Cours de Scoring - Séance 2

Ibrahim TOURE,
Ingénieur Statisticien

Université d'Evry Val d'Essone

1er Décembre 2016

Sommaire

Modèle

Qu'est ce qu'un modèle ?

Mathématiquement, un modèle est un triplet $(\mathcal{H}, \mathcal{A}, \{P, P \in \mathcal{P}\})$

- $\mathcal{H}$ est l'espace des observations (l'ensemble de tous les résultats possibles de l'expérience)
- $\mathcal{A}$ est une tribu de $\mathcal{H}$
- $\mathcal{P}$ est une famille de probabilités définie sur $(\mathcal{H}, \mathcal{A})$

A quoi sert un modèle ?

Expliquer, décrire les mécanismes du phénomène considéré.

- Question : quel est le lien entre la définition mathématique et l'utilité du phénomène ?

Modèle de densité (1/10)

Exemple 1

- On souhaite tester l'efficacité d'un nouveau traitement à l'aide d'un essai clinique
- On traite n=100 patients atteints de la pathologie
- A l'issue de l'étude, 72 patients sont guéris
- Soit p_0 la probabilité de guérison suite au traitement en question
- On est tenté de conclure que $p_0 \approx 0.72$

Un tel résultat n'a cependant guère d'intérêt si on n'est pas capable de préciser l'erreur susceptible d'être commise par cette estimation.

Modèle de densité (2/10)

Exemple 2

- On s'intéresse au nombre de mails reçus par jour par un utilisateur pendant 36 journées
- $\bar{x} = 5.22$ et $S_n^2 = 5.72$

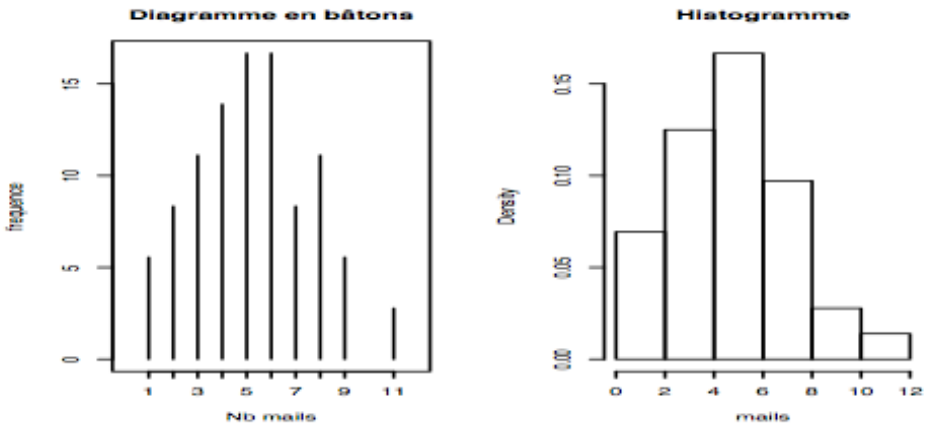


Figure : Histogramme issu de cet exemple

Modèle de densité (3/10)

Exemple 3

- Durée de trajet domicile-travail
- On dispose de n=100 mesures : $\bar{x} = 25.1$, $S_n^2 = 14.46$

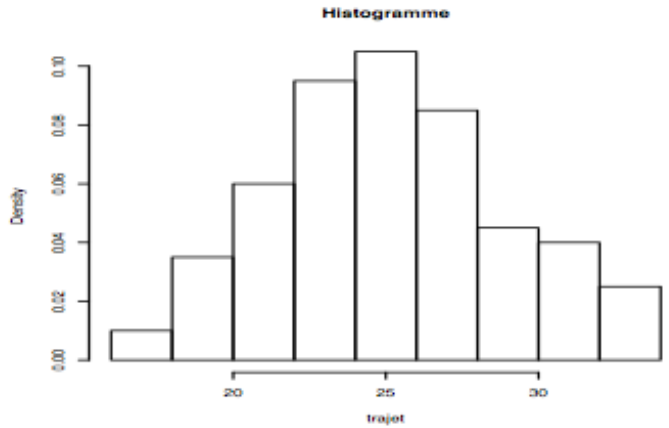


Figure : Histogramme issu de cet exemple

J'ai une réunion à 8h30, quelle est la probabilité que j'arrive en retard si je pars de chez moi à 7h55.

Modèle de densité (4/10)

Problème

- Nécessité de se dégager des observations $x_1, ..., x_n$ pour répondre à de telles questions
- Si on mesure la durée du trajet pendant 100 nouveaux jours, on peut en effet penser que les nouvelles observations ne seront pas exactement les mêmes que les anciennes

Idée Considérer que les *n* valeurs observées $x_1, ..., x_n$ sont des réalisations de variables aléatoires $X_1, ..., X_n$.

Attention

X_i est une variable aléatoire et x_i est une réalisation de cette variable, c'est à dire un nombre !

Modèle de densité (5/10)

Variables aléatoires Une variable aléatoire est une application $X : (\Omega, \mathcal{A}) \to (\mathbb{R}, (B)(\mathbb{R}))$ telle que $\forall \mathcal{B} \in \mathcal{B}(\mathbb{R}, X^{-1}(B) \in \mathcal{A}$

- Lors de la modélisation statistique, l'espace Ω n'est généralement jamais caractérisé
- Il contient tous les phénomènes pouvant expliquer les sources d'aléa (qui ne sont pas explicables...)
- En pratique, l'espace d'arrivée est généralement suffisant

Modèle de densité (6/10)

Loi de probabilité Etant donné **P** une probabilité sur $(\Omega, \mathcal{A})$ et X une variable aléatoire réelle définie sur Ω, on appelle loi de probabilité de X la mesure P_X définie par :
$P_X(B) = P(X^{-1}(A')) = P(X \in B) = P(\{\omega \in \Omega : X(\omega) \in B\}) \forall B \in \mathcal{B}(\mathbb{R})$.
Une loi est caractérisée par :

- sa fonction de répartition : $F_X(x) = P(X \leq x)$
- sa densité : $f_x : \mathbb{R} \to \mathbb{R}^+$ telle $\forall B \in \mathcal{B}(\mathbb{R})$
  $P_X(B) = \int_B f_X(x)dx$

Modèle de densité (7/10)

Un modèle pour l'exemple 1

- on note $x_i = 1$ si le i^{me} patient a guéri, 0 sinon
- On peut supposer que x_i est la réalisation d'une variable aléatoire X_i de loi de bernoulli de paramètre p_0
- Si les individus sont choisis de manière *indépendante* et ont tous la *même probabilité de guérir* (ce qui revient à dire qu'ils en sont au même stade de la pathologie), il est alors raisonnable de supposer que les variables aléatoires $X_1, ..., X_n$ sont indépendantes et de même loi (i.i.d)

On dit que $X_1, ..., X_n$ est un n-échantillon de variables aléatoires indépendantes de même loi $B\ (p_0)$.

Modèle de densité (8/10)

Qu'est ce qu'un modèle statistique (le problème du statisticien) ?

- en partant de *n* variables aléatoires i.i.d, $X_1, ..., X_n$ de loi **P**
- Il s'agit de trouver une famille de loi $\mathcal{P}$ susceptible de contenir **P**
- et que $\mathcal{P}$ soit une loi qui soit la *plus proche* de **P**

Modèle de densité (9/10)

Synthèse des 3 exemples mentionnés

	$\mathcal{H}$	$\mathcal{A}$	$\mathcal{P}$
Exemple 1	$\{0,1\}$	$\mathcal{P}(\{0,1\})$	$\{B(p), p \in [0,1]\}$
Exemple 2	$\mathbb{N}$	$\mathcal{P}(\mathbb{N})$	$\{\mathcal{P}(\lambda), \lambda > 0\}$
Exemple 3	$\mathbb{R}$	$\mathcal{B}(\mathbb{R})$	$\{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+\}$

Figure : synthèse des 3 exemples

Modèle de densité (10/10)

Modèle Paramétrique versus Non Paramétrique

Définition

- Si $\mathcal{P} = \{\mathbf{P}_\theta, \theta \in \oplus\}$ où $\oplus \in \mathcal{R}^d$ alors on parle de modèle paramétrique et $\oplus$ est l'espace des paramètres
- Si $\mathcal{P} = \{\mathbf{P}, \mathbf{P} \in \mathcal{F}\}$ où $\mathcal{F}$ est de dimension infinie, on parle alors de modèle non paramétrique

Exemple : modèle de densité

- $\mathcal{P}$ {suit une loi normale est un modèle paramétrique de paramètre (μ, σ) }
- $\mathcal{P}$ {densités f 2 fois dérivabes } est un modèle non paramétrique

Le problème sera d'estimer (μ, σ) ou f à partir de l'échantillon $X_1, ..., X_n$

Sommaire

Modèle de régression

- On cherche à expliquer une variable Y par p variables explicatives $\mathbf{X}_1, \ldots, \mathbf{X}_p$. On dispose d'un n échantillon i.i.d. $(X_i, Y_i), i = 1, \ldots, n$.

Modèle linéaire (paramétrique)

- On pose

$$Y = \beta_0 + \beta_1 \mathbf{X}_1 + \ldots + \beta_p \mathbf{X}_p + \varepsilon \quad \text{où} \quad \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

- Le problème est d'estimer $\beta = (\beta_0, \ldots, \beta_p) \in \mathbb{R}^{p+1}$ à l'aide de $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Un modèle non paramétrique

- On pose

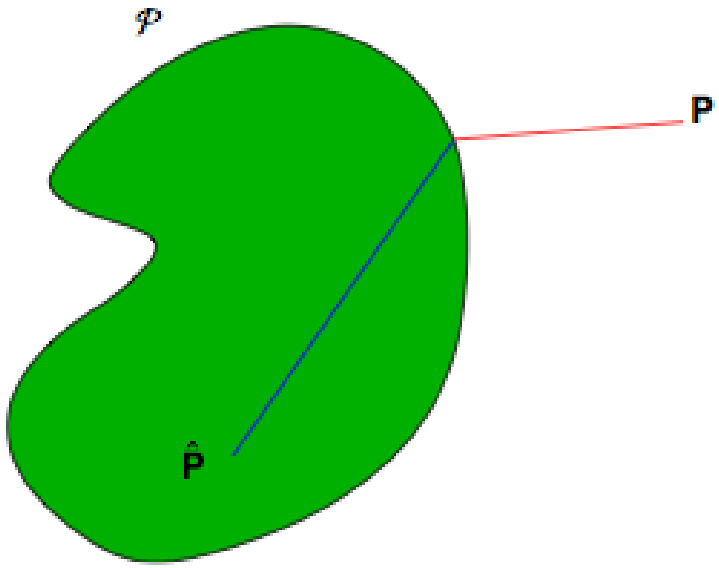
$$Y = m(\mathbf{X}_1, \ldots, \mathbf{X}_p) + \varepsilon$$

où $m : \mathbb{R}^p \to \mathbb{R}$ est une fonction continue.

- Le problème est d'estimer m à l'aide de $(X_1, Y_1), \ldots, (X_n, Y_n)$.

2 types d'erreur

- Poser un modèle revient à choisir une famille de loi candidates pour reconstruire la loi des données **P**.



On distingue deux types d'erreurs :

- Erreur d'estimation : erreur commise par le choix d'une loi dans $\mathcal{P}$ par rapport au meilleur choix.
- Erreur d'approximation : erreur commise par le choix de $\mathcal{P}$.

La modélisation

1. On récolte n observations (n valeurs) $x_1, \ldots, x_n$ qui sont le résultats de n expériences aléatoires indépendantes.

2. **Modélisation :** on suppose que les n valeurs sont des réalisations de n variables aléatoires indépendantes $X_1, \ldots, X_n$ et de même loi $\mathbf{P}_{\theta_0}$.

3. **Estimation :** chercher dans le modèle une loi $\mathbf{P}_{\hat{\theta}}$ qui soit le plus proche possible de $\mathbf{P}_{\theta_0} \Longrightarrow$ chercher un **estimateur** $\hat{\theta}$ de θ_0.

4. **"Validation" de modèle** : on revient en arrière et on tente de vérifier si l'hypothèse de l'étape 2 est raisonnable (test d'adéquation, etc...)

Figure :

Rappels sur le modèle de régression

- On cherche à expliquer une variable Y par p variables $\mathbf{X}_1, \ldots, \mathbf{X}_p$.
- Il s'agit de trouver une fonction $m : \mathbb{R}^p \to \mathbb{R}$ telle que $Y \approx m(\mathbf{X}_1, \ldots, \mathbf{X}_p)$.
- Sauf cas (très) particulier, le lien n'est jamais "parfait"

$$Y = m(\mathbf{X}_1, \ldots, \mathbf{X}_p) + \varepsilon.$$

Modèle de régression

- Poser un modèle de régression revient à supposer que la fonction m appartient à un certain espace $\mathcal{M}$.
- Le problème du statisticien sera alors de trouver la "meilleure" fonction dans $\mathcal{M}$ à l'aide d'un n-échantillon i.i.d. $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Figure :

Exemples de modèles

Modèle non paramétrique

- L'espace $\mathcal{M}$ est de dimension infinie.
- **Exemple** : On pose $Y = m(\mathbf{X}_1, \ldots, \mathbf{X}_p) + \varepsilon$ où m appartient à l'espace des fonctions continues.

Modèle paramétrique

- L'espace $\mathcal{M}$ est de dimension finie.
- **Exemple** : on suppose que la fonction m est linéaire

$$Y = \beta_0 + \beta_1 \mathbf{X}_1 + \ldots + \beta_p \mathbf{X}_p + \varepsilon.$$

Le problème est alors d'estimer $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ à l'aide de $(X_1, Y_1), \ldots, (X_n, Y_n)$.

- C'est le modèle de **régression linéaire**.

Figure :

Notations

- Y : variable (aléatoire) à expliquer à valeurs dans $\mathbb{R}$.
- $X_1, \ldots, X_p$: p variables explicatives à valeurs dans $\mathbb{R}$.
- n observations $(x_1, Y_1), \ldots, (x_n, Y_n)$ avec $x_i = (x_{i1}, \ldots, x_{ip})$.

Le modèle de régression linéaire multiple

Le modèle s'écrit :

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i$$

où les erreurs aléatoires ε_i sont i.i.d. de loi $\mathcal{N}(0, \sigma^2)$.

Figure :

Ecriture matricielle

- On note

$$\mathbb{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbb{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Ecriture matricielle

Le modèle se réécrit

$$\mathbb{Y} = \mathbb{X}\beta + \varepsilon$$

où $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I_n})$.

Figure :

Estimateurs des moindres carrés

Définition

On appelle **estimateur des moindres carrés** $\hat{\beta}$ de β la statistique suivante :

$$\hat{\beta} = \underset{\beta_0,\dots,\beta_p}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_{i1} - \dots \beta_p x_{ip})^2 = \underset{\beta \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \|\mathbb{Y} - \mathbb{X}\beta\|^2.$$

- On note $\mathcal{F}(\mathbb{X})$ le s.e.v. de $\mathbb{R}^n$ de dimension $p+1$ engendré par les $p+1$ colonnes de $\mathbb{X}$.
- Chercher l'estimateur des moindres carrés revient à minimiser la distance entre $\mathbb{Y} \in \mathbb{R}^n$ et $\mathcal{F}(\mathbb{X})$.

Figure :

Réprésentation géométrique

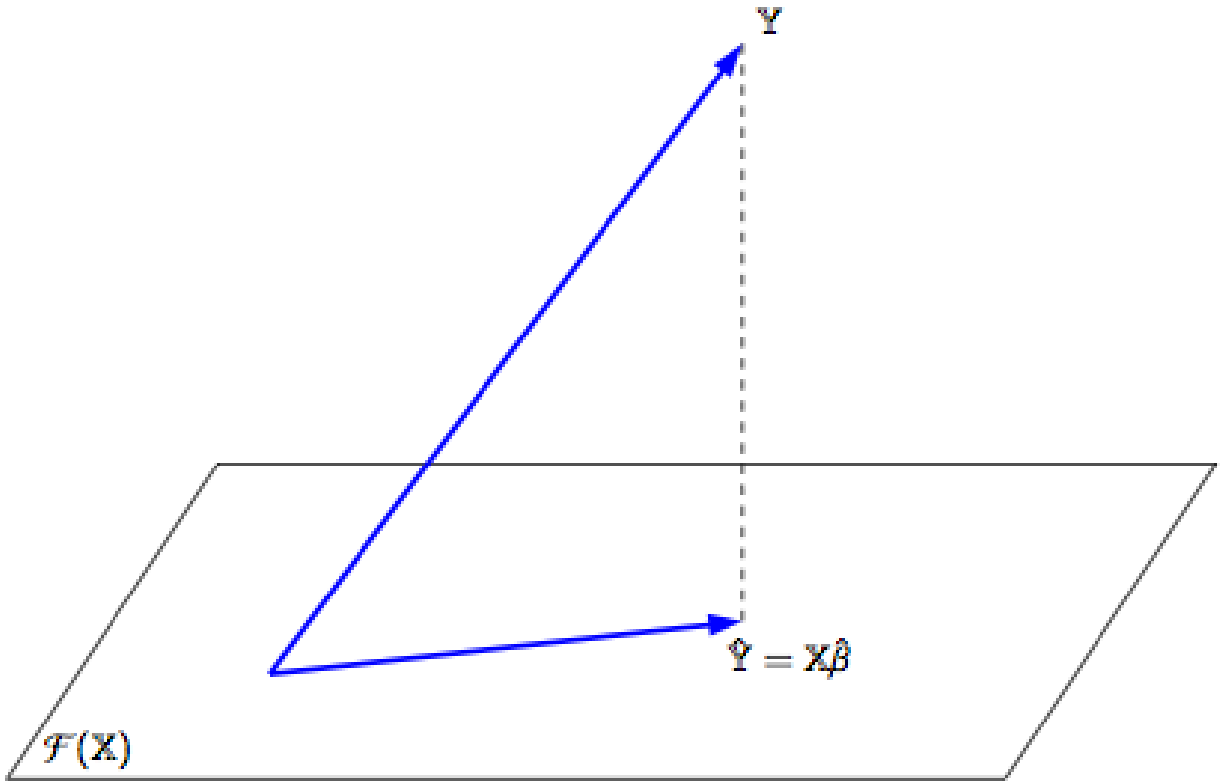


Figure :

Expréssion de l'estimateur des moindres carrés

- On déduit que $\mathbb{X}\hat{\beta}$ est le projeté orthogonal de $\mathbb{Y}$ sur $\mathcal{F}(\mathbb{X})$:

$$\mathbb{X}\hat{\beta} = \mathbf{P}_{\mathcal{F}(\mathbb{X})}(\mathbb{Y}) = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{Y}.$$

Théorème

Si la matrice $\mathbb{X}$ est de plein rang, l'estimateur des MC est donné par :

$$\hat{\beta} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{Y}.$$

Figure :

Propriété

Propriété

1. $\hat{\beta}$ est un estimateur sans biais de β.
2. La matrice de variance-covariance de $\hat{\beta}$ est donnée par

$$\mathbf{V}(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

3. $\hat{\beta}$ est VUMSB.

Loi des estimateurs

- Soit $\hat{\varepsilon} = \mathbb{Y} - \hat{\mathbb{Y}}$ le vecteur des résidus et $\widehat{\sigma^2}$ l'estimateur de σ^2 défini par

$$\widehat{\sigma^2} = \frac{\|\hat{\varepsilon}\|^2}{n-(p+1)}.$$

Proposition

1. $\hat{\beta}$ est un vecteur gaussien d'espérance β et de matrice de variance-covariance $\sigma^2(\mathbb{X}'\mathbb{X})^{-1}$.
2. $(n-(p+1))\frac{\widehat{\sigma^2}}{\sigma^2} \sim \chi^2_{n-(p+1)}$.
3. $\hat{\beta}$ et $\widehat{\sigma^2}$ sont indépendantes.

Figure :

Intervalle de confiance et tests

Corollaire

On note $\widehat{\sigma_j^2} = \widehat{\sigma^2}[\mathbb{X}'\mathbb{X}]^{-1}_{jj}$ pour $j = 0, \ldots, p$. On a

$$\forall j = 0, \ldots, p, \quad \frac{\hat{\beta}_j - \beta_j}{\widehat{\sigma}_j} \sim \mathcal{T}(n - (p+1)).$$

On déduit de ce corollaire :

- des intervalles de confiance de niveau $1 - \alpha$ pour β_j.
- des procédures de test pour des hypothèses du genre $H_0 : \beta_j = 0$ contre $H_1 : \beta_j \neq 0$.

Figure :

Prévision

- On dispose d'une nouvelle observation $x_{n+1} = (x_{n+1,1}, \ldots, x_{n+1,p})$ et on souhaite prédire la valeur $y_{n+1} = x'_{n+1}\beta$ associée à cette nouvelle observation.

- Un estimateur (naturel) de y_{n+1} est $\hat{y}_{n+1} = x'_{n+1}\hat{\beta}$.
- Un intervalle de confiance de niveau $1-\alpha$ pour y_{n+1} est donné par

$$\left[\hat{y}_{n+1} \pm t_{n-(p+1)}(\alpha/2)\hat{\sigma}\sqrt{x'_{n+1}(\mathbb{X}'\mathbb{X})^{-1}x_{n+1} + 1}\right].$$

Figure :

Validation du modèle

- Ajuster un modèle, trouver des estimateurs est un problème relativement "simple".
- Le travail difficile est de trouver un bon modèle, ou encore le meilleur modèle (ce travail est difficile car la notion de meilleur modèle n'existe pas).
- Il est donc nécessaire de trouver des procédures automatiques de choix de modèles (méthodes pas à pas utilisant un critère de type AIC, BIC, régression lasso etc...)
- Puis de vérifier que les hypothèses effectuées (normalité, linéarité) sont raisonnables (analyse des résidus, tests d'adéquation...).

Figure :

Une autre écriture du modèle (1/2)

- Le modèle linéaire

$$Y_i = x_i'\beta + \varepsilon_i, \quad \varepsilon_i \text{ i.i.d de loi } \mathcal{N}(0, \sigma^2)$$

- peut se réécrire pour $i = 1, \ldots, n$

$$\mathcal{L}(Y_i) = \mathcal{N}(x_i'\beta, \sigma^2).$$

Interprétation

Au point x_i la loi de Y est une gaussienne $\mathcal{N}(x_i'\beta, \sigma^2)$.

Figure :

Une autre écriture du modèle (2/2)

- On peut alors calculer la (log)-vraisemblance du modèle

$$\mathcal{L}(y_1, \ldots, y_n; \beta) = \frac{n}{2}\log(\sigma^2) - \frac{n}{2}\log(2\pi) - \frac{1}{2\sigma^2}\|\mathbb{Y} - \mathbb{X}\beta\|^2.$$

- **Conclusion :** l'estimateur du maximum de vraisemblance $\hat{\beta}_{MV}$ coïncide avec l'estimateur des moindres carrés $\hat{\beta}$.

Remarque

- Si les variable explicatives sont aléatoires, ce n'est plus la loi de Y_i qui est modélisée mais celle de Y_i sachant $X_i = x_i$

$$\mathcal{L}(Y_i|X_i = x_i) = \mathcal{N}(x_i'\beta, \sigma^2).$$

- Plus généralement, lorsque les variables explicatives sont supposées aléatoires (économétrie), poser un modèle de régression revient à "mettre" une famille de loi sur Y sachant $X = x$.

Figure :

Sommaire

Détection de clients à risque

- Une chaine de magasin a mis en place une carte de crédit.
- Elle dispose d'un historique de 145 clients dont 40 ont connu des défauts de paiement.
- Elle connait également d'autres caractéristiques de ces clients (sexe, taux d'enttement, revenus mensuels, dépenses effectuées sur certaines gammes de produit...)

Question

Comment prédire si un nouveau client connaîtra des défauts de paiement ?

Figure :

Iris de Fisher

- On a mesuré sur 150 iris de 3 espèces différentes (Setosa, Versicolor, Virginica) les quantités suivantes :
  - Longueur et largeur des pétales
  - Longueur et largeur des sépales

Question

Comment identifier l'espèce d'un iris à partir de ces 4 caractéristiques ?

Figure :

Détection de Spam

- Sur 4 601 mails, on a pu identifier 1813 spams.
- On a également mesuré sur chacun de ces mails la présence ou absence de 57 mots.

Question

Peut-on construire à partir de ces données une méthode de détection automatique de spam ?

Figure :

Pathologie concernant les artères coronaires

- Problème : étudier la présence d'une pathologie concernant les artères coronaires en fonction de l'âge des individus.
- Données : on dispose d'un échantillon de taille 100 sur lequel on a mesuré les variables :
  - chd qui vaut 1 si la pathologie est présente, 0 sinon ;
  - age qui correspond à l'âge de l'individu.

```
> artere[1:5,]
   age agrp chd
1.  20    1   0
2.  23    1   0
3.  24    1   0
4.  25    1   0
5.  25    1   1
```

Figure :

Représentation du problème

- Tous ces problèmes peuvent être appréhendés dans un contexte de régression : on cherche à expliquer une variable Y par d'autres variables $X_1, \ldots, X_p$:

Y	X
Défaut de paiement	caractéristiques du client
Espèce de l'iris	Longueur, largeur pétales et sépales
Spam	présence/absence de mots

- La variable à expliquer n'est plus quantitative mais qualitative.
- On parle de problème de discrimination ou classification supervisée.

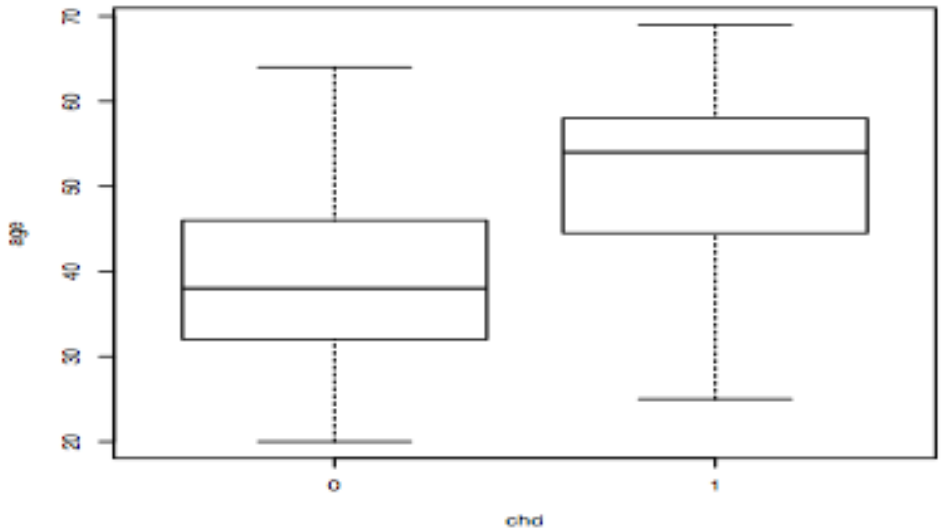
Figure :

Remarque : à l'opposée, on a les méthodes de classification non supervisée (CAH, K-Means, mixte)

Sommaire

Boxplot

```
> plot(age~chd,data=artere)
```



Il semble que la maladie a plus de chance d'être présente chez les personnes agées.

Figure :

Début de modélisation

Question

Comment expliquer la relation entre la maladie et l'âge ?

- On désigne par
  - Y la variable aléatoire qui prend pour valeur 1 si l'individu est atteint, 0 sinon.
  - X la variable (aléatoire) qui correspond à l'âge de l'individu.

Le problème consiste ainsi à tenter de quantifier la relation entre Y et X à partir des données, c'est-à-dire d'un échantillon i.i.d $(X_1, Y_1), \ldots, (X_n, Y_n)$ de taille $n = 100$.

Figure :

Première idée

- On se base sur le modèle linéaire.
- On suppose que les deux variables Y et X sont liées par une relation de la forme

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad (1)$$

où $\beta_0 \in \mathbb{R}$ et $\beta_1 \in \mathbb{R}$ sont les paramètres inconnus du modèle et ε est une variable aléatoire de loi $\mathcal{N}(0, \sigma^2)$.

Problème

La variable Y est ici **qualitative**, l'écriture (1) n'a donc aucun sens.
$\Longrightarrow$ mauvaise idée

Figure :

Remarque : cela revient en effet à dire qu'une gaussienne prend 2 valeurs (0 ou 1) !

Loi conditionnelle

- Chercher à expliquer Y par X revient à chercher de l'information sur la loi de probabilité de Y sachant X.
- En effet, le modèle de régression linéaire peut se réécrire en caractérisant la loi de $Y|X = x$ par la loi $\mathcal{N}(\beta_0 + \beta_1 x, \sigma^2)$.

Idée

- Etendre cette caractérisation à notre contexte (où la variable à expliquer est binaire).
- Une loi candidate naturelle pour la variable $Y|X = x$ est la loi de Bernoulli.

Figure :

Loi de Bernoulli

- On va ainsi caractériser la loi de $Y|X = x$ par la loi de Bernoulli.
- Cette loi dépend d'un paramètre

$$p(x) = \mathbf{P}(Y = 1|X = x).$$

- Sachant $X = x$, on a donc

$$Y = \begin{cases} 1 & \text{avec probabilité } p(x) \\ 0 & \text{avec probabilité } 1 - p(x) \end{cases}$$

La modélisation

Il reste maintenant à caractériser la probabilité $p(x)$.

Figure :

Première idée

- Là encore, on peut se baser sur le modèle linéaire et proposer

$$p(x) = \beta_0 + \beta_1 x.$$

- Cette écriture n'est pas satisfaisante. En effet
  - $p(x) \in [0, 1]$ tandis que $\beta_0 + \beta_1 x \in \mathbb{R}$.
  - Idée : trouver une transformation φ de $p(x)$ telle que $\varphi(p(x))$ prenne ses valeurs dans $\mathbb{R}$.

Figure :

Transformation de p(x)

- On revient sur l'exemple du chd et on représente les fréquences cumulées d'apparition de la maladie en fonction de l'âge :

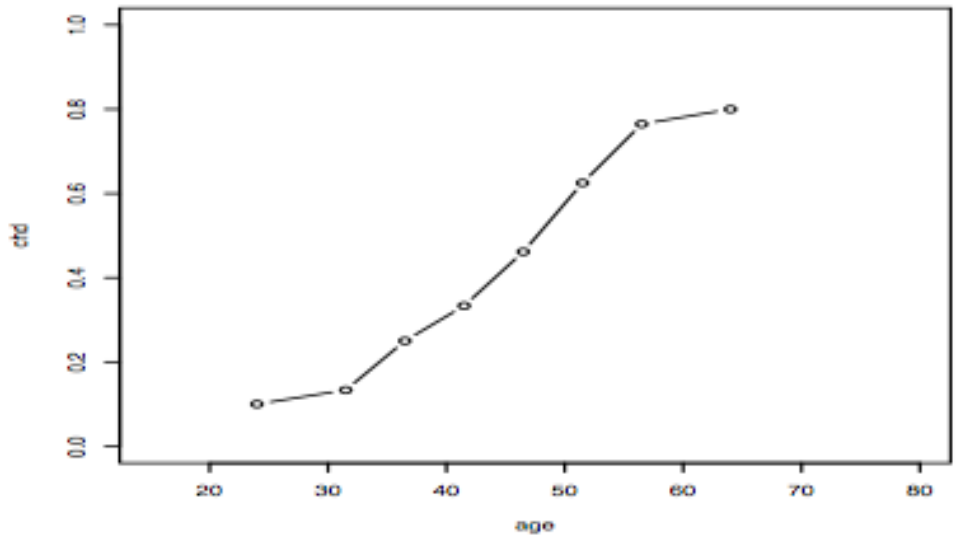
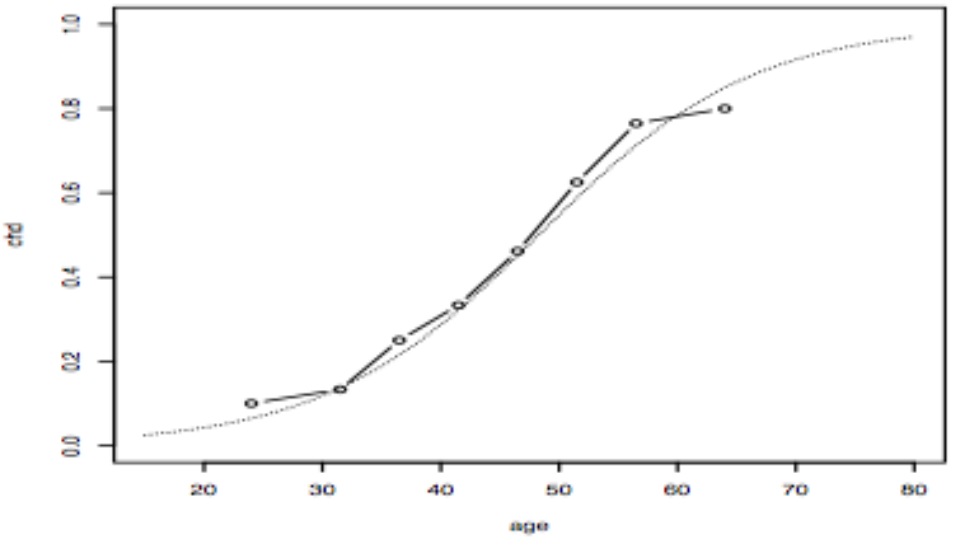


Figure :

Transformation de p(x)

- On revient sur l'exemple du chd et on représente les fréquences cumulées d'apparition de la maladie en fonction de l'âge :



Trouver une transformation de $p(x)$ qui ajuste ce nuage de points.

Figure :

Le modèle de régression logistique

- Il propose de modéliser la probabilité $p(x)$ selon

$$p(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}.$$

- On peut réécrire

$$\text{logit}\, p(x) = \log\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 x.$$

Le modèle de régression logistique

Le modèle de régression logistique consiste donc à caractériser la loi de $Y|X = x$ par une loi de Bernoulli de paramètre $p(x)$ tel que

$$\text{logit}\, p(x) = \log\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 x.$$

Figure :

frame

```
> model <- glm(chd~age,data=artere,family=binomial)
> model

Call:  glm(formula = chd ~ age, family = binomial, data = artere)

Coefficients:
(Intercept)          age
    -5.3095       0.1109

Degrees of Freedom: 99 Total (i.e. Null);  98 Residual
Null Deviance:      136.7
Residual Deviance: 107.4  AIC: 111.4
```

- La fonction **glm** renvoie les estimations de β_0 et β_1.
- On peut ainsi avoir une estimation de la probabilité d'avoir une maladie pour un individu de 30 ans :

$$\hat{p}(x = 30) = \frac{\exp(-5.3095 + 0.1109 * 30)}{1 + \exp(-5.3095 + 0.1109 * 30)} \approx 0.12.$$

Figure :

Sommaire

Introduction

- Le modèle de régression logistique s'ajuste sur R avec la fonction **glm**.
- Le modèle de régression logistique appartient à la famille des modèles linéaires généralisés.
- C'est pourquoi il faut spécifier l'argument `family=binomial` lorsque l'on veut faire une régression logistique.

Figure :

Le modèle linéaire est un GLM

- Le modèle de régression linéaire s'ajuste sur R avec la fonction **lm** :

```
> Y <- rnorm(50)
> X <- runif(50)
> lm(Y~X)

Coefficients:
(Intercept)            X
     0.4245      -0.8547
```

- Mais aussi avec la fonction **glm** :

```
> glm(Y~X,family=gaussian)

Coefficients:
(Intercept)            X
     0.4245      -0.8547
```

Conclusion

Le modèle linéaire appartient également à la famille des modèles linéaires généralisés.

Figure :

2 étapes identiques

- Les modèles linéaires et logistiques sont construits selon le même protocole en 2 étapes :

1. Choix de la loi conditionnelle de $Y|X = x$:
   - Gaussienne pour le modèle linéaire ;
   - Bernoulli pour le modèle logistique.
2. Choix d'une transformation g de l'espérance conditionnelle $\mathbf{E}[Y|X = x]$:
   - Logistique

$$g(\mathbf{E}[Y|X = x]) = g(p(x)) = \text{logit}\, p(x) = x'\beta$$

   - Linéaire

$$g(\mathbf{E}[Y|X = x]) = x'\beta.$$

Figure :

Définitions

Définition

Une loi de probabilité **P** appartient à une famille de lois de type exponentielle $\{\mathbf{P}_\theta\}_{\theta\in\mathbb{R}^p}$ si il existe une mesure dominant μ (Lebesgue ou mesure de comptage le plus souvent) telle que les lois $\mathbf{P}_\theta$ admettent pour densité par rapport à ν

$$f_\theta(y) = c(\theta)h(y)\exp\left(\sum_{j=1}^{p}\alpha_j(\theta)T_j(y)\right)$$

où $T_1, \ldots, T_p$ sont des fonctions réelles mesurables.

Exemple : loi de Bernoulli

La loi de Bernoulli de paramètre p admet pour densité (par rapport à la mesure de comptage)

$$f_p(y) = (1-p)\exp(y\log(p/(1-p))).$$

Figure :

Cadre

- On se place dans un contexte de régression : on cherche à expliquer une variable Y par p variables explicatives $X_1, \ldots, X_p$.
- On dispose d'un n-échantillon $(x_1, Y_1), \ldots, (x_n, Y_n)$ où les $x_i = (x_{i1}, \ldots, x_{ip})$ sont supposées fixes et les Y_i sont des variables aléatoires réelles indépendantes.

Figure :

Modèle linéaire généralisé : GLM

Un modèle linéaire généralisé est constitué de 3 composantes :

1. Composante aléatoire : la loi de probabilité de la réponse Y_i appartient à la famille exponentielle et est de la forme

$$f_{\alpha_i}(y_i) = \exp\left(\frac{\alpha_i y_i - b(\alpha_i)}{a(\phi)} + c(y_i, \phi)\right)$$

où a, b et c sont des fonctions spécifiées en fonction du type de la famille exponentielle.

2. Composante déterministe :

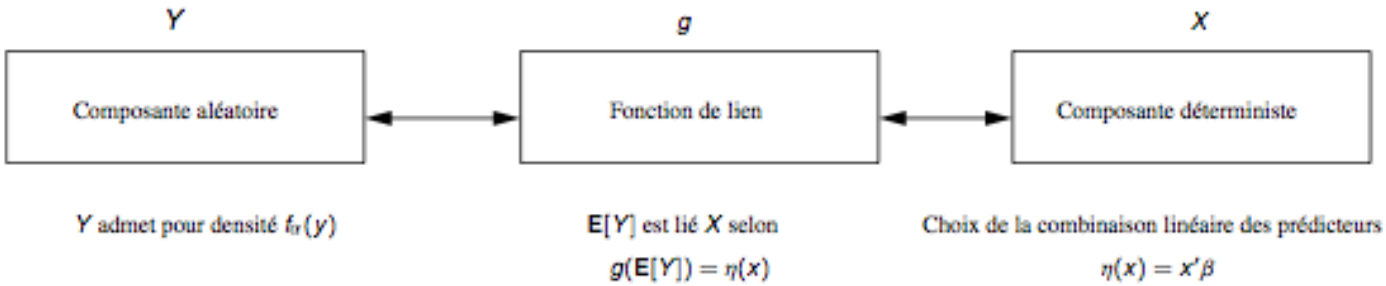
$$\eta(x_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

et précise quels sont les prédicteurs (on peut y inclure des transformations des prédicteurs, des interactions...).

3. Lien : spécifie le lien entre les deux composantes, plus précisément le lien entre l'espérance de Y_i et la composante déterministe : $g(\mathbf{E}[Y_i]) = \eta(x_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$ où g est une fonction inversible appelée **fonction de lien**.

Figure :

Schéma GLM



Un modèle GLM sera caractérisé par le choix de ces trois composantes.

Figure :

Composante déterministe

- Le problème du choix de la combinaison linéaire des variables explicatives est similaire à tous ce qui a été vu dans le modèle linéaire :
  - Utilisation d'indicatrices pour des variables explicatives qualitatives (sans oublier les contraintes d'identifiabilité).
  - Possibilité de prendre en compte des effets quadratique, ou autre transformation des variables explicatives.
  - Possibilité de prendre en compte des interactions.
  - Méthode de sélection de variables (stepwise, lasso...)

Dans la suite, on notera $\eta(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$ la combinaison linéaire choisie.

Figure :

Composante aléatoire et fonction de lien du modèle logistique

Propriété

Le modèle de régression logistique est un GLM.

En effet :

- La loi exponentielle est la loi de Bernoulli de paramètre $p_i = \mathbf{P}(Y_i = 1)$:

$$f_{\alpha_i}(y_i) = \exp[y_i x_i'\beta - \log(1 + \exp(x_i'\beta))].$$

On a donc $\alpha_i = x_i'\beta$ et $b(\alpha_i) = \log(1 + \alpha_i)$.

- La fonction de lien est

$$g(u) = \text{logit}(u) = \log \frac{u}{1-u}.$$

Figure :

Composante aléatoire et fonction de lien du modèle linéaire

Propriété

Le modèle linéaire gaussien est un GLM.

En effet :

- La loi exponentielle est la loi gaussienne de paramètres μ_i et σ^2 :

$$f_{\alpha_i}(y_i) = \exp\left\{\frac{y_i x_i'\beta - 0.5(x_i'\beta)^2}{\sigma^2} - \left(\frac{y_i^2}{2\sigma^2} - \frac{\log(2\pi\sigma^2)}{2}\right)\right\}.$$

- La fonction de lien est l'identité.

Figure :

Premier bilan

- Outre le choix classique de la composante déterministe (choix de la combinaison linéaire des variables explicatives), la modélisation GLM s'effectue à travers 2 choix :

  1. Choix de la loi de Y_i dans la famille exponentielle GLM décrite plus haut.
  2. Choix de la fonction de lien (inversible).

	logistique	linéaire
Loi expo	Bernoulli	Gaussienne
fdl	$g(u) = \text{logit}(u)$	$g(u) = u$

Figure :

Choix de la loi exponentielle et de la fonction de lien

1. Loi exponentielle. Ce choix est généralement guidé par la nature de la variable à expliquer (Binaire : Bernoulli, Comptage : Poisson, continue : normale ou gamma).
2. Fonction de lien. Ce choix est plus délicat. La fonction de lien dite "canonique" $g(u) = (b')^{-1}(u)$ est souvent privilégiée (notamment pour des raisons d'écriture de modèles et de simplicité d'écriture)

Propriété

Les fonctions de lien des modèles logistique et linéaire sont canoniques.

Figure :

Fonctions de lien usuelles

Nom du lien	Fonction de lien
identité	$g(u) = u$
log	$g(u) = \log(u)$
cloglog	$g(u) = \log(-\log(1-u))$
logit	$g(u) = log(u/(1-u))$
probit	$g(u) = \Phi^{-1}(u)$
réciproque	$g(u) = -1/u$
puissance	$g(u) = u^{\gamma}, \gamma \neq 0$

Figure :

GLM sur R

- Il faut bien entendu spécifier à la fonction **glm** les 3 composantes d'un modèle **glm** :

glm(formula=....,family=....(link=....))

1. formula : spécifie la composante déterministe $Y = X_1 + X_2$, $Y = X_1 + X_2 + X_1 : X_2$ (prendre en compte l'interaction entre X_1 et X_2.
2. family : spécifie composante aléatoire (gaussian pour le modèle linéaire gaussien, binomial lorsque la variable à expliquer est binaire...)
3. link : spécifie la fonction de lien (logit pour logistique, probit pour probit...)

Figure :

Exemple

- On cherche à expliquer une variable binaire Y par deux variables continues X_1 et X_2 :

```
> Y <- rbinom(50,1,0.6)
> X1 <- runif(50)
> X2 <- rnorm(50)
```

- On ajuste les modèles

```
> glm(Y~X1+X2,family=binomial)
Coefficients:
(Intercept)           X1           X2
    -0.2849       1.8610      -0.0804

> glm(Y~X1+X2+X1:X2,family=binomial)
Coefficients:
(Intercept)           X1           X2        X1:X2
    -0.3395       2.1175      -0.4568       1.0346

> glm(Y~X1+X2,family=binomial(link = "probit"))
Coefficients:
(Intercept)           X1           X2
   -0.17038      1.11986     -0.04864
```

Figure :

Modèle de Poisson

- On cherche à quantifier l'influence d'un traitement sur l'évolution du nombre de polypes au colon. On dispose des données suivantes :

```
   number    treat age
1      63 placebo  20
2       2     drug  16
3      28 placebo  18
4      17     drug  22
5      61 placebo  13
...
```

où

- `number` : nombre de polypes après 12 mois de traitement.
- `treat` : drug si le traitement a été administré, placebo sinon.
- `age` : age de l'individu.

Le problème est d'expliquer la variable `number` par les deux autres variables à l'aide d'un GLM.

Figure :

GLM

On note

- Y_i la variable aléatoire représentant le nombre de polypes du $i$ème patient après les 12 mois de traitement.
- x_{i1} la variable `treat` pour le $i$ème individu et x_{i2} l'age du $i$ème individu.

GLM

1. La variable Y_i étant une variable de comptage, on choisit comme densité de Y_i la densité (par rapport à la mesure de comptage) de la loi de Poisson de paramètre λ_i :

$$f_{\alpha_i}(y_i) = \exp(-\lambda_i)\frac{\lambda_i^{y_i}}{y_i!} = \exp[y_i \log(\lambda_i) - \exp(\log(\lambda_i)) - \log(y_i!)].$$

2. La fonction de lien canonique est donc donnée par :

$$g(u) = \log(u).$$

Figure :

Modèle de Poisson ou régression log-linéaire

Définition

Le modèle de Poisson modélise la loi de Y_i par une loi de Poisson de paramètre $\lambda_i = \lambda(x_i)$ telle que

$$\log(\lambda(x_i)) = x_i'\beta.$$

- L'ajustement sur R s'effectue toujours à l'aide de la fonction **glm** :

```
> glm(number~treat+age,data=polyps,family=poisson)

Coefficients:
(Intercept)     treatdrug           age
    4.52902      -1.35908      -0.03883
```

- **Prédiction :** pour un individu de 23 ans, ayant reçu le traitement on pourra estimer le nombre de polypes à 12 mois par

$$\exp(4.52902 - 1.35908 - 0.03883 * 23) = 9.745932.$$

Figure :