

$$\widehat{f}(x) = \widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x \tag{4}$$

et l'erreur est donnée par les résidus

$$\widehat{\epsilon}_i = y_i - \widehat{y}_i \tag{5}$$

Parfois, la régression s'écrit avec un Y_i majuscule ($Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$), pour indiquer que Y_i est une variable aléatoire dépendant de la variable aléatoire ϵ_i. Les x_i minuscules indiquent que l'on travaille sachant $[X_i = x_i]$, mais que ϵ_i est toujours considéré en tant que variable aléatoire. Dans la suite, nous utiliserons toujours y_i minuscule et c'est en fonction du contexte qu'il faudra savoir si y_i est une observation (lorsque ϵ_i est une observation déterministe du bruit) ou bien une variable aléatoire (lorsque ϵ_i est une variable aléatoire).

Dans une régression linéaire multiple, la relation liant Y aux variables $X_1, ..., X_p$ est de la forme

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \epsilon \tag{6}$$

Considérons un échantillon $(x_i, y_i)_{i=1,...,n}$ de données issues de (X, Y). On obtient ainsi n équations liant les y_i observés aux x_i et aux β_i.

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + ... + \beta_p x_{1p} + \epsilon_1 \\ \vdots \\ y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \epsilon_i \\ \vdots \\ y_n = \beta_0 + \beta_1 x_{n1} + ... + \beta_p x_{np} + \epsilon_n \end{cases} \tag{7}$$

x_{ij} représente la $i$ème observation du vecteur aléatoire (en majuscule) X_j dans l'échantillon. Chaque vecteur ligne (en minuscule) $x_i = (x_{i1}, ..., x_{ip})$ représente une observation des p vecteurs X_j, pour $j = 1, .., p$.

Il est plus pratique d'utiliser une notation matricielle : en notant $\mathbb{X} = \left(x_{ij}\right)$ on a :

$$\mathbb{X} = \begin{pmatrix} 1 & x_{11} & ... & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & ... & x_{np} \end{pmatrix} \tag{8}$$

Dans la matrice $\mathbb{X} \in \mathbb{R}^{n\times(p+1)}$, la ligne i $(x_{i1}, ...x_{ip})$ représente un individu, c'est à dire la $i$ème observation de toutes ses variables et la colonne j (sauf la première colonne formée de 1 uniquement) représente les différentes observations relatives à une variable X_j. $y \in \mathbb{R}^n$, $\beta \in \mathbb{R}^{p+1}$ et $\epsilon = (\epsilon_1, ..., \epsilon_n) \in \mathbb{R}^n$ sont maintenant des vecteurs et les coordonnées de ϵ sont des copies i.i.d. du bruit défini dans (6). L'équation de régression devient :

$$y = \mathbb{X}\beta + \epsilon \tag{9}$$

dont la solution au sens des moindres carrés est donnée par

$$\widehat{\beta} = \underset{\beta\in\mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \tag{10}$$

$$= \underset{\beta\in\mathbb{R}^{p+1}}{\arg\min} \left((y - \mathbb{X}\beta)^T \times (y - \mathbb{X}\beta) \right) \tag{11}$$

$$= \underset{\beta\in\mathbb{R}^{p+1}}{\arg\min} ||y - \mathbb{X}\beta||_2^2 \tag{12}$$

la solution est le vecteur de coefficients le plus proche - au sens des moindres carrés - des valeurs observées. L'idée de base est un peu d'inverser la relation $y = \mathbb{X}\beta$ et d'écrire $\beta = \mathbb{X}^{-1} y$. Oui mais voilà, la matrice n'est pas inversible (elle n'est pas même pas carrée). Il existe une notion de pseudo-inverse au sens de Moore Penrose qui s'écrit $\mathbb{X}^\dagger$ et qui coïncide exactement avec la solution au sens des moindres carrés (si cette matrice est de rang plein),

$$\widehat{\beta} = \mathbb{X}^\dagger y = \left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T y \tag{13}$$

Si $n \geq p$ et si $\mathbb{X}$ est de rang plein, alors $\mathbb{X}^T\mathbb{X}$ est inversible et l'équation précédente est licite.

Certaines hypothèses du modèle linéaire concerne le bruit ϵ (hypothèses de Gauss-Markov) :

$$\epsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_n) \tag{14}$$

Autrement dit, ϵ est un vecteur aléatoire gaussien, centré, dont les coordonnées sont i.i.d. de variance commune σ^2 (homoscédasticité et non corrélation des résidus). On suppose également que le bruit n'est pas corrélé aux variables explicatives $\mathbb{E}[\epsilon|X] = 0$ (exogénéïté des variables explicatives: le bruit ne doit pas être corrélé aux variables explicatives). On a alors (c'est le théorème de Gauss Markov) :

$$y \sim \mathcal{N}(\mathbb{X}\beta, \sigma^2 \mathrm{I}_n) \tag{15}$$

On peut prouver que sous les hypothèses précédentes $\widehat{\beta}$ est un estimateur sans biais de β de variance minimale égale à $\mathbb{V}(\widehat{\beta}) = \sigma^2(\mathbb{X}^T\mathbb{X})^{-1}$ qui coïncide avec l'estimateur du maximum de vraisemblance du modèle et qu'il est VUMSB.

La matrice $\Pi = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T = \mathbb{X}\mathbb{X}^\dagger$ est la matrice du projecteur orthogonal sur l'espace vectoriel engendré par les colonnes de $\mathbb{X}$ et le vecteur des valeurs ajustées est simplement le projeté orthogonal de y sur l'espace engendré par les colonnes de $\mathbb{X}$.

$$\widehat{y} = \Pi y \tag{16}$$

Le vecteur des résidus est le projeté de y sur l'orthogonal de cet espace, soit:

$$\widehat{\epsilon} = (I - \Pi) y = y - \widehat{y} \tag{17}$$

La somme des carrés résiduelle est

$$SCR = ||\widehat{\epsilon}||^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \tag{18}$$

La somme des carrés totale (recentrée car la colonne de 1 est présente dans la matrice $\mathbb{X}$) est

$$SCT = ||y - \overline{y}||^2 \tag{19}$$

et la somme des carrés estimée (c'est la variance expliquée par le modèle) est

$$SCE = ||\widehat{y} - \overline{y}||^2 \tag{20}$$

Le théorème de Pythagore donne l'équation d'analyse de la variance:

$$SCT = SCE + SCR \tag{21}$$

et le coefficient de détermination R^2 est défini par

$$R^2 = \frac{SCE}{SCT} \tag{22}$$

2.2 Validation du modèle et qualité de la régression

Cette seconde partie est au moins aussi importante que la première et elle est pourtant souvent négligée. Après la modélisation et l'estimation des paramètres, il faut en effet valider le modèle en évaluant la qualité de la régression :

- Vérifier la pertinence du caractère linéaire.
- Faire le choix d'inclure ou non une variable expliquée.
- Analyser les résidus.
- Regarder les valeurs aberrantes.
- Analyser la normalité.
- Vérifier l'homoscédasticité.
- Regarder l'impact d'une variable sur les résidus partiels.
- Etc.

Si l'hypothèse linéaire n'est pas vérifiée ou bien si la matrice $\mathbb{X}$ n'est pas de rang plein, on peut effectuer une régression avec un terme de régularisation (régression «Ridge», LASSO ou autre).

Comme le TP est dédié à la régression logistique, nous ne développons pas plus ce paragraphe. Voir la référence en fin de sujet.

2.3 La régression linéaire avec R

Tout ceci étant posé, nous pouvons maintenant expliquer ce que R calcule avec les fonctions `lm` et `anova`.

`lm(formule,donnees,options)` effectue une regression linéaire selon la formule donnée par `formule` à partir d'unt tableau donné par le paramètre `donnees`. Le résultat est un objet de la classe `lm` qui contient, entre autres, les coefficients de la régression, les résidus, les valeurs ajustées, le rang de la matrice des données, etc.

`summary(lm)` et `anova` permettent d'afficher et de résumer les résultats de `lm`. L'ajustement est effectué de la même façon avec les deux fonctions, mais les tests et les rapports proposés sont différents :

Dans `lm`, le test de Student effectué pour tester la significativité de chaque régresseur mesure l'effet marginal de la variable, étant donné toutes les autres variables du modèle. Dans `anova`, qui effectue une analyse de la variance, les tests de Fisher sont effectués de façon séquentielle, en testant d'abord le premier régresseur par rapport à l'intercept, puis le second par rapport au modèle formé par l'intercept et le premier, etc. Ce test est sensible à l'ordre: si l'on permute la position d'une variable, on change le résultat et on peut même changer la significativité d'une variable.

À chaque fois que la fonction ajoute une nouvelle variable, la variance résiduelle (telle que définie précédemment) diminue et la part de variance expliquée par cette variable est donc la différence entre l'ancienne variance (du modèle précédent, qui ne comprenait pas cette variable) et du modèle courant (dans lequel elle vient d'être ajoutée).

Tout ceci explique que les niveaux de significativité et de façon générale les différentes p-valeurs ne soient pas les mêmes entre `anova` et `summary(lm)`.

La fonction `regr.eval`

Cette fonction évalue la qualité d'une régression en calculant quelques statistiques d'erreurs (erreur quadratique moyenne, absolue, etc.). La syntaxe est la suivante:

`regr.eval(VecVrai,VecPred,VecTrain)`

`VecVrai` est le vecteur contenant les vraies valeurs que le modèle est supposé prédire.

`VecPred` est le vecteur effectivement prédit par le modèle.

`VecTrain` est le vecteur contenant les vraies valeurs de la variable cible sur l'ensemble des données utilisées pour entrainer le modèle.

ce troisième paramètre ne sert que dans le cas où l'on souhaite afficher les erreurs de type `nmse` et `nmae` dont la formule utilise la moyenne du vecteur `VecTrain`.

2.4 La régression linéaire avec Python

La librairie `pandas` est indispensable pour lire facilement les fichiers .csv.

La fonction de régression par défaut est `LinearRegression` de la bibliothèque `Scikit Learn`, à appeler comme suit :

```
from sklearn.linear_model import
LinearRegression
from sklearn.metrics import
mean_squared_error, r2_score
```

Si les données sont stockées dans x, y, l'implémentation s'effectue comme suit:

```
modele = LinearRegression()
modele.fit(x, y)
ypredict = modele.predict(x)
rmse = mean_squared_error(y, ypredict)
r2 = r2_score(y, ypredict)
```

3 Le modèle de régression logistique

3.1 Hypothèses du modèle

Supposons pour commencer que Y soit une variable qualitative binaire pouvant prendre deux valeurs 0 et 1. Le vecteur $X = (1, X_1, .., X_p)$ de variables explicatives a des coordonnées qualitatives ou quantitatives. Nous allons noter

$$p(x) = \mathbb{P}[Y = 1|X = x] = \mathbb{E}[Y/X = x] \tag{23}$$

la probabilité conditionnelle que Y prenne la valeur 1 sachant l'observation x (comme Y est binaire, c'est aussi l'espérance conditionnelle sachant $X = x$). Le rapport de vraisemblance logarithmique (LLR) de Y sachant X est

$$L(x) = \ln\left(\frac{\mathbb{P}[Y = 1|X = x]}{\mathbb{P}[Y = 0|X = x]}\right) = \ln\left(\frac{p(x)}{1 - p(x)}\right) \tag{24}$$

Les économètres appellent ce rapport de vraisemblance logarithmique un « logit ». La fraction à l'intérieur du logarithme s'appelle un «odds ratio» (ou rapport de cotes, de chances ou encore risque relatif). L'idée de la régression logistique est de considérer que le modèle est linéaire non pas par rapport à Y, mais par rapport à son logit :

$$L(x) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p = x.\beta \tag{25}$$

de sorte que

$$p(x) = \frac{\exp(x.\beta)}{1 + \exp(x.\beta)} = \frac{1}{1 + \exp(-x.\beta)} = \Lambda(x.\beta) \tag{26}$$

avec

$$\Lambda(x) = \frac{1}{1 + e^{-x}} \tag{27}$$

La fonction $\Lambda(x)$ est la fonction logistique qui est une bijection continue de $\mathbb{R}$ dans $]0, 1[$. On a :

$$\Lambda = L^{-1} \tag{28}$$

On ne régresse pas directement Y, mais la probabilité qu'elle prenne une valeur donnée. Dans le modèle linéaire généralisé, L s'appelle la fonction de lien («link function») qui exprime le prédicteur linéaire en fonction de la moyenne. Le modèle de régression logistique peut donc s'écrire

$$Y = \Lambda(\mathbb{X}.\beta) + \epsilon = p(x) + \epsilon \tag{29}$$

où ϵ est un bruit aléatoire. Puisque Y ne peut prendre que deux valeurs 0 et 1, $\epsilon = p(x)$ ou $1 - p(x)$. Le bruit est donc ici une variable de Bernoulli prenant uniquement deux valeurs. Par ailleurs, $\mathbb{E}[Y|X = x] = p(x)$, ce qui implique $\mathbb{E}[\epsilon|X] = 0$. Il faut

noter qu'en principe, on n'écrit pas l'équation (29) avec un bruit additif ϵ car celui-ci dépend directement de la distribution (il est donc hétéroscédastique et n'est pas gaussien).

Lorsque la variable Y prend plus de deux modalités on parle de régession logistique polytomique multinomial (appelé également «softmax regression» en apprentissage supervisé). On compare la probabilité de chaque modalité $1,...,K$ par rapport à une modalité de référence (mettons K) :

$$L_k(x) = \left(\frac{\mathbb{P}[Y=k/X=x]}{\mathbb{P}[Y=K/X=x]}\right),\ \forall k = 1,...,K-1 \tag{30}$$

$$= \beta_{0,k} + \beta_{1,k}x_1 + ... + \beta_{p,k}x_p = x.\beta_k \tag{31}$$

Ainsi,

$$\mathbb{P}[Y=k] = \frac{\exp(x.\beta_k)}{\sum_{k=1}^{K}\exp(x.\beta_k)},\ i = k,...,K-1 \tag{32}$$

On peut toujours se débrouiller pour que $\beta_K = 0$, quitte à remplacer β_k par $\beta_k - \beta_K$.

3.2 Estimation des paramètres

L'estimation des paramètres β_i s'effectue par la méthode du maximum de vraisemblance. Y étant une variable de Bernoulli, le modèle d'échantillonnage a pour vraisemblance

$$\prod_{i=1}^{n}\mathbb{P}[Y=y_i|X=x_i] = \prod_{i=1}^{n}\left(p(x_i)^{y_i}(1-p(x_i))^{1-y_i}\right) \tag{33}$$

La log-vraisemblance est donc donnée par

$$l(x,y,\beta) = \sum_{i=1}^{n}\left(y_i \ln p(x_i) + (1-y_i)\ln(1-p(x_i))\right) \tag{34}$$

$$= \sum_{i=1}^{n}\left(y_i x_i.\beta - \ln(1+e^{x_i.\beta})\right) \tag{35}$$

Le gradient de l a pour coordonnées les

$$\frac{\partial l}{\partial \beta_j}(x,y,\beta) = \sum_{i=1}^{n}\left(y_i x_{ij} - \frac{x_{ij}e^{x_i\beta}}{1+e^{x_i\beta}}\right) \tag{36}$$

$$= \sum_{i=1}^{n}\left(x_{ij}(y_i - p(x_i))\right) \tag{37}$$

De façon matricielle, en notant $P = (p(x_1),...,(x_n)) = \Lambda(\mathbb{X}\beta)$ et $W = \text{diagg}(p_i(1-p_i))$, le gradient et la matrice hessienne s'écrivent:

$$\nabla l(x,y,\beta) = \mathbb{X}^T(y-P), \tag{38}$$

$$H(x,y,\beta) = -\mathbb{X}^T W\mathbb{X}. \tag{39}$$

Les équations de vraisemblance n'étant pas linéaires, il faut les résoudre par des méthodes numériques. On obtient alors un estimateur $\widehat{\beta}$.

Sous de bonnes hypothèses de régularité du modèle, on a

$$\sqrt{n}\left(\widehat{\beta}-\beta\right) \rightsquigarrow \mathcal{N}\left(0, \mathbb{I}(\beta)^{-1}\right) \tag{40}$$

3.3 Interprétation des résultats

Les paramètres d'une régression ne sont pas interprétables directement, on doit les comparer les uns aux autres. Dans l'expression de $L(x)$, la fraction

$$\pi(x) = \frac{p(x)}{1-p(x)} \tag{41}$$

s'appelle la cote. Ce n'est rien d'autre qu'un rapport de vraisemblance. Quand la cote est supérieure à 1, un succès est plus probable qu'un échec. Le quotient ρ de deux cotes s'appelle un «odds ratio» (ou rapport de cotes, rapport de chances, ou encore risque relatif). Dans le cas binaire,

$$\rho(1,0) = \frac{\pi(1)}{\pi(0)} = \frac{p(1)/[1-p(1)]}{p(0)/[1-p(0)]} = \exp\beta_1 \tag{42}$$

Si β_1 est positif, la probabilité que Y soit égale à 1 augmente quand x augmente. Si x augmente de 1, le rapport de cotes augmente de $\exp\beta_1$.

Dans le cas d'une régression multiple, on voit facilement que:

$$\rho(1,0) = \exp(\beta_1 + ... + \beta_p) \tag{43}$$

Si l'on considère deux observations qui diffèrent seulement par la j-ième variable, une variation d'une unité de cette variable correspond à un rapport de cotes de $\exp\beta_j$, de sorte que β_j mesure l'influence de la j-ième variable sur la cote $\pi(x)$.

Prenons l'exemple d'une unique variable exogène binaire. $Y = 1,0$ mesure la présence ou l'absence d'une maladie et $X = 1,0$ la présence ou l'absence d'un symptôme.

$$L(x) = \ln\left(\frac{\mathbb{P}[Y=1|X=x]}{\mathbb{P}[Y=0|X=x]}\right) \tag{44}$$

$$= \beta_0 + \beta_1 x \tag{45}$$

$$\rho(1,0) = \exp(\beta_1) \tag{46}$$

Si $\rho(1,0) = 1$ la maladie est indépendante du symptôme. Si $\rho(1,0) > 1$ elle est plus fréquente pour les individus qui ont le symptôme et si $\rho(1,0) < 1$ elle est plus fréquente chez les individus qui n'ont pas le symptôme.

3.4 Validation du modèle et qualité de la régression

Les techniques présentées ici pour évaluer la qualité de l'estimation sont valables pour n'importe quelle méthode d'apprentissage supervisé.

• Matrice de confusion : c'est une matrice 2×2 qui évalue le nombre de bonnes et de mauvaises prédictions entre les valeurs prédites et les valeurs observées. On y distingue les vrais/faux (V/F) positifs/négatifs (P/N), le taux d'erreur, le taux de succès, la sensibilité (TVP), la prévision (TVP parmi ceux classés P) et la spécificité (proportion de N détectés).

• Pseudo R^2 : analogue du R^2 pour la régression logistique. Le R^2 évalue la part de la variance expliquée par le modèle étudié en comparant ses performances avec le modèle de base réduit à la constante. Dans le cas de la régression logistique, il s'agit d'une comparaison de vraisemblance entre les taux d'erreurs. Un pseudo R^2 proche de 1 va dans le sens d'un modèle de bonne qualité.

• Test de Hosmer-Lemeshow : il compare les probabilités $\widehat{p}(x_i)$ estimées par le modèle, aux probabilités observées $p(x_i)$ en quantifiant cet écart par un indice. On pourrait le représenter par un diagramme (de fiabilité) dont les deux axes sont gradués de 0 à 1 et l'on observe l'écart entre les deux distributions de probabilités. Les probabilités sont ordonnées et groupées par décile pour donner la statistique d'un test du χ^2.

• Test de population de Mann-Whitney : il quantifie la différence entre les observations positives et négatives en testant si la population positive est significativement plus élevée.

• Courbe ROC : c'est un outil graphique qui évalue les performances de la régression et compare la sensibilité à la spécificité, par l'intermédiaire d'un nuage de points à deux dimensions. Elle produit également un indice AUC (Area Under Curve) quantifiant ses performances, qui est égal à la probabilité qu'un individu positif soit devant un individu négatif.

• Analyse des résidus.

• L'analogue pour la régression logistique, de la somme des carrés des résidus, est la déviance dont la définition est:

$$D = -2l(x, \beta) \tag{47}$$

où l est la log-vraisemblance. La déviance résiduelle est $d = 2(l(x, \beta) - l(x, \widehat{\beta}))$.

Il faut ensuite effectuer des tests de significativité (globale) des coefficients pour décider si une variable explicative doit être intégrée ou non au modèle : test du rapport de vraisemblance, test de Wald, etc. Ces tests évaluent l'influence (c'est à dire la contribution) d'un régresseur sur la variable à expliquer. Par exemple pour le test de Wald, on pose $Z = \hat{\beta}_i / \sigma(\hat{\beta}_i)$. Sous l'hypothèse nulle $H_0 : \beta_i = 0$, Z étant l'e.m.v., il converge vers une loi normale $Z \rightsquigarrow \mathcal{N}(0,1)$ et la p-valeur associée au test vaut $\mathbb{P}[|Z| \geq |z|]$.

Modèles emboîtés : on utilise souvent des méthodes par récurrence («backward», «forward» ou «both») qui suppriment ou ajoutent une variable puis recalcule l'ensemble des coefficients pour analyser à nouveau leur significativité. Ces méthodes reposent sur des critères statistiques, par exemple AIC (Akaike) ou BIC (Schwartz), évalués à partir de la déviance du modèle.

Il faut garder à l'esprit le fait que moins un modèle possède de variables, plus il sera robuste et facile à interpréter.

3.5 La régression logisitique avec R

On rappelle que la régression linéaire s'obtient avec la commande `reg=lm`($Y \sim X1 + X2 + X3$).

`reg` permet d'obtenir l'estimation ponctuelle des coefficients de β.
`reg$coeff[i]` donne l'estimation de β_{i-1}.
`predict(reg)` calcule les valeurs prédites moyennes de Y prises aux valeurs des données de X_i.
Si β_0 n'a pas de sens, on peut le retirer en tapant :
`reg=lm`($Y \sim X1 + X2 + X3 - 1$.
Enfin, `summary(reg)` affiche les résultats des tests statistiques effectués sur les coefficients pour en mesurer la qualité.

La régression logistique binaire s'obtient avec la commande `glm` :

```
reg=glm(Y~ X1+X2+X3,family=binomial).
reg
```

Avec l'option `family=gaussian` on retrouve la régression linéaire. On peut préciser la forme de la fonction de lien (fonction «link») de la façon suivante: `family=binomial(link="logit")`.

Nous allons tester ces fonctions sur un jeu de données classique. Il s'agit d'une étude médicale datant de 1983, dans laquelle 462 patients ont été suivis pour étudier le risque d'apparation de maladies cardiaques. 12 variables sont disponibles dans le jeu de données complet, mais nous ne garderons que deux d'entre elles : l'âge (age) et la présence ou non de la maladie (chd). Nous travaillerons également avec un échantillon de 100 unités extraites du jeu de données.

1°. À l'aide de l'instruction `read.table` charger l'échantillon `sample.txt` dans R et visualiser les données.

2°. Expliquer pourquoi une régression linéaire n'est pas adaptée, puis visualiser sur le même graphique que celui de la question précédente la proportion de malades en fonction de l'âge.

3°. Effectuer la régression logistique expliquant l'apparition de la maladie en fonction de l'âge. En déduire la valeur des coefficients β_i et les interpréter.

4°. Mesurer la qualité de l'ajustement en calculant différents coefficients R^2 et la déviance D.

5°. Évaluer le pouvoir discriminant du modèle en calculant la sensibilité et la spécificité.

6°. Évaluer la calibration du du modèle en effectuant le test de Hosmer et Lemeshow.

7°. Évaluer la significativité des coefficients en effectuant un test de Wald, puis un test de vraisemblance.

8°. Effectuer une regression logistique multiple à l'aide d'un jeu de données de votre choix. Par exemple, le jeu de données « Iris de Fisher » s'y prête bien:

https://archive.ics.uci.edu/ml/datasets/Iris

3.6 La régression logisitique avec Python

Premier jeu de données : les iris de Fisher

Télécharger le notebook Python contenant la première partie du TP via le lien suivant : https://cpmath.fr/RegLogIRIS.ipynb. Lire les instructions.

1°. Charger le jeu de données à partir des commandes `datasets` de la bibliothèque `scikit learn`. Étudier le fichier de données en effectuant des statistiques descriptives de base et visualiser le nuage de points à l'aide de l'instruction `scatter`.

2°. À l'aide de la fonction `PCA`, effectuer une analyse en composantes principales en 3 dimensions.

3°. À l'aide de la fonction `pairplot`, visualiser les différents nuages de points affichés en fonction de chaque paire de variables. Interpréter ce que vous voyez.

4°. Afin d'effectuer une première régression logistique sur une variable binaire, fusionner les deux premières classes d'iris et afficher le nuage de points correspondant. À l'aide des fonctions `LogisticRegression` de `scikit learn` et `logit` de la bibliothèque `statmodels`, construire un modèle de régression logistique binaire et le tester sur quelques données.

5°. Effectuer une régression logistique polytomique sur les trois classes d'iris.

6°. À l'aide de la fonction `classification_report`, évaluer (très grossièrement) la qualité de la régression.

Second jeu de données : la prédiction des maladies cardio-vasculaires

1°. Télécharger le jeu de données (lien). Décrire la problématique de l'étude. Étudier la base des données en effectuant des statistiques descriptives simples et visualiser le nuage de points de l'âge en fonction de la variable `chd` (utiliser `scatterplot`).

2°. Expliquer pourquoi une régression linéaire n'est pas adaptée, puis visualiser sur le même graphique que celui de la question précédente la proportion de malades en fonction de l'âge (vous pouvez définir, par exemple, des classes d'âge).

3°. À l'aide de la fonction `glm` de la bibliothèque `statmodels`, effectuer la régression logistique expliquant l'apparition de la maladie en fonction de l'âge. En déduire la valeur des coefficients β_i et les interpréter. La syntaxe est la suivante :

```
reglog1 = smf.glm('chd ~ age', data=maladie,
family=sm.families.Binomial()).fit()
print(reglog1.summary())
```

4°. Étudier le tableau contenant les résultats des tests de nullité des coefficients. Que représentent les coefficients ? L'écart type ? le z-score ? Rappeler précisément ce que représente la p-valeur.

5°. Effectuer une nouvelle régression logistique avec l'ensemble des variables et reprendre les questions du 4°. Que pensez-vous de la significativité des variables `sbp` (systolic blood pressure) et `obesity` ?

6°. À l'aide de la fonction `RFE` de `scikit learn`, revoir le modèle en supprimant les variables les moins significatives, par élimination régressive («backward elimination»). Utiliser les tests AIC-BIC ou bien les tests basés sur la probabilité d'erreur. Que représente la déviance ? La déviance résiduelle ? Le résultats des tests AIC-BIC ? Effectuer les tests d'adéquation de la déviance et examiner les résidus. Le modèle final est-il identifiable ?

7°. Effectuer les tests de Wald, du score, du rapport de vraisemblance.

8°. Évaluer la calibration du modèle en effectuant le test de Hosmer et Lemeshow.

9°. Interpréter les résultats de la régression, en particulier les valeurs de certaines cotes. Vous pourrez, par exemple, indiquer comment est calculée la probabilité pour un individu d'un certain âge d'avoir une maladie. Produire des intervalles de confiance.

10°. Découper les données en un sous-ensemble d'apprentissage (d'entrainement) et un sous-ensemble de validation (de test). Produire la table de confusion et évaluer la qualité de la régression.

11°. Tracer la courbe ROC du classifieur binaire.

12°. Conclure.

Bibliographie

• Régression avec R, 2e édition, de P.A. Cornillon, N. Hengartner, E. Matzner-Lober, L. Rouvière (2019). EDP Sciences.

• Logistic Regression, a self learning text, 3rd edition, Kleinbaum, Klein. Springer.

• Applied Logistic Regression, 3rd edition, Hosmer, Lemeshow, Sturdivant (2013). Wiley.