Apprentissage statistique supervisé

La Régression Logistique

Dr. Modeste Dayé

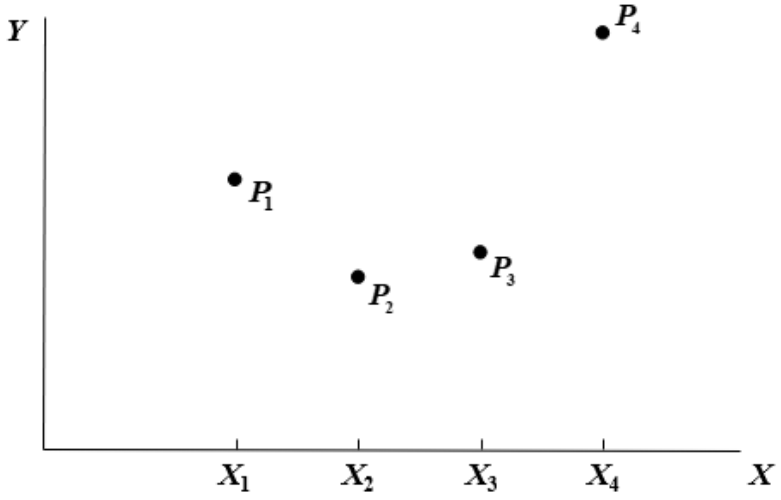
EEIA 2023

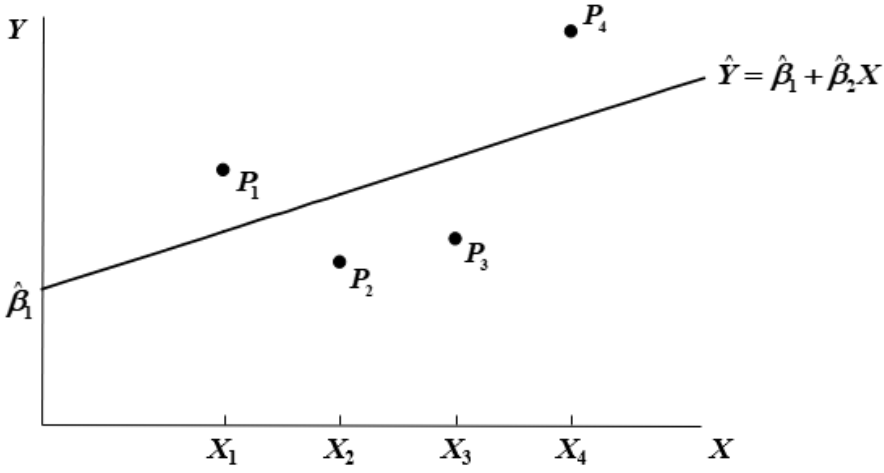
31 juillet 2023

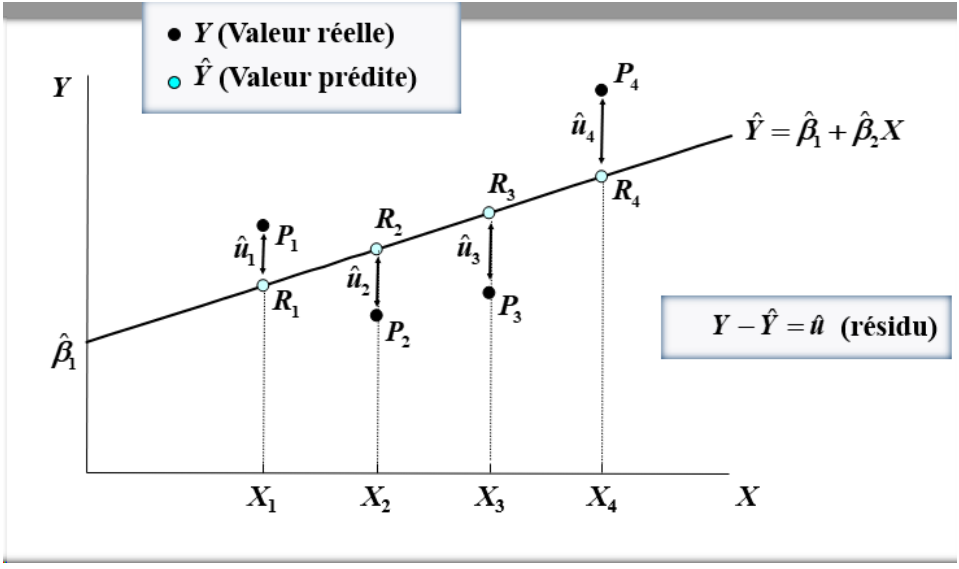
Plan

Prédiction ou explication d'une variable cible

- Dans un jeu de données contenant une variable (cible) d'intérêt (continue ou catégorielle) et des caractéristiques pertinentes qui lui sont liées :
  $y = f(X) + \epsilon$
  - Exercice de prédiction de y
  - Exercice d'explication de la mesure dans la quelle chacune des caractéristiques affecte y et sa pertinence lorsqu'on rapporte à la population : inférence statistique ;
- Dans les 2 cas, on veut comprendre le processus de génération des données : quelle est la logique qui lie les caractéristiques à la variable cible : est-ce une logique linéaire, non-linéaire, paramétrique, non-paramétrique ?
  - Il faut pouvoir utiliser un bon algorithme, et faire usage de la théorie liée au domaine (modèle économique, voir littérature économique) pour bien choisir son $f()$ ou ne même pas en imposer.


Y
P4
P1
P3
P2
X1
X2
X3
X4
X


Y
P_4
$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X$
P_1
P_2
P_3
$\hat{\beta}_1$
X_1
X_2
X_3
X_4
X


Y (Valeur réelle)
Ŷ (Valeur prédite)
Y
$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X$
P_4
$\hat{u}_4$
R_3
R_4
R_2
$\hat{u}_1$
P_1
$\hat{u}_3$
$\hat{u}_2$
R_1
P_3
P_2
$Y - \hat{Y} = \hat{u}$ (résidu)
$\hat{\beta}_1$
X_1
X_2
X_3
X_4
X

Régression

Objectif : On veut comprendre la variance, i.e. les différentes valeurs prises par une **variable continue à partir d'une ou de plusieurs caractéristiques**

- Cas simple : Régression linéaire simple :

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

,

- Moindre Carrées Ordinaires (MCO) : meilleur adjustment possible d'un nuage de points entre variables continues : $\min \sum_{i=1}^{n} \hat{u}_i^2 = (y_i - \hat{y}_i)^2$
- y continue

Questions de classification

- Certains poursuivent leurs études après le bac d'autres pas.
  - *Nous disposons des caractéristiques des élèves réussissant au bac (age, notes en classe et au bac, caractéristiques socio-démographiques des parents, etc.)*

Questions de classification

- Certains poursuivent leurs études après le bac d'autres pas.
  - *Nous disposons des caractéristiques des élèves réussissant au bac (age, notes en classe et au bac, caractéristiques socio-démographiques des parents, etc.)*
  - Peut-on construire un modèle qui nous dise les chances qu'un nouveau bachelier aille à l'université (ou dans une filière spécifique) ? Autrement dit, classer les nouveaux bacheliers.

Questions de classification

- Certains poursuivent leurs études après le bac d'autres pas.
  - *Nous disposons des caractéristiques des élèves réussissant au bac (age, notes en classe et au bac, caractéristiques socio-démographiques des parents, etc.)*
  - Peut-on construire un modèle qui nous dise les chances qu'un nouveau bachelier aille à l'université (ou dans une filière spécifique) ? Autrement dit, classer les nouveaux bacheliers.

- Certaines candidatures ont été retenues pour l'EEIA 2023 et d'autres pas.
  - *Peut-on prédire la probabilité d'être sélectionné (au moins pour la phase entretien) d'un candidat à l'EEIA étant données ses caractéristiques et la qualité de la LM (plagiat éliminatoire) ?*
  - Les formateurs de l'EEIA qui passent de nombreuses heures à lire les LM et discuter/creuser les motivations présentées par les candidats seront sans doute intéressés.

- Les variables binaires sont composées de 2 catégories : Oui/Non, malade/non-malade, admis/non-admis, etc.
  $\Rightarrow$ Deux réponses possibles : Non/Oui, que l'on pourra encoder dans une variable binaire numérique 0/1 pour exploitation dans les algorithmes de classification.

Plan

Le Modèle de Probabilité Linéaire (MLP)

- Soit le modèle linéaire simple :

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \text{pour un individu i}$$

où :

- y_i : Variable à expliquer (cible ou *target*) qui prend la valeur 1 si l'apprenant i avait déjà étudié la régression logistique par le passé et 0 sinon.
- x_i : le domaine de formation de l'apprenant i (variable catégorielle ici : stat, maths, info, agronomie, etc...)
- ϵ_i l'erreur de spécification du modèle (terme d'erreur)
- β_0 et β_1 sont les paramètres à estimer et qui permette de comprendre le processus de génération des données (y_i).

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Avec un modèle linéaire et une variable cible dichotomique, on prédit des probabilités :

- Si on respecte l'hypothèse d'espérance nulle du terme d'erreur ($E(\epsilon_i/x_i) = 0$) alors $E(y_i/x_i) = \beta_0 + \beta_1 x_i$
- On peut réécrire cette espérance sous forme de probabilité. Soit p_i la probabilité que $y_i = 1$ alors $(1 - p_i)$ est la probabilité que $y_i = 0$
  On a :

$$E(y_i/x_i) = 1 \times p_i + 0 \times (1 - p_i) = p_i = \beta_0 + \beta_1 x_i$$

D'où le fait qu'on parle de Modèle linéaire de probabilité :

$$\hat{p}_i(y_i = 1/x) = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

MLP :

$$\hat{p}_i(x) = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

MLP :

$$\hat{p}_i(x) = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

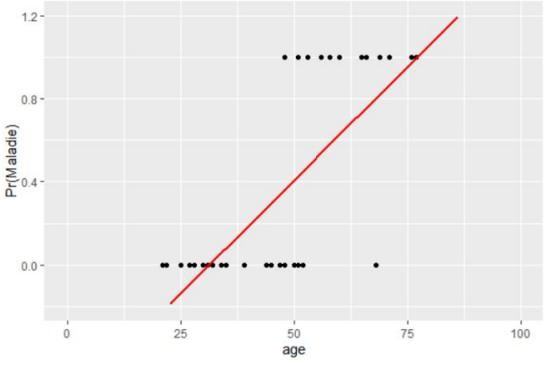
Problème : p $\in [0, 1]$ alors que :

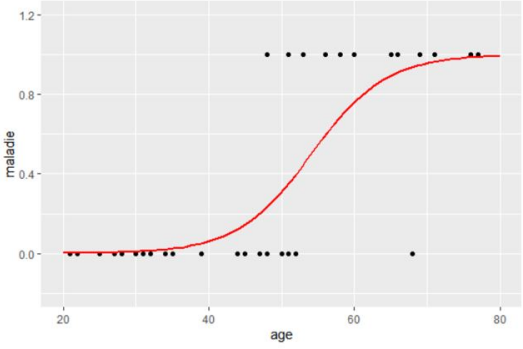
$$\hat{\beta}_0 + \hat{\beta}_1 x_i \in R$$

Idée : trouver une transformation ϕ *de* $p_i(x)$ telle que $\phi(p_i(x))$ prenne ses valeurs dans R.

$$\phi(\hat{p}_i(x)) = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Prédictions MLP parfois abérrantes (Voir TP plus tard)





Modélisation par les CDF : fonctions de répartition

$$\hat{p}_i(y_i = 1/x) = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Nous souhaitons que :

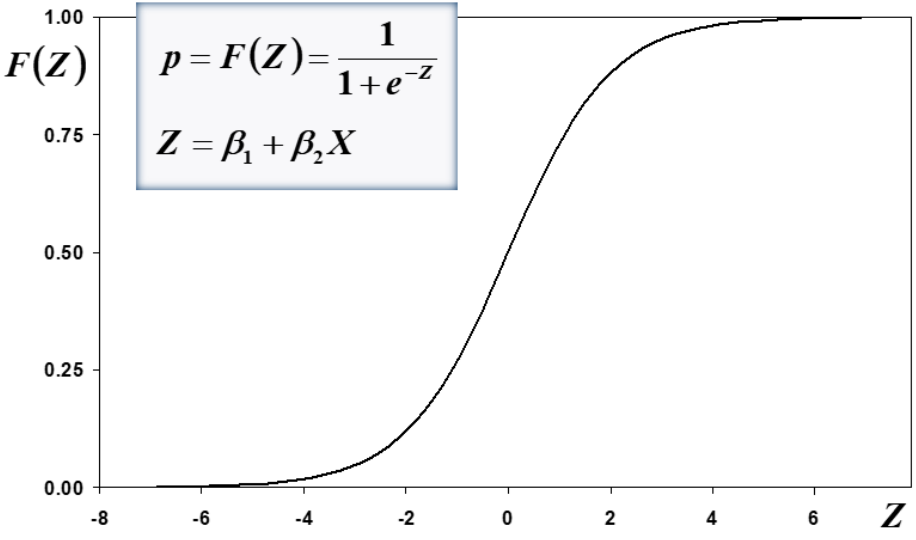
- $E(y_i/x_i) = p_i$ soit compris entre 0 et 1

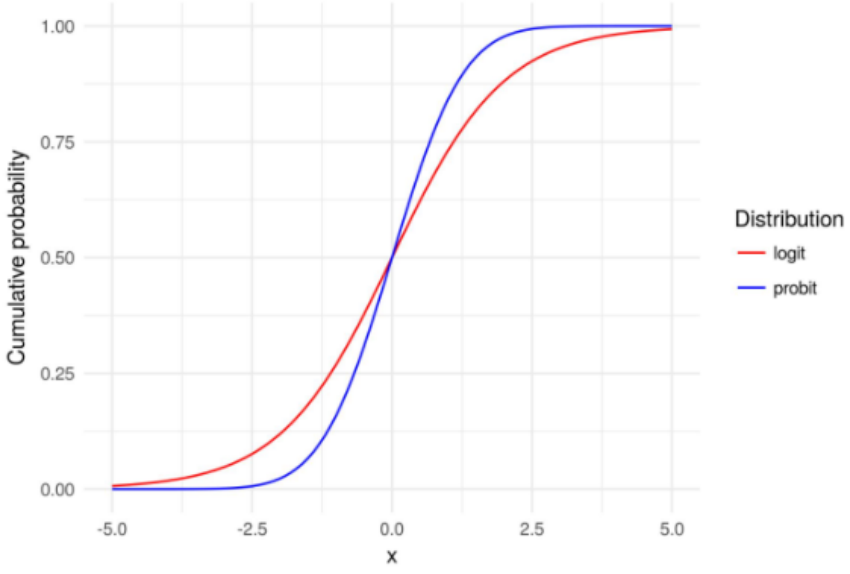
Deux modélisations sont utilisées en économétrie/statistique :

- Probit (CDF, loi normale)
- **Logistique (CDF, loi logistique)**

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

avec z la variable aléatoire suivant une distribution logistique.


$p = F(Z) = \frac{1}{1+e^{-z}}$
$Z = \beta_1 + \beta_2 X$
F(Z)
1.00
0.75
0.50
0.25
0.00
-8
-6
-4
-2
0
2
4
6
Z


1.00
0.75
0.50
0.25
0.00
Cumulative probability
-5.0
-2.5
0.0
2.5
5.0
x
Distribution
logit
probit

Figure

Régression Logistique

$$\hat{p}_i(y_i = 1/x) = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

La régression logistique modélise donc : $p_i(x) = \dfrac{exp(\beta_0 + \beta_1 x_i)}{1 + exp(\beta_0 + \beta_1 x_i)}$

Avec

- $\lim_{z_i \to +\infty} p_i = 1$
- $\lim_{z_i \to -\infty} p_i = 0$

Où : $z_i = \beta_0 + \beta_1 x_i$

Régression Logistique

$$\hat{p_i}(y_i = 1/x) = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

La régression logistique définit

- $p_i(x) = \dfrac{exp(\beta_0 + \beta_1 x_i)}{1 + exp(\beta_0 + \beta_1 x_i)} = \dfrac{1}{1 + exp(-(\beta_0 + \beta_1 x_i))}$

Régression Logistique

$$\hat{p}_i(y_i = 1/x) = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

La régression logistique définit

- $p_i(x) = \dfrac{exp(\beta_0 + \beta_1 x_i)}{1 + exp(\beta_0 + \beta_1 x_i)} = \dfrac{1}{1 + exp(-(\beta_0 + \beta_1 x_i))}$
- Et donc
  $1 - p_i(x) = 1 - \dfrac{1}{1 + exp(-(\beta_0 + \beta_1 x_i))} = \dfrac{exp(-(\beta_0 + \beta_1 x_i))}{1 + exp(-(\beta_0 + \beta_1 x_i))}$

Régression Logistique

$$\hat{p}_i(y_i = 1/x) = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

La régression logistique définit

- $p_i(x) = \dfrac{exp(\beta_0 + \beta_1 x_i)}{1 + exp(\beta_0 + \beta_1 x_i)} = \dfrac{1}{1 + exp(-(\beta_0 + \beta_1 x_i))}$
- Et donc
  $1 - p_i(x) = 1 - \dfrac{1}{1 + exp(-(\beta_0 + \beta_1 x_i))} = \dfrac{exp(-(\beta_0 + \beta_1 x_i))}{1 + exp(-(\beta_0 + \beta_1 x_i))}$
- $\dfrac{p_i}{1 - p_i} = exp(\beta_0 + \beta_1 x_i)$

Régression Logistique

- on a : $\frac{p_i}{1-p_i} = exp(\beta_0 + \beta_1 x_i)$

Régression Logistique

- on a : $\frac{p_i}{1-p_i} = exp(\beta_0 + \beta_1 x_i)$ **odds** : proba (chances) d'observer y=1 par rapport à y=0 pour une caractéristique x donnée.

Régression Logistique

- on a : $\frac{p_i}{1-p_i} = exp(\beta_0 + \beta_1 x_i)$ **odds** : proba (chances) d'observer y=1 par rapport à y=0 pour une caractéristique x donnée.
- $log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 x_i$ avec $log(\frac{p_i}{1-p_i})$, la fonction logit de p_i ou encore le log-Odds.

Régression Logistique

- on a : $\frac{p_i}{1-p_i} = exp(\beta_0 + \beta_1 x_i)$ **odds** : proba (chances) d'observer y=1 par rapport à y=0 pour une caractéristique x donnée.
- $log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 x_i$ avec $log(\frac{p_i}{1-p_i})$, la fonction logit de p_i ou encore le log-Odds.
- **Rappel** : On voulait trouver une transformation ϕ *de* $p(x)$ telle que $\phi(p_i(x))$ prenne ses valeurs dans R et donc que notre prédiction $\hat{p}_i(y_i = 1/x) = \hat{\beta}_0 + \hat{\beta}_1 x_i$ soit définie.

Régression Logistique

- on a : $\frac{p_i}{1-p_i} = exp(\beta_0 + \beta_1 x_i)$ **odds** : proba (chances) d'observer y=1 par rapport à y=0 pour une caractéristique x donnée.
- $log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 x_i$ avec $log(\frac{p_i}{1-p_i})$, la fonction logit de p_i ou encore le log-Odds.
- **Rappel** : On voulait trouver une transformation ϕ *de* $p(x)$ telle que $\phi(p_i(x))$ prenne ses valeurs dans R et donc que notre prédiction $\hat{p}_i(y_i = 1/x) = \hat{\beta}_0 + \hat{\beta}_1 x_i$ soit définie.
- Faire une régression logistique revient donc estimer : $\phi(p) = log(\frac{p}{1-p}) = \beta_0 + \beta_1 x$, qui fait dont partie de la classe des modèles linéaires généralisés.

Plan

Estimation par maximum de vraisemblance

- Vraisemblance (intuitivement) : Probabilité que le processus de génération des données décrit par le modèle ait produit les données réellement observées. (objectif en termes d'optimisation d'une telle fonction ?)

Estimation par maximum de vraisemblance

- Vraisemblance (intuitivement) : Probabilité que le processus de génération des données décrit par le modèle ait produit les données réellement observées. (objectif en termes d'optimisation d'une telle fonction ?)
  $\rightarrow$ la maximiser !

Formalisation de fonction de vraisemblance :

On a :

- $proba(y_i = 1|x) = F(\beta_0 + \beta_1 x_i) = \dfrac{exp(\beta_0 + \beta_1 x_i)}{1 + exp(\beta_0 + \beta_1 x_i)}$
- $proba(y_i = 0|x) = 1 - F(\beta_0 + \beta_1 x_i)$

On peut alors écrire :

$$proba(y_i = h_i|x) = (F(\beta_0 + \beta_1 x_i))^{h_i} (1 - F(\beta_0 + \beta_1 x_i))^{1-h_i}$$

avec $h_i = 0$ ou $h_i = 1$ (remplacer h_i par 0 et par 1 et observer qu'on retrouve bien les formules de probabilité données ci-dessus).

Si on généralise cela à l'ensemble de observations en supposant que les individus sont **iid** on obtient la fonction de vraisemblance :

$$L = \prod_{i=1}^{n} (F(\beta_0 + \beta_1 x_i))^{h_i} (1 - F(\beta_0 + \beta_1 x_i))^{1-h_i} = \prod_{i=1}^{n} p_i^{h_i} (1 - p_i)^{1-h_i}$$

Estimation par maximum de vraisemblance

- Fonction de vraisemblance :

$$L = \prod_{i=1}^{n} \left(F(\beta_0 + \beta_1 x_i)\right)^{h_i} \left(1 - F(\beta_0 + \beta_1 x_i)\right)^{1-h_i} = \prod_{i=1}^{n} p_i^{h_i} (1 - p_i)^{1-h_i}$$

- Fonction de log-vraisemblance :

$$ln(L) = \sum_{i=1}^{n} \left[y_i ln\left(F(\beta_0 + \beta_1 x_i)\right) + (1 - y_i) ln\left(1 - F(\beta_0 + \beta_1 x_i)\right)\right]$$

Estimation par maximum de vraisemblance

- Fonction de vraisemblance :

$$L = \prod_{i=1}^{n} \left(F(\beta_0 + \beta_1 x_i)\right)^{h_i} \left(1 - F(\beta_0 + \beta_1 x_i)\right)^{1-h_i} = \prod_{i=1}^{n} p_i^{h_i} (1 - p_i)^{1-h_i}$$

- Fonction de log-vraisemblance :

$$ln(L) = \sum_{i=1}^{n} \left[y_i ln\left(F(\beta_0 + \beta_1 x_i)\right) + (1 - y_i) ln\left(1 - F(\beta_0 + \beta_1 x_i)\right)\right]$$

Estimation par maximum de vraisemblance

- Fonction de vraisemblance :

$$L = \prod_{i=1}^{n} \left(F(\beta_0 + \beta_1 x_i)\right)^{h_i} \left(1 - F(\beta_0 + \beta_1 x_i)\right)^{1-h_i} = \prod_{i=1}^{n} p_i^{h_i} (1 - p_i)^{1-h_i}$$

- Fonction de log-vraisemblance :

$$ln(L) = \sum_{i=1}^{n} \left[y_i ln\left(F(\beta_0 + \beta_1 x_i)\right) + (1 - y_i) ln\left(1 - F(\beta_0 + \beta_1 x_i)\right)\right]$$

- On peut montrer que la fonction de log-vraisemblance est concave et qu'elle admet donc un maximum

Estimation par maximum de vraisemblance

- Fonction de vraisemblance :

$$L = \prod_{i=1}^{n} \left(F(\beta_0 + \beta_1 x_i)\right)^{h_i} \left(1 - F(\beta_0 + \beta_1 x_i)\right)^{1-h_i} = \prod_{i=1}^{n} p_i^{h_i} (1 - p_i)^{1-h_i}$$

- Fonction de log-vraisemblance :

$$ln(L) = \sum_{i=1}^{n} \left[y_i ln\left(F(\beta_0 + \beta_1 x_i)\right) + (1 - y_i) ln\left(1 - F(\beta_0 + \beta_1 x_i)\right)\right]$$

- On peut montrer que la fonction de log-vraisemblance est concave et qu'elle admet donc un maximum
- Les coefficients $\widehat{\beta}_i$ sont ceux qui maximisent la fonction de log-vraisemblance. $\Rightarrow$ Calculer les dérivées partielles de la fonction et les égaliser à 0 (pas facile à la main...)

Estimation par maximum de vraisemblance

- Fonction de vraisemblance :

$$L = \prod_{i=1}^{n} \left(F(\beta_0 + \beta_1 x_i)\right)^{h_i} \left(1 - F(\beta_0 + \beta_1 x_i)\right)^{1-h_i} = \prod_{i=1}^{n} p_i^{h_i} (1 - p_i)^{1-h_i}$$

- Fonction de log-vraisemblance :

$$ln(L) = \sum_{i=1}^{n} \left[y_i ln\left(F(\beta_0 + \beta_1 x_i)\right) + (1 - y_i) ln\left(1 - F(\beta_0 + \beta_1 x_i)\right)\right]$$

- On peut montrer que la fonction de log-vraisemblance est concave et qu'elle admet donc un maximum
- Les coefficients $\widehat{\beta}_i$ sont ceux qui maximisent la fonction de log-vraisemblance. $\Rightarrow$ Calculer les dérivées partielles de la fonction et les égaliser à 0 (pas facile à la main...)
- Mais bon, si F est simple, possible de s'amuser à dériver à la main....mais, le algorithmes et logiciels pour le faire sont nombreux.

Estimation par minimisation de la fonction de coût (logLoss)

- Fonction de log-vraisemblance :

$$ln(L) = \sum_{i=1}^{n} [y_i ln\left(F(\beta_0 + \beta_1 x_i)\right) + (1 - y_i) ln\left(1 - F(\beta_0 + \beta_1 x_i)\right)]$$

Estimation par minimisation de la fonction de coût (logLoss)

- Fonction de log-vraisemblance :

$$ln(L) = \sum_{i=1}^{n} [y_i ln\left(F(\beta_0 + \beta_1 x_i)\right) + (1 - y_i) ln\left(1 - F(\beta_0 + \beta_1 x_i)\right)]$$

- Fonction de log-loss (cross entropy loss function) :

$$log{-}loss = -ln(L) = -\sum_{i=1}^{n} [y_i ln\left(F(\beta_0 + \beta_1 x_i)\right) + (1 - y_i) ln\left(1 - F(\beta_0 + \beta_1 x_i)\right)]$$

$\rightarrow$ Descente de gradient

Plan

En pratique, 3 étapes essentielles après avoir préparé vos données (encodage des variable catégorielles notamment, gestion des NaN (missing values)) :

1. Diviser (split) la base de données en échantillon d'entraînement () et en échantillon test

```
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.2, random_state = 21)
```

2. Entraîner le modèle sur l'échantillon "*train*" :
   le modèle apprend le processus de génération des données, i.e, comment est-ce que les caractéristiques utilisées (prédicteurs ou *features*) classifient les individus suivant la variable dépendante dans la catégorie 1 ou dans la catégorie 0.

```
LR = LogisticRegression().fit(X_train,Y_train)
```

3. Tester le modèle en évaluant sa qualité prédictive.
   Dans quelle mesure est ce que le modèle a "bien appris" le processus de génération des valeurs de la variable cible ?) :
   - En pratique (par défaut), si $\widehat{Prob}(Y_i = 1) > 0.5$, on retient que l'individu concerné est classé par le modèle dans la catégorie 1 de la variable cible, étant données ses caractéristiques.

Plan

Qualité/performance de l'exercice

1 Première approximation : Matrice de confusion.

	y_i **observé** $= 0$	y_i **observé** $= 1$
y_i **estimé** $=0$	nb estimation correcte	nb estimation incorrecte
y_i **estimé** $=1$	nb estimation incorrecte	nb estimation correcte

Qualité/performance de l'exercice

$$\Rightarrow \text{Count} - R^2 = \frac{\text{nombre total d'estimations correctes}}{\text{nombre total d'estimations}}$$

Problème : les valeurs des y_i estimés sont comprises entre 0 et 1 mais seront la plupart du temps différentes des valeurs 0 et 1 $\Rightarrow$ à partir de quel seuil de $\widehat{y}_i$ peut on considérer que $\widehat{y}_i = 1$? (pas de réponse universelle, généralement > 0.5).

Exemple : Soit la matrice de confusion post estimation :

Observé/estimé	**1**	**0**
1	57	21
0	12	60

$\rightarrow$ A quoi est égal le count-R^2 ou score de prédiction ou accuracy (exactitude).

Exemple : Soit la matrice de confusion post estimation :

Observé/estimé	**1**	**0**
1	57	21
0	12	60

$\rightarrow$ A quoi est égal le count-R^2 ou score de prédiction ou accuracy (exactitude).
$\text{Count} - R^2 = \frac{57+60}{57+21+12+60} = \frac{117}{150} = 0.78$

2 pseudoR^2 de Mc-Fadden : $R^2_{MF} = 1 - \frac{ln(L_{NR})}{ln(L_R)}$
Où : $ln(L_{NR})$ est la fonction de log-vraissembleance dans le modèle non restreint (avec tous les régresseurs) et $ln(L_R)$ est la fonction de log-vraissemblance dans le modèle restreint (avec uniquement la constante β_0)

3 Différents taux d'intérêt (Précision, Recall ou Sensibilité, spécificité,...) : Voir TP.

4 Receiver Operator Characteristic ROC, basé sur l'Area Under the Curve (Air sous la Courbe)-AUC : *taux de Vrais positifs* ou sensibilité en fonction du taux de "faux positifs" (1-spécificité) à différents seuils de classement (Voir TP).

5 Validation croisée (non couvert)