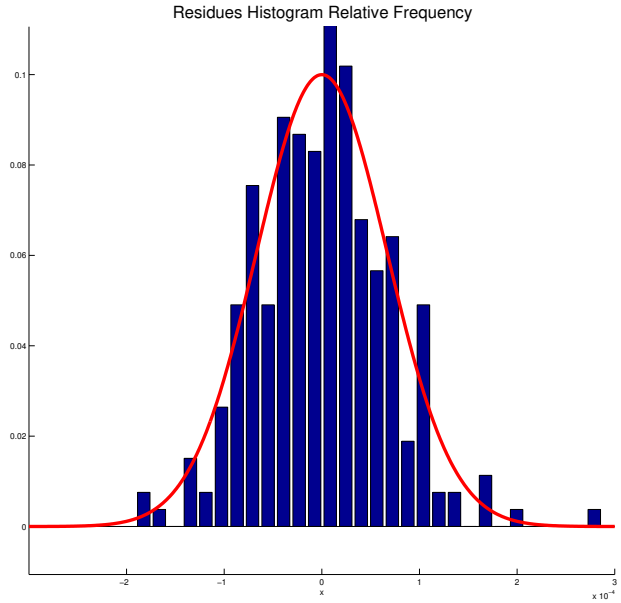
Introduction Stats Econométrie

Régression multivariée

marco.cuturi@ensae.fr

Approche Statistique pour la Régression Linéaire

- Propriétés de l'EMV: convergence de $\hat{\alpha}_{\text{EMV}}$ vers $\alpha = (\omega, \beta)$?
- Intervalles de Confiance pour les coefficients,
- Tests pour verifier que les hypothèse sur les résidus réalistes



- Approches bayésiennes: au lieu de ne considérer qu'**une** estimation du paramètre optimal $\theta = (\omega, \beta)$, on étudie une densité (a posteriori) des paramètres $\theta = (\mathbf{w}, b)$ à partir d'une distribution a priori et des données.

L'estimateur $\hat{\alpha}$ est sans biais.

$$
\begin{aligned}
\mathbb{E}[\,\hat{\boldsymbol{\alpha}}] &= \mathbb{E}\Big[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{Y})\Big] \\
&= \mathbb{E}\Big[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\alpha + \boldsymbol{\varepsilon})\Big] \\
&= \alpha + \mathbb{E}\Big[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}\Big] \\
&= \alpha
\end{aligned}
$$

Variance de l'estimateur $\hat{\alpha}$.

Notons que:

$$\begin{aligned}\hat{\alpha} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{Y}) \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\alpha + \boldsymbol{\varepsilon}) = \alpha + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}\end{aligned}$$

Ainsi,

$$\hat{\alpha} - \alpha = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}.$$

et

$$\begin{aligned}\mathbb{E}[\,(\hat{\boldsymbol{\alpha}} - \alpha)(\hat{\boldsymbol{\alpha}} - \alpha)^T] &= \mathbb{E}\Big[((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon})((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon})^T\Big] \\ &= \mathbb{E}\Big[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\Big] \\ &= \mathbb{E}\Big[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\sigma^2 I_d\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\Big] \\ &= \mathbb{E}\Big[\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\Big] \\ &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}.\end{aligned}$$

Distribution de $\hat{\alpha}$.

- Nous avons vu que

$$\mathbb{E}[\,\hat{\boldsymbol{\alpha}}] = \alpha$$
$$\mathbb{E}[\,(\hat{\boldsymbol{\alpha}} - \alpha)(\hat{\boldsymbol{\alpha}} - \alpha)^T] = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}.$$

- En fait... $\boxed{\hat{\alpha} = \alpha + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon} \sim \mathcal{N}(\alpha, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})}$

- En effet:

$$\text{Si } u \sim \mathcal{N}(0, \Sigma), \quad \text{alors } \mu + Lu \sim \mathcal{N}(\mu, L\Sigma L^T)$$

Estimation de σ.

- Les estimations de résidus sont

$$\begin{aligned}\hat{\boldsymbol{\varepsilon}} &= \mathbf{Y} - \mathbf{X}\hat{\alpha} \\ &= \mathbf{Y} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\ &= M\mathbf{Y}\end{aligned}$$

  en écrivant $M = I_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$

- Ainsi, l'estimateur des moindres carrés de la variance des résidus est

$$\begin{aligned}\hat{\boldsymbol{\sigma}}^2 &= \frac{1}{n-1}\mathbf{Y}^T M^T M\mathbf{Y} \\ &= \frac{1}{n}\mathsf{RSS}(\hat{\alpha})\end{aligned},$$

  où $\mathsf{RSS}(\alpha) = \|\mathbf{Y} - \mathbf{X}^T\alpha\|^2$.

- Cet estimateur est biaisé. Le bon estimateur tient compte des degrés de liberté

$$\boldsymbol{s}^2 = \frac{\mathsf{RSS}(\hat{\alpha})}{n-d}$$

Distribution de $\hat{\sigma}$.

$$\hat{\boldsymbol{\sigma}}^{\mathbf{2}} = \frac{1}{n-1}\mathbf{Y}^T M^T M \mathbf{Y}$$

- Plus précisément,

$$\begin{aligned}\hat{\boldsymbol{\sigma}}^{\mathbf{2}} &= \frac{1}{n-1}(\mathbf{X}\alpha + \boldsymbol{\varepsilon})^T M^T M (\mathbf{X}\alpha + \boldsymbol{\varepsilon}) \\ &= \frac{1}{n-1}\boldsymbol{\varepsilon}^T M \boldsymbol{\varepsilon}\end{aligned}$$

- Et ainsi,

$$\frac{(n-1)}{\sigma^2}\hat{\boldsymbol{\sigma}}^{\mathbf{2}} = \frac{\boldsymbol{\varepsilon}}{\sigma}^T M \frac{\boldsymbol{\varepsilon}}{\sigma}$$

- Que peut on dire de la distribution de $\frac{n-1}{\sigma^2}\hat{\boldsymbol{\sigma}}^{\mathbf{2}}$?

Distribution de $\hat{\sigma}$.

- $M = I_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ est un **projecteur**
- En effet $M^2 = (I_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^2 = I_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$
- M est une matrice symmétrique, à valeurs propres réelles.
- Si (λ, e) est un couple (valeur, vecteur) propre,

$$M^2 e = Me, \text{ et donc } M(\lambda e) = \lambda^2 e = \lambda e$$

  et donc λ est soit 0 ou 1.
- et donc $M = U^T \Delta U$ où $UU^T = I_n$ et Δ diagonale avec $(\text{rank}M)$ valeurs propres à 1 et le reste à 0.
- Ainsi, $\frac{\boldsymbol{\varepsilon}}{\sigma}^T M \frac{\boldsymbol{\varepsilon}}{\sigma} = \left(U\frac{\boldsymbol{\varepsilon}}{\sigma}\right)^T \Delta \left(U\frac{\boldsymbol{\varepsilon}}{\sigma}\right)$
- Si $Z \sim \mathcal{N}(0, I_n)$ et $U^TU = I_n$ alors $UZ \sim \mathcal{N}(0, I_n)$.
- Ainsi $\left(U\frac{\boldsymbol{\varepsilon}}{\sigma}\right) \sim \mathcal{N}(0, I_n)$
- Et donc

$$\frac{n-1}{\sigma^2}\hat{\boldsymbol{\sigma}}^2 = \frac{\boldsymbol{\varepsilon}}{\sigma}^T M \frac{\boldsymbol{\varepsilon}}{\sigma} = \sum_{i=1}^{\text{rank}(M)} \left(\frac{\varepsilon_i}{\sigma}\right)^2 \sim \chi^2_{\text{rank}M}.$$

Indépendance de $\hat{\alpha}$ & $\hat{\sigma}$ conditionellement à $\mathbf{X}$

- En résumé:

$$\hat{\boldsymbol{\alpha}} = \alpha + P\boldsymbol{\varepsilon}, P = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$
$$\hat{\boldsymbol{\sigma}}^{\mathbf{2}} = \|M\boldsymbol{\varepsilon}\|^2, M = I_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

- Notons que P and M sont orthogonaux:

$$\begin{aligned} PM &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(I_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T) \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = 0 = MP^T \end{aligned}$$

- Ainsi $P\varepsilon$ et $M\varepsilon$ sont des vecteurs alatoires $\mathbb{R}^n$ décorrélés, Gaussiens, et donc indépendants.

- Il est par la suite possible de définir des intervalles de confiance pour chacun des coefficients:

$$\alpha_j \in \left[\hat{\alpha}_j \pm q^{\mathcal{N}(0,1)}_{1-\frac{\alpha}{2}} \sqrt{\hat{\sigma}^2 \left[(\mathbf{X}^T\mathbf{X})^{-1}\right]_{jj}}\right]$$

  avec probabilité $1 - \alpha$.

Techniques de Régression Avancées

Rappel sur les normes vectorielles

- Pour un vecteur $\mathbf{a} \in \mathbb{R}^d$, la norme Euclidienne est

$$\|\mathbf{a}\|_2 = \sqrt{\sum_{i=1}^{d} a_i^2}.$$

- Plus généralement $q > 0$ (seules les valeurs $q \geq 1$ sont des normes)

$$\|\mathbf{a}\|_q = \left(\sum_{i=1}^{d} |a_i|^q \right)^{\frac{1}{q}}.$$

- En particulier, pour $q = 1$,

$$\|\mathbf{a}\|_1 = \sum_{i=1}^{d} |a_i|$$

- Quet $q \to \infty$ ou $q \to 0$,

$$\|\mathbf{a}\|_\infty = \max_{i=1,\cdots,d} |a_i|. \quad \|\mathbf{a}\|_0 = \#\{i | a_i \neq 0\}.$$

Régularization de Tikhonov '43 - Régression Ridge '62

- Motivation de Tikhonov : résoudre des **problèmes inverses mal posés** grâce à la **régularisation**
- Si $\min_\alpha L(\alpha)$ est minimisé sur plusieurs points... considérons plutôt

$$\min_\alpha L(\alpha) + \lambda \|\alpha\|_2^2$$

- On peut montrer que cette approche aboutit à un estimateur

$$\hat{\alpha} = (\mathbf{X}^T\mathbf{X} + \boldsymbol{\lambda I_{d+1}})^{-1}\mathbf{X}\mathbf{Y}$$

- Le conditionnement de la matrice devient maintenant

$$\frac{\lambda_{\max}(\mathbf{X}^T\mathbf{X}) + \lambda}{\lambda_{\min}(\mathbf{X}^T\mathbf{X}) + \lambda}.$$

Sélection de Variables: Recherche Exhaustive

- Selon le rasoir d'Ockham, nous voudrions pouvoir évaluer pour un p donné

$$\min_{\alpha, \|\alpha\|_0 = p} L(\alpha)$$

- $\rightarrow$ selectionner le **meilleur** vecteur α qui décrit des coefficients non-nuls **exclusivement** sur p **variables**.
- $\rightarrow$ Trouver la **meilleure** combinaison de p variables.

En pratique

- Avec $p \leq n$, $\binom{n}{p}$ combinaisons possibles de p variables.
- Approche gloutonne: générer $\binom{n}{p}$ problèmes de régression, séléctionner celui qui aboutit à la meilleure somme résiduelle de carrés.

Prohibitif pour un nombre modéré de variables n et p... $\binom{30}{5} = 150.000$

Sélection de variables : Approche Forward

Puisque la recherche **exacte** est **prohibitif en pratique**, utiliser une heuristique **"forward"**.

- **Recherche Forward**:
  - Définir $I_1 = \{0\}$.
  - étant donné un ensemble $I_k \subset \{0, \cdots, d\}$ de k variables, **quelle est la variable maximalement informative qui pourrait être ajoutée?**
    - ▷ Calculer pour chaque variable i dans $\{0, \cdots, d\} \setminus I_k$

$$t_{\boldsymbol{i}} = \min_{(\alpha_k)_{k \in I_k}, \boldsymbol{\alpha}} \sum_{j=1}^{N} \left(y_j - \left(\sum_{k \in I_k} \alpha_k x_{k,j} + \boldsymbol{\alpha} x_{\boldsymbol{i},j} \right) \right)^2$$

    - ▷ Actualiser $I_{k+1} = I_k \cup \{i^\star\}$, où $i^\star$ est tel que $i^\star = \mathbf{min}\ t_i$.
    - ▷ $k = k + 1$ jusqu'à obtenir le nombre voulu de variables.

Sélection de variables: Heuristique Backward

... ou l'heuristique **backward**.

- **Recherche Backward**:
  - Définir $I_d = \{0, 1, \cdots, n\}$.
  - étant donné un ensemble $I_k \subset \{0, \cdots, d\}$ de k variables, **quelle est la variable minimalement informative qui pourrait être enlevée?**
    - ▷ Calculer pour chaque variable i in I_k

$$t_{\boldsymbol{i}} = \min_{(\alpha_k)_{k \in I_k \setminus \{\boldsymbol{i}\}}} \sum_{j=1}^{N} \left\| y_j - \left(\sum_{k \in I_k \setminus \{\boldsymbol{i}\}} \alpha_k x_{k,j} \right) \right)^2$$

    - ▷ Actualiser $I_{k-1} = I_k \setminus \{i^\star\}$ où $i^\star$ est tel que $i^\star = \mathbf{max}\ t_i$.
    - ▷ $k = k - 1$ jusqu'à obtenir le nombre voulu de variables.

Comment choisir le nombre voulu de variables?

Trois approches classiques:

- Critère d'information d'Akaike
- Critère d'information Bayésien
- Validation croisée

Akaike Information Criterion ('72)

$$\min_{\alpha_k} k - \log \mathcal{L}(\alpha_k)$$

- Utiliser une approche forward, ajouter progressivement des variables.
- Arrêter la sélection quand la somme des résidus + nombre de paramètres remonte.

AIC Corrigé: $\min_{\alpha_k} L(\alpha_k) + k + \frac{k(k+1)}{n-k-1}$

Bayesian Information Criterion ('78)

$$\min_{\alpha_k} \frac{\log n}{2} k - \log \mathcal{L}(\alpha_k)$$

- Utiliser une approche forward, ajouter progressivement des variables.
- Arrêter la sélection quand la somme des résidus + nombre de paramètres remonte.
- N'a d'intérêt que quand $n \gg k$.

Validation Croisée

- Scinder l'ensemble de données en

$$\mathbf{X}_1 \in \mathbb{R}^{n_1 \times d}, \mathbf{X}_2 \in \mathbb{R}^{n_2 \times d},$$

$$\mathbf{Y}_1 \in \mathbb{R}^{n_1}, \mathbf{Y}_2 \in \mathbb{R}^{n_2}.$$

- Utiliser une approche forward en utilisant

$$L_1(\alpha) = \|\mathbf{Y}_1 - \mathbf{X}_\mathbf{1}^\mathbf{T}\alpha\|^2,$$

  comme critère. Ajouter progressivement des variables pour former $\hat{\alpha}_k$.

- En définissant la SRC sur l'autre échantillon

$$L_2(\alpha) := \|\mathbf{Y}_1 - \mathbf{X}_\mathbf{1}^\mathbf{T}\alpha\|^2,$$

  arrêter d'ajouter des variables dès que cette somme remonte, soit dès que $L_2(\hat{\alpha}_{k+1}) \geq L_2(\hat{\alpha}_k)$.

Ouverture: Approches Régularisées

Les techniques décrites ci-dessous sont vues comme plus modernes, et impliquent l'utilisation d'une régularisation du problème des moindres carrés. Elles seront plus amplement détaillées en 2A et 3A.

Moindre Carrés Naïf'

$$\min_{\alpha} L(\alpha)$$

Meilleure combinaison de p variables (Occam!)

$$\min_{\alpha, \|\boldsymbol{\alpha}\|_0 = p} L(\alpha)$$

Régularisation de Tikhonov

$$\min_{\alpha} L(\alpha) + \boldsymbol{\lambda}\|\boldsymbol{\alpha}\|_2^2$$

LASSO (least absolute shrinkage et selection operator)

$$\min_{\alpha} L(\alpha) + \boldsymbol{\lambda}\|\boldsymbol{\alpha}\|_1$$

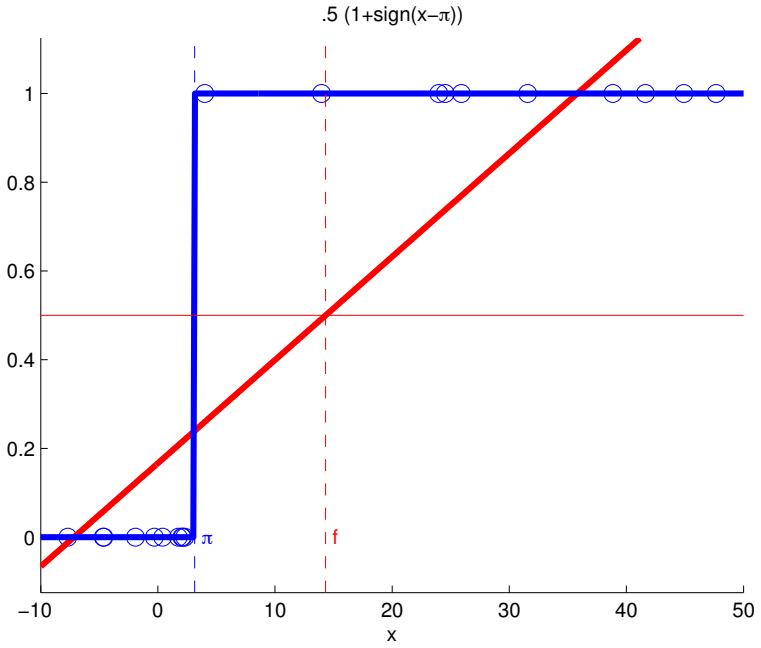
Extension:
Régression Logistique

Quet la régression de moindre carrés ne marche pas

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

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

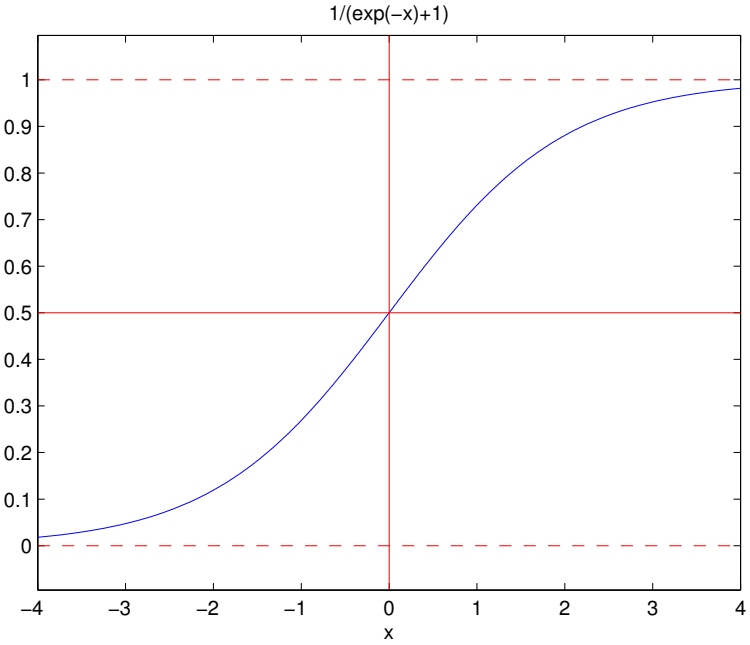
- Que se passe-t-il si nous utilisons ces données directement l'estimateur des moindres carrés?... **demo**



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

Interpretation 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 $y = 1$.

Interpretation 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.

Interpretation 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,

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

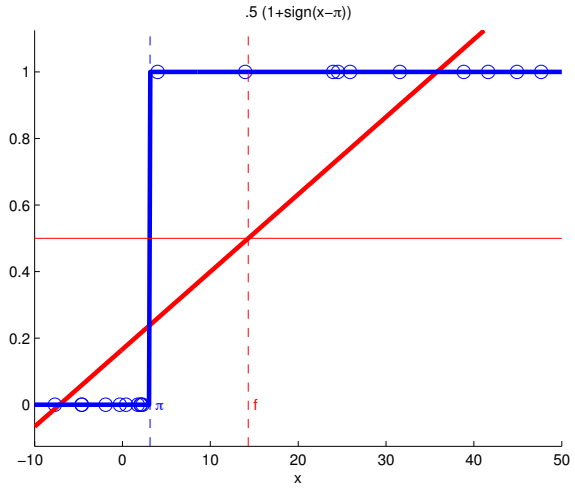
- Maximiser cette log-vraisemblance est équivalent à

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

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

Estimation de c et b par MV

Comparons...



...avec