Régression logistique

M. Clertant [1]

Statistiques biomédicales

1. clertant@math.univ-paris13.fr, Lab. Analyse, Géométrie et Applications

Plan

Les données

	age	sexe	sport	smoke	peak	arrhythmya	ECV
1	56	0	90	0	0.5	1.16	0
2	35	1	0	0	0.84	1.01	0
3	83	1	30	1	0.39	1.03	0
4	71	1	0	2	−0.23	0.72	1
5	44	1	120	0	1.19	0.59	0
6	76	0	0	2	2.15	0.41	1
7	41	0	60	0	−1.54	1.78	0
8	94	1	0	2	1.17	1.14	1
9	81	1	30	1	0.66	1.56	0
10	70	1	60	0	0.07	1.86	0
11	55	1	0	0	1.09	0.49	0
12	25	1	60	2	1.24	0.99	0
13	53	0	30	0	−1.61	0.37	0
14	62	0	60	0	2.82	1.11	0
15	29	1	90	0	−0.96	1.55	0

Figure – Les quinze premières lignes de la dataframe `data_ECV` qui en contient 2000. La variable `age` est en année, `sport` est en minute/semaine, `sexe` : 0 pour une femme, 1 pour un homme ; `smoke` : 0 pour non-fumeur, 1 pour fumeur, 2 pour fumeur intensif ; `peak` et `arrhythmya` sont des scores attribués à des observations cliniques sur le système cardiaque des individus ; `ECV` : 1 si la personne à un problème cardiaque (a eu un ECV), 0 sinon.

Comment expliquer ou prédire `ECV` à partir des autres variables ?

Tentative de régression

Y est la variable à expliquer ou à prédire,
aussi appelée variable dépendante (ex : variable `ECV`).
X regroupe les variables explicatives ou covariables (ex : `age`, `sport`, `peak` ...).

Les solutions présentées dans ce cours s'intéresse au cas suivant : $Y \mid X = x$ suit une loi de Bernoulli ou une binomiale ou une multinomiale. Cas d'une Bernoulli : Comment modéliser $Y \in \{0, 1\}$ à partir de covariables X continues ou non ?

Première idée : Si notre modèle prend la forme $Y = f(X)$, on cherche à minimiser sur notre échantillon qq chose de la forme $(Y - f(X))^2$ ou une vraisemblance, mais pas de dérivation possible, f prend des valeurs dans un espace discret.
Solution : Modéliser la probabilité

$$\pi(X) = \mathbb{P}(Y = 1 \mid X).$$

Seconde mauvaise idée (régression linéaire) : $\pi(X) = A^\top X$ avec $A = (a_0, a_1, \ldots, a_n)$ et $X = (1, X_1, \ldots, X_n)$.
Exemple : $A = (-1, 2, 4)^\top$,

$$X_1 = (1, 0.3, 0.2)^\top \Rightarrow \pi(X_1) = 0.4$$

$$X_2 = (1, 0.3, 0.4)^\top \Rightarrow \pi(X_2) = ?$$

"On ne veut pas que la probabilité double quand les covariables sont doublées",
ou dit autrement : $A^\top X \notin [0, 1]$ pose problème.

Théorème de Bayes et odds ratio (OR)

$$\mathbb{P}(Y \mid X) = \frac{\mathbb{P}(Y) \times \mathbb{P}(X \mid Y)}{\mathbb{P}(X)} = \frac{\mathbb{P}(Y) \times \mathbb{P}(X \mid Y)}{\sum_{k=1}^{l} \mathbb{P}(X \mid Y = y_k)\mathbb{P}(Y = y_k)}$$

L'Odds (th. de Bayes avec 2 classes) :

$$\frac{\pi(x)}{1-\pi(x)} = \frac{\mathbb{P}(Y=1 \mid X)}{\mathbb{P}(Y=0 \mid X)} = \frac{\mathbb{P}(Y=1)}{\mathbb{P}(Y=0)} \times \frac{\mathbb{P}(X \mid Y=1)}{\mathbb{P}(X \mid Y=0)}.$$

En rouge : quantité facile à estimer à partir des données.
Comment rendre l'estimation du rapport bleu possible ?

$$\log\left(\frac{\mathbb{P}(X \mid Y=1)}{\mathbb{P}(X \mid Y=0)}\right) = b_0 + b_1 X_1 + \ldots + b_p X_p$$

Modéle semi-paramétrique : pas d'hypothèse directe sur les distributions mais sur le rapport. Le modéle inclus les lois normales, exponentielles, gamma, beta, de Poisson de Bernoulli et mélange de variables binaires (ou a à support fini).

Quelles hypothèse doit-on rajouter pour que $\mathbb{P}(X \mid Y = i), i \in \{0, 1\}$ puissent suivre des lois normales ?

Logit, sigmoïde et régression logistique

On cherche à modéliser : $\pi(X) = \mathbb{P}(Y = 1 \mid X)$.

Fonction logit :

$$\text{logit}(\pi(X)) = \log\left(\frac{\pi(X)}{1-\pi(X)}\right).$$

Modéle de la régression logistique :

$$\text{logit}(\pi(X)) = \log\left(\frac{\pi(X)}{1-\pi(X)}\right) = A^\top X = a_0 + a_1 X_1 + \ldots + a_n X_n.$$

En inversant *logit*, la sigmoïde :

$$\pi(X) = \frac{\exp(A^\top X)}{1+\exp(A^\top X)} = \frac{1}{1+\exp(-A^\top X)}$$

L'Odds

$$\text{O}(X) = \frac{\pi(X)}{1-\pi(X)}$$

Simple à interpréter : $\text{O}(X) = 2$, L'individu de covariable X a deux fois plus de chances d'être positif que négatif.

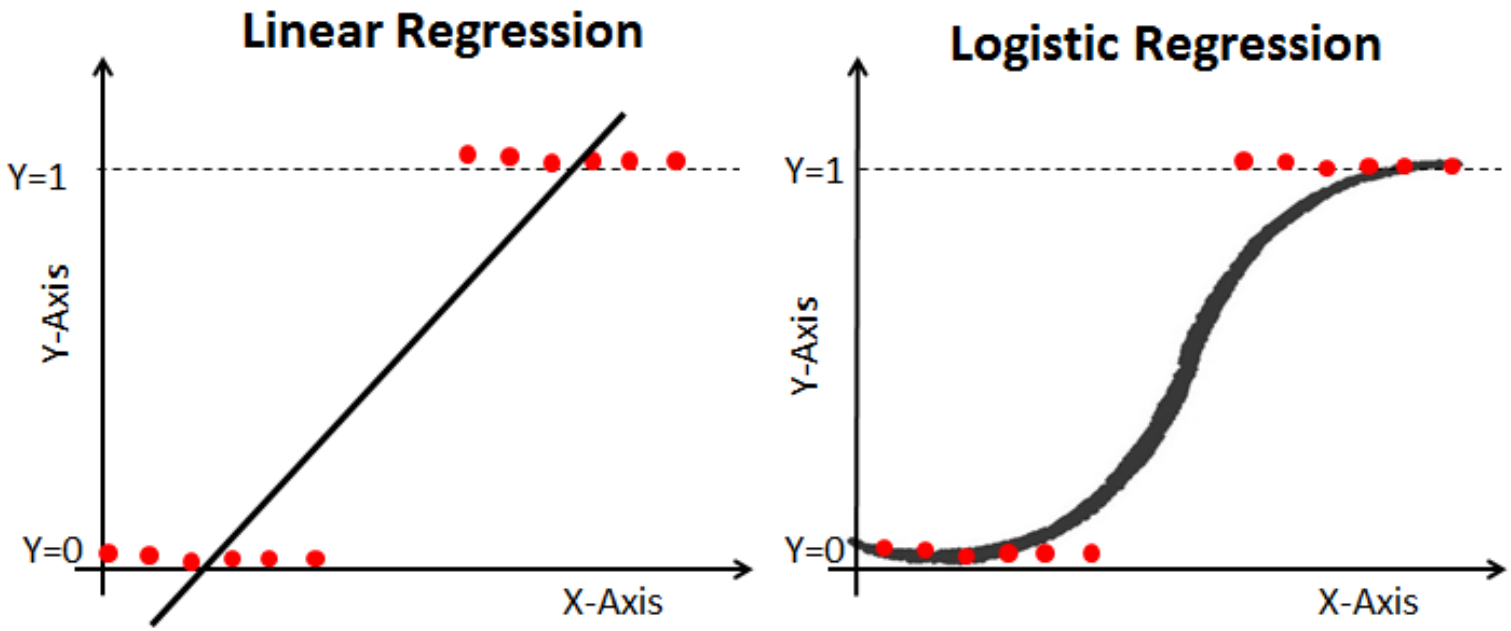

Figure – En rouge, les vraies Y_i. Les deux courbes représentent la probabilité que Y soit positif sachant X : $\pi(X) = 1/(1 + e^{-(a_0+a_1X)})$. (source Datacamp)

- $A^\top X$ varie de $-\infty$ à $+\infty$.
- $0 \leq \pi(X) \leq 1$ est une probabilité

Règle d'affectation : $\pi(X) > 0.5$ ou $\mathrm{O}(X) > 1 \Longrightarrow Y = 1,$; sinon $Y = 0$.

Frontière de décision (1/2)

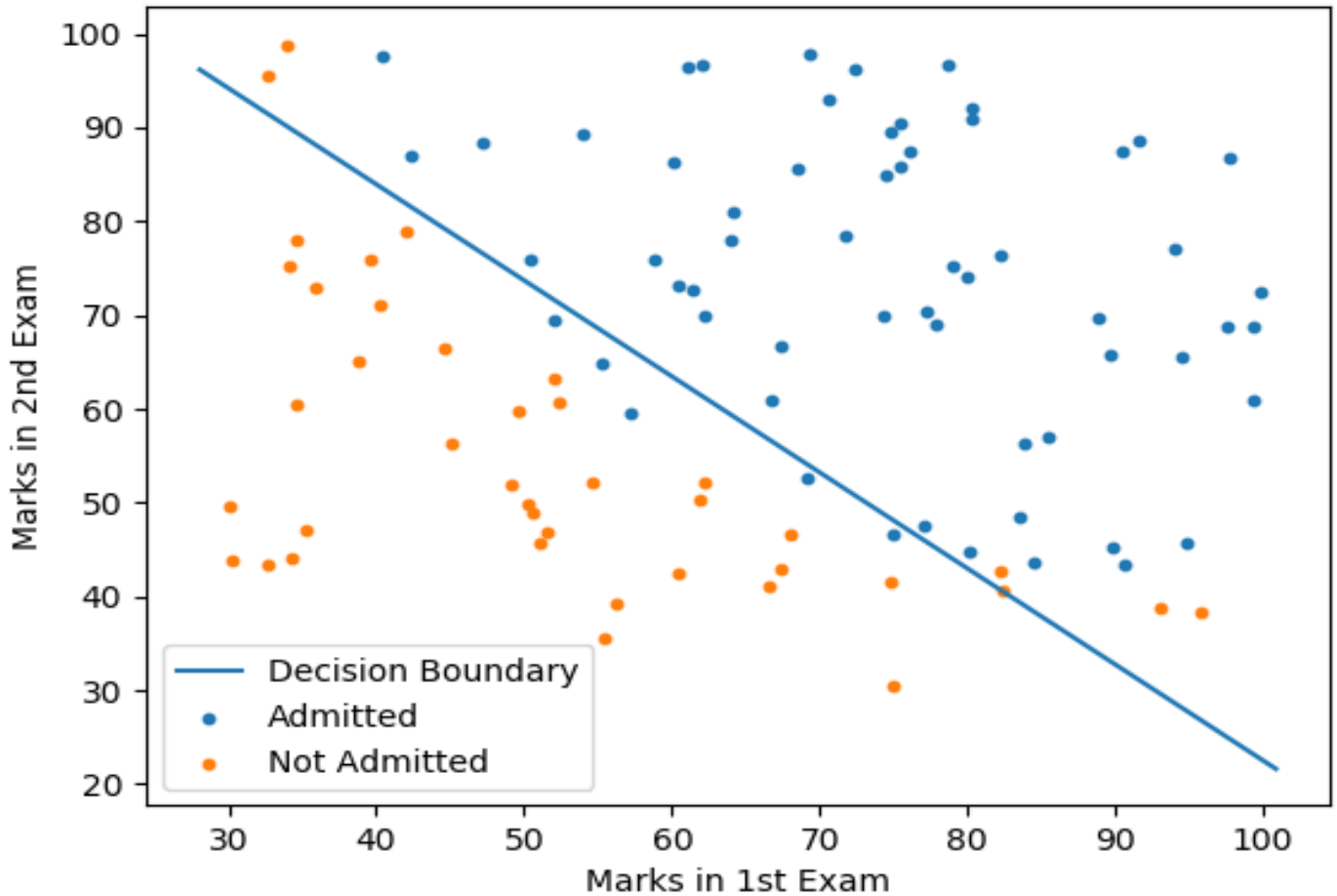

Figure – Y : admission à l'université ; X : notes aux deux premier examens d'un cours en ligne de Andrew Ng.

La frontière est linéaire : $\pi(X) > 0.5 \Leftrightarrow A^\top X = 0$ (hyperplan affine).

Frontière de décision (2/2)

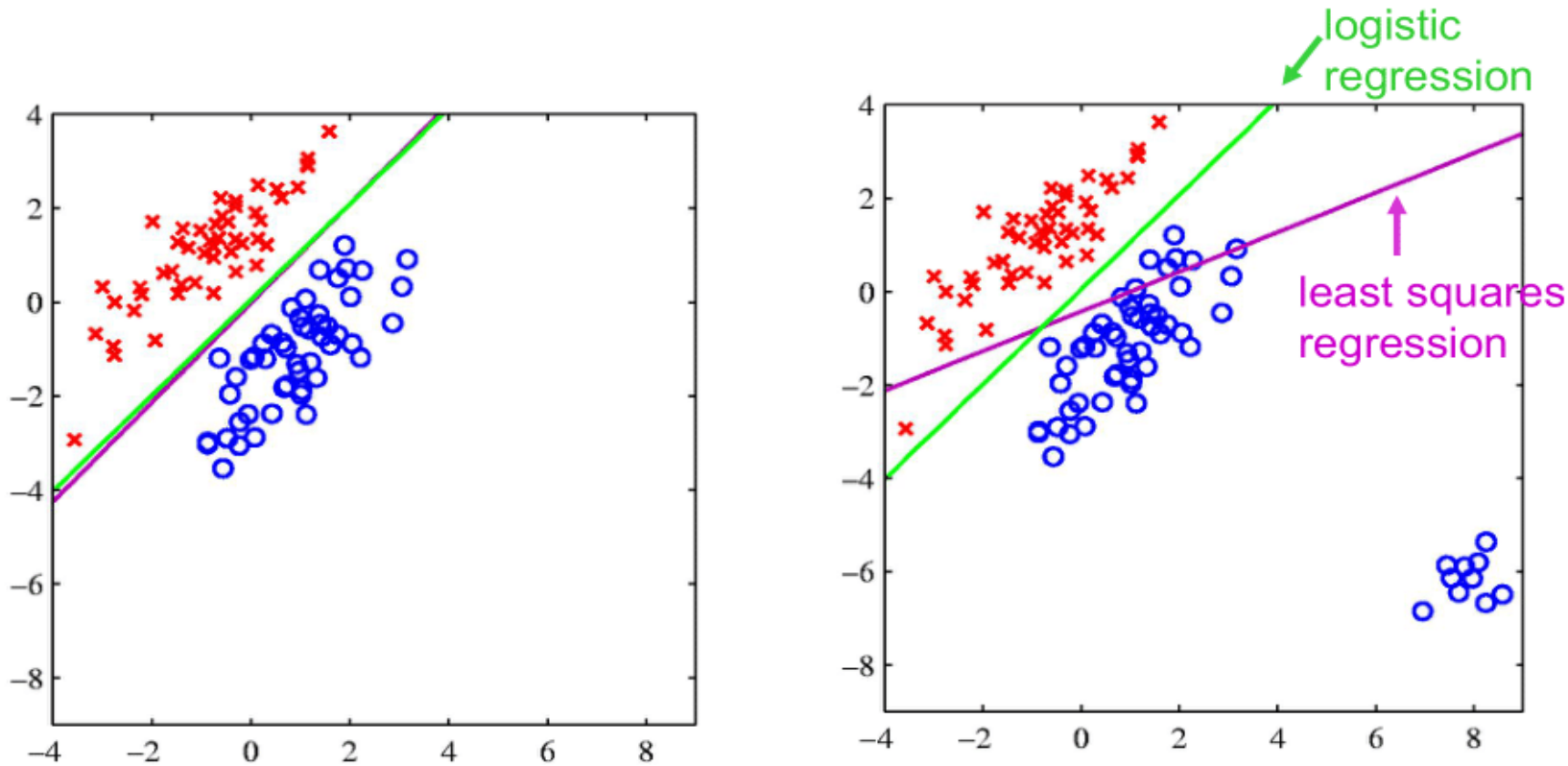

Figure – Les frontières de décision pour la régression logistique et la régression linéaire ajustée par la méthode des moindre carrés. Dans la deuxième image, les points bleus éloignés ont un grand impact sur la position de la frontière de la régression linéaire.

Plan

Un premier modèle avec glm (generalized linear model)

```
 > sum(data_ECV$ECV)
[1] 969
> model <- glm(ECV ~ ., data = data_ECV, family = binomial)
> print(model)
Coefficients:
(Intercept)   age       sexe      sport      smoke
-14.3029    0.21922  -0.12795  -0.04289    0.54349
   peak    arrhythmya
  1.3986     1.58197
Degrees of Freedom: 1999 Total (i.e. Null);  1993 Residual
Null Deviance:      2771
Residual Deviance: 597.5        AIC: 611.5
```

Questions : 0. Combien de personnes dans la table ont-elles connus un EVC dans les trois ans ?
1. D'après le modèle généré ci-dessus, quelle est la probabilité qu'une personne ait connu un épisode cardio-vasculaire lorsque ses caractéristiques sont : femme de 40 ans non sportive et fumeuse occasionnelle avec `peak` et `arrhytmya` à 0 ?
2. Comment varie cette probabilité si cette femme fait 30 minutes de sport par semaine ?

Vraisemblance

Les observations $(y^{(1)}, x^{(1)}), \ldots, (y^{(n)}, x^{(n)})$ sont i.i.d., alors la vraisemblance (fonction des paralmètres A) s'écrit :

$$L(A) = \mathbb{P}(y^{(1)}, \ldots, y^{(n)} | x^{(1)}, \ldots, x^{(n)}; A) = \prod_{i=1}^{n} \mathbb{P}(y^{(i)} \mid x^{(i)}; A).$$

En supposant que $Y \mid X$ suit une Bernoulli (cas binaire) :

$$\begin{aligned}\mathbb{P}(y^{(i)} \mid x^{(i)}; A) &= \mathbb{P}(Y = 1 \mid x^{(i)}; A)^{y^{(i)}} \times \mathbb{P}(Y = 0 \mid x^{(i)}; A)^{1-y^{(i)}} \\ &= \pi(x^{(i)}, A)^{y^{(i)}} \times \big(1 - \pi(x^{(i)}, A)\big)^{1-y^{(i)}}\end{aligned}$$

Estimateur du maximum de vraisemblance

L'apprentissage du modèle se fait en maximisant la vraisemblance :

$$\hat{A} = \max_{A \in \mathcal{A}} L(A) = \max_{A \in \mathcal{A}} \prod_{i=1}^{n} \pi(x^{(i)}, A)^{y^{(i)}} \times \big(1 - \pi(x^{(i)}, A)\big)^{1-y^{(i)}}$$

On peut aussi maximiser la log-vrais., $\log(L(A))$, ou minimiser $-\log(L(A))$.

Comment obtenir l'optimum $\hat{A}$ ou presque ? (1/3)

La solution existe car $-\log(L(A))$ est convexe (on peut trouver son minimum).

Algorithme de Newton-Raphson

On cherche la solution de : $\left(\frac{\partial L(A)}{\partial a_j}\right)_j = 0$

- Choisir une valeur initiale du paramètre : $A^{(0)}$ (au hasard ou selon un a priori)
- Faire une boucle jusqu'à ce que la solution ne change presque pas :

$$A^{(k+1)} = A^{(k)} - \left[\left(\frac{\partial^2 L(A^{(k)})}{\partial a_i \partial a_j}\right)_{i,j}\right]^{-1} \times \left(\frac{\partial L(A^{(k)})}{\partial a_i}\right)_i$$

En rouge, l'inverse de la matrice hessienne peut être difficile à calculer ou à approximer (temps de computation). Quand il n'y a qu'une seule covariable, il s'agit de l'inverse de la dérivée seconde, l'algorithme est rapide.
En bleu, le gradient ne pose pas de problème (ou presque).

Règle d'arrêt : $A^{(k)}$ ne décroît presque plus pendant un nombre fixé d'étapes ; nombre maximum d'étapes.

Comment obtenir l'optimum $\hat{A}$ ou presque ? (2/3)

Algorithme du gradient (gradient descent)

On cherche à minimiser $-\log(L(A))$

- Choisir une ou des valeurs initiales du paramètre A (au hasard ou selon un a priori)
- Faire une boucle en allant dans la direction de la pente (pas λ) :

$$A^{(k+1)} = A^{(k)} - \lambda \times \left(\frac{\partial - logL(A^{(k)})}{\partial a_j} \right)_j$$

Règle d'arrêt : Le gradient devient très petit (plus de déplacement ou presque).

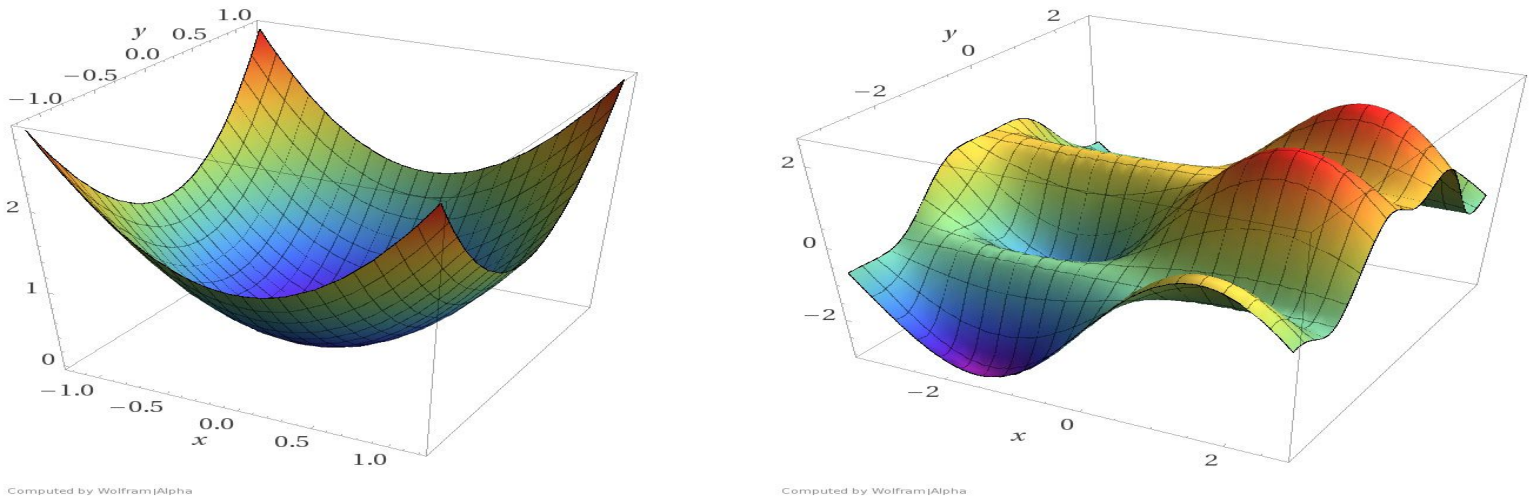

Figure – Une surface convexe (la descente de gradient atteint son minimum) ; une surface non-convexe (minima-locaux).

Exercice : 1. Dans le cadre de l'algorithme du gradient, démontrer que :

$$A_j^{(k+1)} = A_j^{(k)} + \lambda \sum_{i=1}^{n} x_j^{(i)} \left(y^{(i)} - \pi(x^{(i)}; A^{(k)}) \right)$$

avec $x_j^{(i)}$ la j-ième composante de l'observation i.

2. On retient que le gradient est $-\sum_{i=1}^{n} x^{(i)} \left(y^{(i)} - \pi(x^{(i)}; A^{(k)}) \right)$.

Montrer que la matrice hessienne est :

$$\sum_{i=1}^{n} x^{(i)} x^{(i)\top} \pi(x^{(i)}; A^{(k)}) \left(1 - \pi(x^{(i)}; A^{(k)}) \right).$$

Comment obtenir l'optimum $\hat{A}$ ou presque ? (3/3)

Cadre bayésien

- Un a priori sur le paramètre A : $\mathbb{P}(A)$.
- Loi a posteriori :
$$\mathbb{P}(A \mid Y, X) \propto \mathbb{P}(Y \mid X, A)\mathbb{P}(A).$$
- Permet de limiter le surajustement, "diminution de la dimension du paramètre".

Exercice : Soit $\mathbb{P}_A \sim \mathcal{N}(0, \alpha^{-1} Id)$ On cherche à maximiser $\mathbb{P}(A \mid Y, X)$ grâce à un algorithme du gradient. Montrer que :

$$A_j^{(k+1)} = A_j^{(k)} - \lambda \sum_{i=1}^{n} x_j^{(i)} \left(y^{(i)} - \pi(x^{(i)}; A^{(k)}) \right) - \lambda \alpha A_j^k.$$

α est alors un hyperparamètre de notre modèle.

Plan

Rappel

Soient $X_1, \ldots, X_n$ des observations iid suivant la loi f_θ : $X_k \sim f_\theta$.
L'estimateur du maximum de vraisemblance conditionnellement aux X_i est $\hat{\theta}_n$.

Normalité asymptotique de l'emv

Sous des conditions de régularité, on a :

$$\sqrt{n}(\hat{\theta}_n - \theta) \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}(0, I(\theta)^{-1}),$$

où $I(\theta)$ est l'information de Fisher : $I(\theta) = -\mathbb{E}\left[\dfrac{\partial^2}{\partial\theta^2}\log(f_\theta(X))\right]$.

On pose : $\theta = (\theta_1, \ldots, \theta_k) \in \Theta$ et $\theta_0 = (\theta_1, \ldots, \theta_{k'}) \in \Theta_0$ avec $k' < k$ de sorte que f_{θ_0} est un sous-modèle de f_θ : $\Theta_0 \subset \Theta$. On note L la fonction de vraisemblance et $\hat{\theta}$ et $\hat{\theta}_0$ les emv respectifs.

Théorème de Wilks

Sous des conditions de régularité et θ étant le vrai paramètre, si l'hypothèse $H_0 : \theta \in \Theta_0$ est vraie, on a :

$$-2\log\left(\frac{L(\hat{\theta}_0)}{L(\hat{\theta})}\right) \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} \chi^2(k-k').$$

Pseudo-R^2

La log-vraisemblance du modèle de paramètre A est $LL(A) = log(L(A))$ (On la considèrera évaluée en l'estimateur du maximum de vraisemblance $\hat{A}$.)
Si A est réduit à a_0, la log-vraisemblance du modèle est notée $LL(a_0)$.

La déviance : $D = -2LL(\hat{A})$.

McFadden's R^2

$$R^2_{MF} = 1 - \frac{LL(\hat{a}_0)}{LL(\hat{A})}$$

Cox and Snell's R^2

$$R^2_{CS} = 1 - \left(\frac{L(\hat{a}_0)}{L(\hat{A})}\right)^{2/n}$$

$\max R^2_{CS} = 1 - (L(\hat{a}_0))^{2/n}$

Nagelkerke's R^2

$$R^2_{MF} = \frac{R^2_{CS}}{\max R^2_{CS}}$$

Question : Entre quelles valeurs varient les différents R^2? Quels résultats auraient un modèle bien ajusté ?

Test du rapport de vraisemblance

Idée : utiliser le thèorème de Wilks pour construire un test asymptotique sur l'utilité des variables explicatives.

Comparer le modèle avec p covariables contre un modéle réduit avec p' covariables ($p < p'$). Les p' covariables sont incluses dans les p covariables. Les paramètres respectifs de ces modèles sont $\hat{A}_p$ et $\hat{A}_{p'}$.

On a :

$$-2 \times \log\left(\frac{L(\hat{A}_{p'})}{L(\hat{A}_p)}\right) \sim \chi^2(p - p').$$

Question : Pourquoi cette statistique est-elle positive ?

Idées : - Pour évaluer la pertinence d'une variable, on retire cette variable dans le modèle réduit ; si la variable a_i n'est pas significative au seuil α (p-value supérieure à α), l'hypothèse $a_i = 0$ est compatible avec le modèle.
- De la même façon, on peut évaluer un groupe de variables.

Exercice : On rappelle que lorsqu'un modèle de régression logistique a été ajusté à `data_ECV`, le coefficient a_i de la variable `sexe` était de -0.12795 et la déviance du modèle était de 597.5.

1. Comment le coefficient de la variable `sexe` peut-il s'interpréter ?
2. On ajuste le modèle alternatif suivant. Que nous apprend le théorème de Wilks au vu de ces résultats ?

```
> model0 <- glm(ECV ~ .-sexe, data = data_ECV, family = binomi
> print(model0)

Coefficients:
(Intercept)         age       sport       smoke        peak  arrhythmya
   -14.37777     0.21929    -0.04282     0.54275     1.39858     1.58592

Degrees of Freedom: 1999 Total (i.e. Null);  1994 Residual
Null Deviance:      2771
Residual Deviance: 597.9        AIC: 609.9
```

3. Le `model0` ci-dessus pourrait-il être remplacer par un modèle nul (sans variable) ?

Test de Wald

On rappelle que la matrice hessienne de la vraisemblance est donnée par :

$$H = \left(\frac{\partial^2 L(A^{(i)})}{\partial a_i \partial a_j} \right)_{i,j} = -\sum_{i=1}^{n} x^{(i)} x^{(i)\top} \pi(x^{(i)}; A^{(k)}) \left(1 - \pi(x^{(i)}; A^{(k)})\right).$$

Son inverse est la matrice de variance covariance des coefficients estimés (emv).

On a :

$$\hat{A}^\top H \hat{A} \sim \chi^2(p).$$

Si on veut tester l'hypothèse $H_0 : a_j = a_{j+1} = \ldots = a_{j+t} = 0$, on se restreint à ces t coefficients :

$$\hat{A}_{(t)}^\top H_{(t)} \hat{A}_{(t)} \sim \chi^2(t).$$

Pour un seul coefficient a_j : on note $\hat{\sigma}_j^2$ l'inverse du j-ème coefficient de la matrice hessienne. On a :

$$\frac{\hat{a}_j^2}{\hat{\sigma}_j^2} \sim \chi^2(1).$$

Exercice : La fonction `summary` donne un accès au test de Wald composante par composante. Interpréter les résultats ci-dessous.

```
> summary(model)
Coefficients:
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept) -14.302932    0.880025 -16.253  < 2e-16 ***
age           0.219221    0.012709  17.249  < 2e-16 ***
sexe         -0.127951    0.210536  -0.608 0.543360
sport        -0.042886    0.003933 -10.905  < 2e-16 ***
smoke         0.543486    0.156229   3.479 0.000504 ***
peak          1.398635    0.131807  10.611  < 2e-16 ***
arrhythmya    1.581971    0.193721   8.166 3.18e-16 ***
```

Valable pour tous les algorithmes de classification.

Apprendre le modèle sur une partie de l'échantillon (training set) et le tester sur la partie restante (validation set).

Matrice de confusion

n=165	**Predicted: NO**	**Predicted: YES**	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

Validation croisée à k blocs (k-folds cross-validation)

- Découpé l'échantillon en k sous échantillons
- Sélectionner un bloc comme échantillon test et utilisér les $k-1$ restants comme échantillon d'apprentissage
- Répéter l'étape précédentes pour chacun des blocs et faire la moyenne des estimations obtenus (matrice de confusion).

En pratique avec R

Pour obtenir, les matrices de confusions suivantes, on a crée une partition d'indice dans le jeu de données de 2000 lignes à l'aide du package `caret`. Les modèles ont été entraînés sur `data.A` (échantillon d'apprentissage) et sont testé sur `data.T` (échantillon test). L'évaluation a été faîte à l'aide la fonction `predict` qui permet d'appliquer le modèle ajusté à une dataframe.

```
> Empiric_eval(model, data.T)
    pred
       0   1
  0  294  16
  1   13 277
[1] 0.04833333

> Empiric_eval(model_I, data.T)
    pred
       0   1
  0  296  14
  1   10 280
[1] 0.04
```

Dans la section suivante, on explique comment construire le `model_I`.

Plan

Odds Ratio et Risque relatifs

Nombre de personnes	Malades	Non Malades
Exposés	a	b
Non-exposés	c	d

Risque de tomber malade :

Risque chez les exposés $= \frac{a}{a+b}$

Risque chez les non-exposés $= \frac{c}{c+d}$

Risque Relatif $= \frac{\text{Risque chez les exposés}}{\text{Risque chez les non-exposés}} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$

Différence de Risque $= \frac{a}{a+b} - \frac{c}{c+d}$

= Risque chez les exposés – Risque chez les non-exposés

Odds : $O(M = 1 \mid E = 1) = \dfrac{\mathbb{P}(M = 1 \mid E = 1)}{\mathbb{P}(M = 0 \mid E = 1)} = \dfrac{a/(a+b)}{b/(a+b)}.$

Odds-ratio : $OR = \dfrac{O(M = 1 \mid E = 1)}{O(M = 1 \mid E = 0)} = \dfrac{a \times d}{b \times c}.$

Prévalence faible ($a, c << b, d$) : $OR \approx RR$. OR est moins sensible à l'échantillonnage que RR.

Interprétation des coefficients

Le coefficient $\hat{a}_i$ s'interprète comme le logarithme de l'odds pour la variable explicative x_i.

$$\frac{\pi(X;A)}{1-\pi(X;A)} = \exp(a_0 + a_1 x_1 + \ldots + a_i x_i + \ldots + a_p x_p)$$

Si x_i est une variable explicative binaire (homme femme, traités placebo), $\exp(\hat{a}_i)$ mesure le surcroît de risque introduit par le facteur x_i.

Dans le cas des variables quantitatives (âge, poids, biofacteurs), $\exp(\hat{a}_i)$ mesure le surcroit de risque dû à l'augmentation de $+1$ sur la variable.
Comme on a :

$$\pi(X;A) = \frac{1}{1+e^{-(a_0+a_1 x_1+\ldots+a_i x_i+\ldots+a_p x_p)}},$$

il est aussi possible d'obtenir l'augmentation de la probabilité de $Y = 1$ quand la covariable x_i change conditionnellement aux autres covariables (moyennes sur les covariables).

Échantillonnage 1/2

```
>  predict(model, newdata = data.frame(age=40, sexe=0,
  sport=0, smoke=1, peak=1, arrhythmya=1, ECV=0), type="respons
          1
0.1273379
```

Une femme de 40 ans en bonne santé, qui fume modérément et ne fais pas de sport a une probabilité de 12.7% d'avoir un problème cardiaque. C'est beaucoup ... et cela est dû à l'échantillonnage :

- Dans la dataframe, les 969 personnes ayant connu un événement cardiaque proviennent d'une base de données d'un hôpital. Ce sont des patients.
- Les autres sont des personnes tirées au sort dans la population "non-malades".

On ne peut plus interpréter les probabilités données par le modèle en l'état et peut-être est-ce le cas des coefficients a_i aussi.

La régression logistique permet de se sortir de ce mauvais pas élégamment.

Pour cela, on suppose que les personnes ont été tirées de la population standard selon le schéma suivant :

- Tirer une personne de la population standard.
- Si la personne est "malade", la conserver dans l'échantillon avec une probabilité τ_1,
- Si la personne est "non-malade", la conserver dans l'échantillon avec une probabilité τ_0.

Échantillonnage 2/2

Soit la variable $S = 1$, si la personne fait partie de l'échantillon, 0 sinon.
La régression logistique ajusté à l'échantillon est :

$$\text{logit}(\pi(X)) = \text{logit}(\mathbb{P}(Y = 1 \mid X, S = 1)) = A^\top X.$$

On souhaiterait connaître le modèle :

$$\text{logit}(\pi_0(X)) = \text{logit}(\mathbb{P}(Y = 1 \mid X)) = B^\top X.$$

Propriété

Sous les conditions d'échantillonnage citées, on a :

$$\textit{logit}(\pi_0(X)) = \textit{logit}(\pi(X)) - \log(\frac{\tau_1}{\tau_0}) = A^\top X - \log(\frac{\tau_1}{\tau_0})$$

Exercice :1. Montrer la propriété ci-dessus. On utilisera le théorème de Bayes, notamment :

$$\pi(x) = \frac{\pi_0(x)\mathbb{P}(S = 1 \mid Y = 1)}{\mathbb{P}(S = 1 \mid X = x)}.$$

2. Il n'y a que 2 % de la vraie population qui a des problèmes cardiaques. Quelle est la probabilité π_0 pour une femme de 40 ans en bonne santé, qui fume modérément et ne fais pas de sport ($\pi(x) = 12.7\%$)

Les données explicatives

Que les variables explicatives soient nominales ($K > 2$ modalités) ou ordinales (modalités ordonnées), on peut "binariser" les données pour les interpréter. Attention à la conservation de l'ordre pour les données ordinales.

Exemple : Maximum de glycémie rangé en quatre classes :< 1.2, ≥ 1.2, ≥ 2 et ≥ 2.5

- $x_1 = 1$ si glycémie ≥ 1.2, 0 sinon.
- $x_2 = 1$ si glycémie ≥ 2, 0 sinon.
- $x_3 = 1$ si glycémie ≥ 2.5, 0 sinon.

On peut aussi modéliser les interactions.

Exemple : $X = (1, x_1, x_2)$

$$\frac{\pi(X;A)}{1-\pi(X;A)} = \exp(a_0 + a_1x_1 + a_2x_2 + a_3(x_1 \times x_2))$$

a_3 est le coefficient modélisant l'interaction positive de x_1 et x_2.

Exercice : 1. Quelle(s) variable(s) pourraient être binarisées dans le jeu de données `data_ECV` ? Expliquer comment le faire si cela est possible.

```
> head(data_ECV,5)
  age sexe sport smoke  peak arrhythmya ECV
1  33    1    90     0 -0.73       1.02   0
2  75    1    60     0  0.50       0.00   1
3  65    0     0     1  0.65       1.27   1
4  31    0    90     0  1.33       1.37   0
5  95    0     0     2  0.50       0.59   1
```

Exercice : 1. Quelle(s) variable(s) pourraient être binarisées dans le jeu de données `data_ECV` ? Expliquer comment le faire si cela est possible.

```
> head(data_ECV,5)
  age sexe sport smoke  peak arrhythmya ECV
1  33    1    90     0 -0.73       1.02   0
2  75    1    60     0  0.50       0.00   1
3  65    0     0     1  0.65       1.27   1
4  31    0    90     0  1.33       1.37   0
5  95    0     0     2  0.50       0.59   1
```

2. Après avoir binariser, la variable `smoke` sous la forme `(smoke.1, smoke.2)`, on obtient le coefficient a_i pour `smoke.2` égal à -0.12. Interpréter ce résultat. Que peut-on (doit-on) alors faire ?

3. Les médecins suggère que la conjonction d'un haut score pour `peak` et `arrhythmya` est le signe d'un état faible du système cardio-vasculaire. Expliquer comment modéliser cela dans `R`.

Données catégorielles à expliquer

Y prend ses valeurs dans K classes : $Y \in 1, \ldots, K$

On utilise une classe pivot ; par ex. la classe K.

On pose :

$$\log\left(\frac{\mathbb{P}(Y=1 \mid X)}{\mathbb{P}(Y=K \mid X)}\right) = A_1^\top X$$

$$\ldots\ldots$$

$$\log\left(\frac{\mathbb{P}(Y=K-1 \mid X)}{\mathbb{P}(Y=K \mid X)}\right) = A_{K-1}^\top X.$$

Il s'agit alors de déterminer les coefficients de $K-1$ régression logistique. Les probabilités sont alors facilement déterminés en remarquant que :

$$\mathbb{P}(Y=K \mid X) = 1 - \sum_{i=1}^{K-1} \mathbb{P}(Y=i \mid X) = 1 - \mathbb{P}(Y=i \mid X) \sum_{i=1}^{K-1} e^{A_i^\top X}$$

Hypothèse d'indépendance des alternatives non pertinentes : À la fin du repas, la probabilité relative de choisir îles flottantes ou tarte tatin en dessert n'est pas modifiée par la présence ou non sur la carte de la crème au caramel.

Sélection automatique de variables

Problème : Parmi p variables explicatives, comment sélectionner un sous-ensemble de k variables qui permettent une bonne performance du modèle tout en évitant un sur-ajustement (le modéle "colle aux données" mais ne prédit pas correctement) ?

Critère basé sur la vraisemblance maximale :
Lorsqu'on l'on ajoute un paramètre, passant de k à $k+1$, le maximum de la fonction de vraisemblance augmente : $L(\hat{A}_k) \leq L(\hat{A}_{k+1})$.

La vraisemblance est un bon critère d'ajustement, mais il faut le compenser.

Akaike Information Criterion (AIC) :

$$AIC = -2\log(L(\hat{A})) + 2k$$

Bayesian Information Criterion (BIC) :

$$AIC = -2\log(L(\hat{A})) + k\log(n)$$

Question : Quelle lien y a-t-il entre l'AIC et le test basé sur le théorème de Wilks ?

On cherche à minimiser l'AIC et le BIC. Cependant, parmi p variables, il y a 2^p sous-ensembles possibles et autant de modèle à comparer.
Dans la fonction `step` de `R`, le paramètre `direction` à trois options : `backward`, `upward` et `both` ; il s'agit de sélection pas à pas, on ajoute ou on retire une variable.

Modéle linéaire généralisé - Generalized Linear Model (GLM)

On cherche à modéliser $\mathbb{E}(Y \mid X)$ par une transformation inversible d'une relation linéaire entre X et un paramètre A. Trois ingrédients pour un GLM :

Famille exponentielle

$Y \mid X = x$ suit une loi de densité :

$$f(y \mid \theta, \eta) = h(y \mid \eta, \theta) \times \exp\left(\frac{b(\theta, \eta)T(y) - a(\theta)}{c(\eta)}\right)$$

Loi normale, exponentielle, gamma, beta, khi^2, wishart, de Poisson, de Bernoulli, binomiale, multinomiale ...

Relation linéaire

$A^\top X$ avec $X = (1, x_1, \ldots, x_p)^\top$ et $A = (a_0 \ldots, a_p)^\top$.

Fonction de lien g

$$\mathbb{E}(Y \mid X) = g^{-1}(A^\top X)$$

Régression linéaire : $Y \mid X = x$ suit une loi normale et g est la fonction identité.

Exemple GLM : régression de Poisson

Distribution

$Y \mid X = x$ suit une loi de Poisson $\mathcal{P}(\lambda_x)$.
Hypothèse : La variance et l'espérance sont égales à λ_x.

Fonction de lien : logarithme

$$\log(\lambda_X) = \log\left[\mathbb{E}(Y \mid X)\right] = A^\top X$$

On a donc :

$$\mathbb{P}(Y = y \mid X = x) = \frac{\lambda_x^y}{y!} e^{-\lambda_x} = \frac{e^{yA^\top x} e^{-e^{A^\top x}}}{y!}.$$

et la vraisemblance pour le $n-$ échantillon $(x^{(i)}, y^{(i)})_{1 \leq i \leq n}$ est :

$$L\left(A; (x^{(i)}, y^{(i)})_{1 \leq i \leq n}\right) = \prod_{i=1}^{n} \frac{e^{y^{(i)} A^\top x^{(i)}} e^{-e^{A^\top x^{(i)}}}}{y^{(i)}!}$$

Remarque : La régression de Poisson est notamment utilisée pour modéliser le nombre d'événements dans un laps de temps donné ou dans une zone donnée ; conditionnellement à X, les survenues des événements sont considérées indépendantes (hypothèse du modèle).