Introduction Stats Econométrie

Régression Logistique + Tests du χ_2

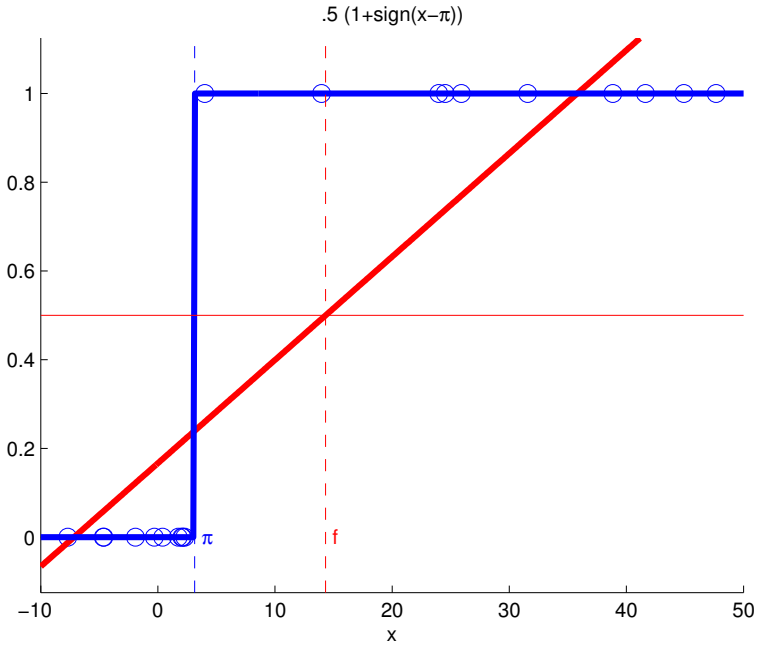
marco.cuturi@ensae.fr

Quand la régression par moindre carrés ne marche pas

- Considérons le problème suivant:
  - Des points $\mathbf{x}_j$ sont pris aléatoirement entre -10 et 50.
  - Le label associé

$$y_j = \begin{cases} 0 \text{ if } \mathbf{x}_j < \pi, \\ 1 \text{ if } \mathbf{x}_j > \pi. \end{cases}$$

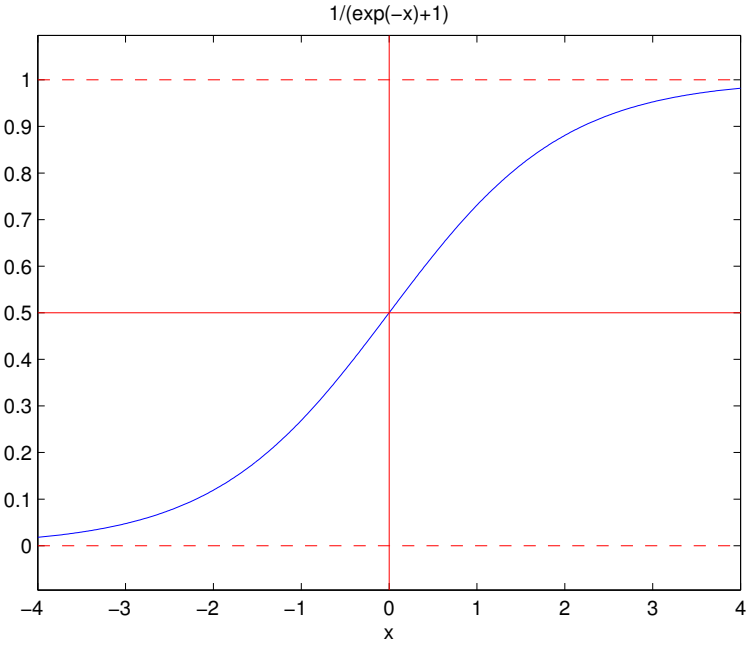
- Que se passe-t-il si nous utilisons ces données directement l'estimateur des moindres carrés?... **demo**



Comment adapter la régression? courbe logistique

- Courbe logistic :

$$g(z) = \frac{e^z}{e^z + 1} = \frac{1}{e^{-z} + 1}$$



  - Pour tout z, $0 \leq g(z) \leq 1$

Comment adapter la régression? fonction logistique

Idée fondamentale

- Plutôt que de trouver $\mathbf{c}$ et b tels que

$$f(\mathbf{x}_j) = \mathbf{c}^T\mathbf{x}_j + b \quad \approx \quad y_j \in \{0, 1\}$$

- La régression logistique considère plutôt les $\mathbf{c}$ et b tels que

$$g \circ f(\mathbf{x}_j) = \frac{1}{e^{-(\mathbf{c}^T\mathbf{x}_j+b)} + 1} \approx y_j \in \{0, 1\}.$$

- Si pour une valeur $\mathbf{x}$,
  - $g \circ f(\mathbf{x}) > 1/2$, on prédit une valeur 1
  - $g \circ f(\mathbf{x}) < 1/2$, on prédit une valeur 0

Interprétation Probabiliste de la Régression Logistique

- Supposons qu'il existe une densité $\boldsymbol{p}(X, Y)$ des couples $(\mathbf{x}, y) \in \mathbb{R}^d \times \{0, 1\}$.
- Supposons que nous connaissions $\boldsymbol{p}$.

- Le ratio

$$r(\mathbf{x}) = \frac{p(Y = 1|X = \mathbf{x})}{p(Y = 0|X = \mathbf{x})}$$

  est le odds-ratio ("rapport des chances") d'un point $\mathbf{x}$.

- Bien évidemment,
  - Si $r(\mathbf{x}) > 1$, il est plus probable que $y = 1$ plutôt que $y = 0$.
  - Si $r(\mathbf{x}) < 1$, il est plus probable que $y = 0$ plutôt que $y = 1$.

Interprétation Probabiliste de la Régression Logistique

- En d'autres terms...

$$\log \frac{p(Y=1|X=\mathbf{x})}{p(Y=0|X=\mathbf{x})}, \quad \begin{cases} > 0 \text{ alors } y=1 \text{ est la réponse probable} \\ < 0 \text{ alors } y=0 \text{ est la réponse probable} \end{cases}$$

- La régression logistique **assume** que le log odds-ratio évolue selon une relation **linéaire**:

$$\log \frac{p(Y=1|X=\mathbf{x})}{p(Y=0|X=\mathbf{x})} \approx \mathbf{c}^T\mathbf{x} + b$$

- Ceci implique que la surface décision est linéaire/

La régression logistique n'assume
l'existence d'un modèle que pour le log-odds ratio,
pas pour la probabilité p dans son intégralité.

Interprétation Probabiliste de la Régression Logistique

- Puisque $p(Y = 0|X = \mathbf{x}) = 1 - p(Y = 1|X = \mathbf{x})$, nous avons

$$\log \frac{p(Y = 1|X = \mathbf{x})}{1 - p(Y = 1|X = \mathbf{x})} = \mathbf{c}^T\mathbf{x} + b$$

- ce qui implique

$$p(Y = 1|X = \mathbf{x}) = \frac{1}{e^{-(\mathbf{c}^T\mathbf{x}+b)} + 1} = g(\mathbf{c}^T\mathbf{x} + b).$$

Les variables prédictives contribuent **linéairement**
à l'acroissement du log odds-ratio, et donc à la probabilité $y = 1$.

Estimation de $\mathbf{c}$ et b par MV

- Loi Bernoulli, si $p(y=1)=p$ et $p(y=0)=1-p$ pour variable binaire aléatoire y,
  - Vraisemblance d'un tirage y sachant que le paramètre est p:

$$p^y(1-p)^{1-y}$$

- Dans le contexte de la **régression logistique**, p dépend de $\mathbf{c}, b$ et $\mathbf{x}_j$ pour chaque point,

$$\mathcal{L}(\mathbf{c}, b) = \prod_{j=1}^{N} g(\mathbf{c}^T\mathbf{x}_j + b)^{y_j}(1 - g(\mathbf{c}^T\mathbf{x}_j + b))^{1-y_j}$$

Estimation de $\mathbf{c}$ et b par MV

- En passant aux log,

$$\log \mathcal{L}(\mathbf{c}, b) = \sum_{j=1}^{N} y_j \log g(\mathbf{c}^T \mathbf{x}_j + b) + (1 - y_j) \log(1 - g(\mathbf{c}^T \mathbf{x}_j + b)).$$

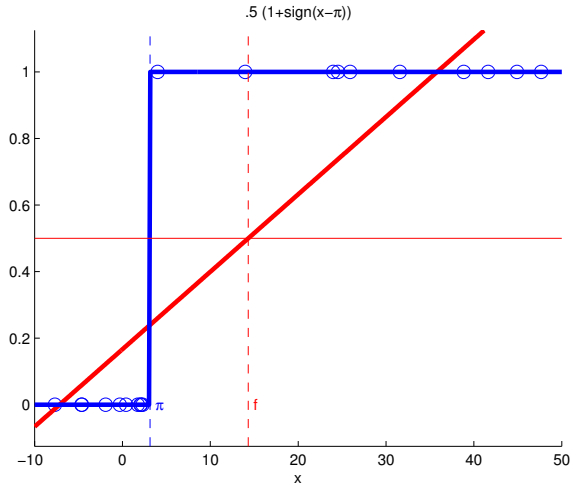
- Maximiser cette log-vraisemblance est équivalent à

$$\max_{\mathbf{c},b} \log \mathcal{L}(\mathbf{c}, b) \Leftrightarrow \max_{\mathbf{c},b} \sum_{j=1}^{N} y_j(\mathbf{c}^T \mathbf{x}_j + b) - \log(1 + e^{\mathbf{c}^T \mathbf{x}_j + b}).$$

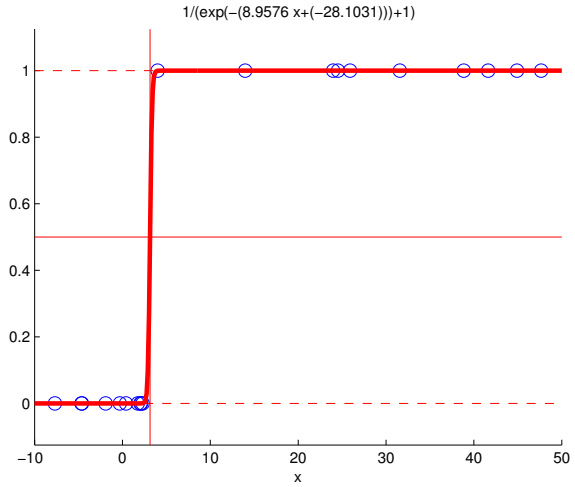
- Pas de forme close!!!... besoin d'un algorithme d'optimisation efficace.
- Pour des jeux de données raisonnables, méthode de Newton.

Estimation de $\mathbf{c}$ et b par MV

Comparons...



...avec



Tests du χ_2

Test d'adéquation à une loi multinomiale

- On observe: échantillon de données $y_1, \ldots, y_N$ d'une variable aléatoire Y qui prend un nombre fini K de valeurs.
- On veut tester l'hypothèse nulle selon laquelle les probabilités que Y prenne les valeurs 1 à K sont respectivement $p_1, \ldots, p_K$ avec $\sum_{j=1}^{K} p_j = 1$.
- Soit $\hat{p}_j$ la probabilité empirique que Y prenne la valeur j, *i.e.* le nombre d'observations qui prennent la valeur j dans l'échantillon, divisé par N :

$$\hat{p}_j = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{y_i = j}.$$

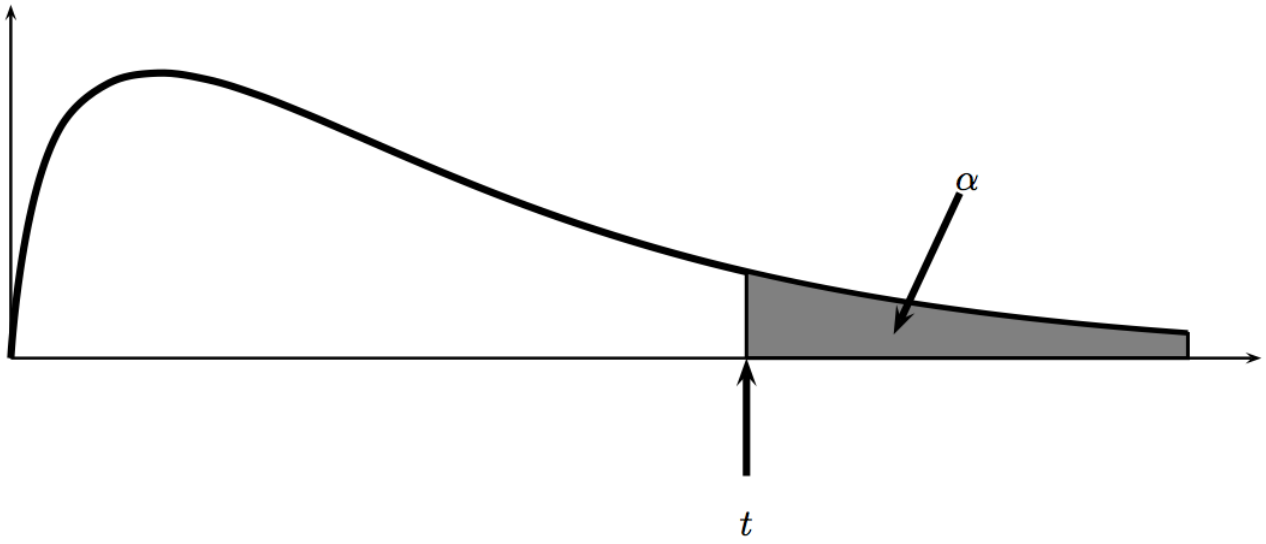
- Statistique du χ^2 :

$$T = \sum_{j=1}^{K} \frac{(N\hat{p}_j - Np_j)^2}{Np_j} = N \sum_{j=1}^{K} \frac{(\hat{p}_j - p_j)^2}{p_j} =$$

Test d'adéquation à une loi multinomiale

- Sous H_0, T suit *asymptotiquement* une loi du χ_2 à $(K-1)$ degrés de libertés
- On peut donc construire un test de niveau α en rejetant l'hypothèse nulle lorsque la statistique de test est plus grande que $F^{-1}_{\chi^2(K-1)}(1-\alpha)$, le quantile d'ordre $1-\alpha$ de la loi du χ_2 à $(K-1)$ degrés de libertés :

$$T \geqslant F^{-1}_{\chi^2(K-1)}(1-\alpha).$$



Preuve

- Soit donc Y une v.a. telle que, pour $1 \leq j \leq K$,

$$P(Y = j) = p_j.$$

- On considère le vecteur aléatoire suivant:

$$Z = \left(\frac{\mathbf{1}_{Y=1} - p_1}{\sqrt{p_1}}, \frac{\mathbf{1}_{Y=2} - p_2}{\sqrt{p_2}}, \ldots, \frac{\mathbf{1}_{Y=K} - p_K}{\sqrt{p_K}} \right).$$

- Si on note $Z = (Z^{(1)}, \ldots, Z^{(K)})$ les composantes de Z, nous avons:
  - $\forall i,\ \mathbb{E}(Z^{(i)}) = 0$
  - $\forall i, \mathrm{Var}(Z^{(i)}) = 1 - p_i$
  - $\forall i \neq j, \mathrm{Cov}(Z^{(i)}, Z^{(j)}) = -\sqrt{p_i p_j}$
- Z est un vecteur aléatoire centré dont la matrice de variance-covariance est

$$\Gamma = I_K - \sqrt{p}\sqrt{p}^T$$

.

Preuve

- Si l'on dispose d'un échantillon $y_1, \ldots, y_N$ de la variable Y, on en déduit un échantillon $z_1, \ldots, z_N$ de la variable Z.
- Le théorème central limite dit alors que quand N tend vers l'infini,

$$\frac{Z_1 + \ldots + Z_N}{\sqrt{N}} \sim_{N \to \infty} \mathcal{N}(0_K, \Gamma)$$

- Mais cette loi n'est autre que celle du projeté d'un vecteur aléatoire de $\mathbb{R}^K$ suivant une loi normale centrée réduite sur l'hyperplan orthogonal à $\sqrt{p}$ (espace de dimension $K-1$).

$$\Gamma = I_K - \sqrt{p}\sqrt{p}^T = (I_K - \sqrt{p}\sqrt{p}^T)(I_K - \sqrt{p}\sqrt{p}^T)$$

.

- $Z \sim_N PX$ où $P = I_K - \sqrt{p}\sqrt{p}^T$ est un projecteur de rang $K-1$.
- $\|PX\|^2$ suit alors une loi du χ_2 à $(K-1)$ degrés de libertés.
- C'est la loi limite de $\left(\frac{Z_1+\ldots+Z_N}{\sqrt{N}}\right)^2$ qui n'est autre que la statistique T.

Example

- Le lancement d'un dé 600 fois de suite a donné les résultats suivants :

numéro	1	2	3	4	5	6
effectifs	88	109	107	94	105	97

- Degrés de liberté: $6 - 1 = 5$.
- hypothèse: le dé n'est pas truqué. Risque $\alpha = 0{,}05$.
- Statistique T:

$$\frac{(88-100)^2}{100} + \frac{(109-100)^2}{100} + \frac{(107-100)^2}{100} + \frac{(94-100)^2}{100} + \frac{(105-100)^2}{100} + \frac{(97-100)^2}{100} = 3{,}44.$$

- La loi du χ_2 à cinq degrés de liberté donne la valeur en dessous de laquelle on considère le tirage comme normal avec un risque $\alpha = 0{,}05$:
  $P(T < 11{,}07) = 0{,}95$. Comme $3{,}44 < 11{,}07$, on ne considère pas ici que le dé soit truqué.

Test d'indépendance

- But: vérifier l'absence de lien statistique entre deux variables X et Y.
- X et Y indépendants $\Leftrightarrow$ aucun lien statistique
- H_0 : les deux v.a. X et Y sont indépendantes.
- Par exemple, X et Y désignent des qualités physiques (couleur des yeux, boxy mass index).
- L'hypothèse à tester est l'indépendance entre ces deux qualités.
- X et Y prennent un nombre fini de valeurs, I pour X, J pour Y.
- On dispose d'un échantillon de N individus. Notons O_{ij} l'effectif observé d'individus pour lesquels $X = i$ et $Y = j$.
- Sous l'hypothèse d'indépendance, on s'attend à une valeur espérée

$$E_{ij} = \frac{O_{i+} \times O_{+j}}{N}$$

où

$$O_{i+} = \sum_{j=1}^{J} O_{ij}, \qquad O_{+j} = \sum_{i=1}^{I} O_{ij}$$

Test d'indépendance

- On calcule la distance χ_2 entre les valeurs empiriques O_{ij} et celles attendues en cas indépendance E_{ij}

$$T = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = N \sum_{i,j} \frac{\left((O_{i,j}/N) - p_{i+}p_{+j}\right)^2}{p_{i+}p_{+j}},$$

  où $p_{i+} = O_{i+}/N, p_{+j} = O_{+j}/N$.

- On montre ici aussi que $T \sim \chi^2_{(I-1)(J-1)}$ degrés de liberté.

Example: Vaccination

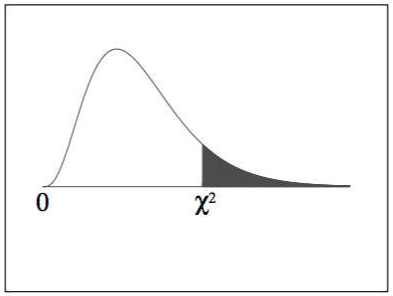
Health Outcome	Not vaccinated Col 1	Vaccinated Col 2	Row marginals (Row sum)
Sick with pneumococcal pneumonia	23	5	**28**
Sick with non-pneumococcal pneumonia	8	10	**18**
Stayed healthy	61	77	**138**
Column marginals (Sum of the column)	**92**	**92**	**N = 184**

Cell expected values and (cell Chi-square values).

Health outcome	Not vaccinated	Vaccinated
Sick with pneumococcal pneumonia	13.92 (5.92)	12.57 (4.56)
Sick with non-pneumococcal pneumonia	8.95 (0.10)	9.05 (0.10)
Stayed healthy	69.12 (0.95)	69.88 (0.73)

- Statistique du χ_2 est $T = 12.35$
- Degrés de liberté: $(3-1)(2-1) = 2$.
- Rejet de l'hypothèse nulle.

Chi-Square Distribution Table



The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750