École nationale de la statistique et de l'analyse de l'information

STATISTIQUE MATHÉMATIQUE - MS ENSAI
TP1 : Régression logistique - CORRIGÉ.
Oct 2026

1 Le modèle logistique

1.1 Pour quoi faire ?

Une variable binaire $Y \in \{0,1\}$ est naturellement modélisée conditionnellement aux variables explicatives $X = x$ par une loi de Bernoulli :

$$Y \mid X = x \sim \text{Bernoulli}(p(x)), \tag{1}$$

$$\mathbb{E}[Y \mid X = x] = p(x), \tag{2}$$

$$\mathbb{V}(Y \mid X = x) = p(x)(1 - p(x)). \tag{3}$$

En régression linéaire, $\mathbb{E}[Y \mid X = x] = x^\top \beta$ peut sortir de $[0,1]$; de plus, la variance conditionnelle est hétéroscédastique. Le modèle logistique pose

$$\underbrace{\log\left(\frac{p(x)}{1-p(x)}\right)}_{\text{logit}(p(x))} = \beta_0 + \sum_{j=1}^{q} \beta_j x_j =: L(x), \tag{4}$$

$$p(x) = \underbrace{\frac{1}{1+e^{-L(x)}}}_{\Lambda(L(x))}. \tag{5}$$

La fonction Λ (*logistique*) impose $0 < p(x) < 1$.

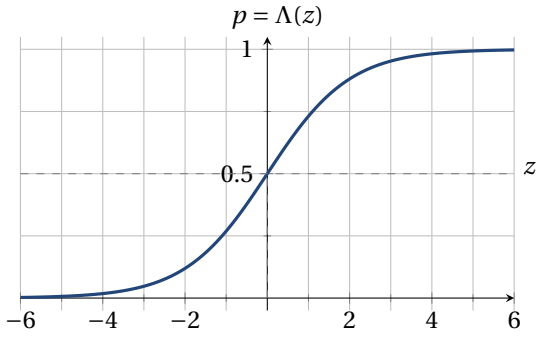


Figure 1: La fonction logistique.

1.2 Cote, rapport de cotes et interprétation des coefficients

La *cote* (*odds*) est $o(x) = p(x)/(1-p(x))$; il ne faut pas la confondre avec la probabilité $p(x)$. D'après (5), $o(x) = \exp(L(x))$. Si seule la variable x_j augmente d'une unité, toutes choses égales par ailleurs :

$$\frac{o(x + e_j)}{o(x)} = e^{\beta_j} =: \text{OR}_j, \text{ et pour } \Delta \text{ unités, } \text{OR}_j(\Delta) = e^{\Delta \beta_j}. \tag{6}$$

Un OR de 1.20 signifie que *la cote* augmente de 20% et non que la probabilité augmente de 20 points ni de 20%. En particulier, la dérivée est $\partial p(x)/\partial x_j = \beta_j p(x)(1-p(x))$: l'effet marginal sur la probabilité dépend de x.

1.3 Estimation par maximum de vraisemblance

Pour des observations indépendantes (x_i, y_i), posons $p_i(\beta) = \Lambda(x_i^\top \beta)$, avec une constante dans x_i. La vraisemblance et la log-vraisemblance sont

$$L(\beta) = \prod_{i=1}^{n} p_i(\beta)^{y_i} (1 - p_i(\beta))^{1-y_i}, \tag{7}$$

$$\ell(\beta) = \sum_{i=1}^{n} \big[y_i \log p_i(\beta) + (1-y_i) \log(1 - p_i(\beta)) \big]. \tag{8}$$

Le maximum de vraisemblance $\widehat{\beta}$ maximise ℓ numériquement. Le score et l'information (observée et attendue, ici identiques) valent

$$U(\beta) = \frac{\partial \ell}{\partial \beta} = X^\top (y - p), \tag{9}$$

$$\mathbb{I}(\beta) = -\frac{\partial^2 \ell}{\partial \beta \partial \beta^\top} = X^\top W X, \tag{10}$$

$$W = \text{diag}\big(p_i(1-p_i)\big). \tag{11}$$

Sous les hypothèses de régularité et si le modèle est identifiable, $\widehat{\beta}$ est asymptotiquement normal, de matrice de covariance estimée $\mathbb{V}(\widehat{\beta}) = (X^\top \widehat{W} X)^{-1}$. Les écarts types sont les racines carrées des termes diagonaux. La méthode IRLS (*iteratively reweighted least squares*) utilisée par `GLM` résoud numériquement et itérativement les équations du score.

2 Partie I — Les iris de Fisher

2.1 Charger et décrire les données

Le jeu de données `Iris` contient $n = 150$ fleurs, réparties à parts égales entre les trois espèces *setosa, versicolor* et *virginica.* Chaque fleur est décrite par quatre variables quantitatives (en centimètres) : longueur et largeur des sépales, longueur et largeur des pétales. On note

$$X_i = (X_{i1}, X_{i2}, X_{i3}, X_{i4}) \in \mathbb{R}^4, \tag{12}$$

$$Y_i \in \{0,1,2\},\ i = 1, \ldots, 150. \tag{13}$$

L'objet retourné par `load_iris()` est un `Bunch`, proche d'un dictionnaire Python avec accès par attributs : `iris.data` contient la matrice 150×4, `iris.target` les 150 étiquettes, `iris.feature_names` le nom des mesures et `iris.target_names` les noms des espèces, dans l'ordre des codes $0, 1, 2$.

Un `DataFrame` est un tableau à deux dimensions, avec noms de colonnes et types éventuellement différents : sa logique ressemble à celle d'une feuille de calcul Excel. Il facilite les opérations de tri, filtrage, calcul de moyennes et visualisation. Le premier nuage représente *uniquement les sépales* : la séparation est moins nette qu'avec les pétales ; *setosa* est facilement reconnaissable à cause de ses pétales.

2.2 ACP en trois dimensions

Une ACP recherche les directions orthogonales maximisant successivement la variance projetée. Après centrage-réduction, les valeurs propres de la matrice de corrélation fournissent la variance expliquée.

2.3 Matrice de graphiques `pairplot`

Le `pairplot` présente une matrice de graphiques : *sur la diagonale*, les histogrammes indiquent les distributions empiriques de chaque mesure selon l'espèce ; *hors diagonale*, les nuages de points montrent la répartition des fleurs pour chaque paire de mesures. On contstate que :

1. *Setosa* se distingue nettement des deux autres espèces par la longueur et la largeur de ses pétales, nettement plus petits.
2. *Versicolor* et *virginica* se recouvrent partiellement, notamment si l'on n'observe que les sépales. Une classification parfaite par une seule mesure est donc peu plausible.
3. La longueur et la largeur des pétales sont *fortement corrélées* positivement : les pétales plus longs sont généralement plus larges. Ce lien peut être dû en partie aux différences entre espèces et doit être distingué de la corrélation *au sein* de chaque espèce.

Ce diagnostic exploratoire motive la construction de classifieurs (régression logistique, k plus proches voisins, etc.) et aide à choisir les variables. Il ne dispense pas d'une évaluation prédictive sur observations nouvelles.

2.4 Fusion des deux premières espèces : logit binaire

L'instruction `iris.data[:, 0:2]` sélectionne toutes les lignes (symbole `:`) et uniquement les colonnes d'indices 0 et 1 : la borne supérieure 2 est exclue. On retient donc X_1, longueur du sépale, et X_2, largeur du sépale (en cm), et $X \in \mathbb{R}^{150\times 2}$. La condition `iris.target != 2` produit des booléens ; la conversion `.astype(int)` transforme `False` en 0 et `True` en 1. La sélection `X[y==0][:,0]` récupère les abscisses des fleurs de classe 0, et `X[y==0][:,1]` leurs ordonnées. La commande `model.fit(X,y)` ajuste les coefficients en maximisant la log-vraisemblance *pénalisée* dans `scikit-learn` (pénalisation L^2 par défaut). Le modèle impose

$$P[Y=1 \mid X=x] = \frac{1}{1+\exp[-(\beta_0+\beta_1 x_1+\beta_2 x_2)]}. \qquad (14)$$

Les probabilités de `predict_proba` apparaissent dans l'ordre `model.classes_`, ici $[0,1]$. Pour les quatre points d'essai, les prédictions sont $[1,0,1,1]$ et les probabilités estimées de la classe 1 sont, aux arrondis près, $(0.831, 0.185, 0.999, 0.951)$.

Interprétation : à largeur de sépale fixée, un centimètre de plus de longueur multiplie la cote $\mathbb{P}[Y=1 \mid x]/\mathbb{P}[Y=0 \mid x]$ par $e^{-2.6025} \simeq 0.074$ (diminution de 92.6% de la *cote*, non de la probabilité). À longueur fixée, un centimètre de plus en largeur la multiplie par $e^{0.75} \simeq 2.11$; mais l'estimation est incertaine ($p \simeq 0.24$, intervalle de confiance du coefficient contenant 0). L'intervalle de confiance de l'OR contient donc 1 : on ne peut pas conclure à une association conditionnelle non nulle au seuil 5%.

Le test du *rapport de vraisemblance global* fournit $p \simeq 5.03 \times 10^{-17}$, tandis que le pseudo-$R^2$ de McFadden vaut environ 0.3931. La p-valeur porte sur l'hypothèse $H_0 : \beta_1 = \beta_2 = 0$ et non sur chaque coefficient séparément. Le pseudo-R^2 mesure un gain de log-vraisemblance par rapport à la constante seule, *pas* la proportion de variance expliquée.

Terme	$\widehat{\beta}$	Erreur type	z	p-valeur	$e^{\widehat{\beta}}$
Constante	14.1836	3.117	4.550	< 0.001	—
Longueur sépale	-2.6025	0.441	-5.907	< 0.001	0.0741
Largeur sépale	0.7458	0.641	1.163	0.245	2.108

Table 1: Maximum de vraisemblance `statsmodels` : $Y=1$ pour setosa ou versicolor, $Y=0$ pour virginica ; X_1, X_2 désignent les sépales, $n=150$.

Quand la séparation rend le résumé inexploitable

Si l'on code `y=(iris.target != 0).astype(int)`, on cherche à séparer *setosa* des deux autres espèces. Sur les sépales, une droite séparatrice parfaite existe pour cet échantillon : l'EMV logistique *non régularisé* n'a pas de maximum fini. Un ajustement `sm.Logit` peut alors donner de très grands coefficients, des écarts-types énormes, ainsi que des p-valeurs proches de 1. Les statistiques de Wald usuelles deviennent inutilisables. La situation est également fréquente avec les dimensions des pétales. Il faut surveillers les indicateurs de non-convergence, la séparation et éventuellement recourir à une estimation pénalisée (ou à une méthode adaptée à la séparation).

2.5 Régression polytomique sur trois espèces

La réponse comporte maintenant trois catégories $Y \in \{0,1,2\}$ et les quatre variables explicatives sont utilisées simultanément. On écrit

$$P(Y=k \mid X=x) = \frac{\exp(\alpha_k + x^\top \beta_k)}{\sum_{j=0}^{2} \exp(\alpha_j + x^\top \beta_j)}, \qquad k=0,1,2. \quad (15)$$

Pour assurer l'identifiabilité, on fixe par exemple $\alpha_0 = 0$ et $\beta_0 = 0$, ce qui revient à prendre *setosa* comme référence. La classe prédite est $\widehat{Y}(x) = \arg\max_k \mathbb{P}[Y=k \mid X=x]$. La fonction `LogisticRegression` ajuste ici un classifieur multinomial régularisé. Pour l'inférence sur les coefficients, on utilise `sm.MNLogit` sans pénalisation.

La standardisation soustrait la moyenne empirique de chaque caractéristique puis divise par son écart type. Elle rend les ordres de grandeur comparables, et favorise l'optimisation et la régularisation. Le tableau `MNLogit.summary()` fournit, par rapport à la classe de référence, les coefficients, les erreurs types, les statistiques z et les p-valeurs. Attention : les espèces Iris sont presque ou parfaitement séparables dans certaines projections. Même si `MNLogit` annonce une convergence, des valeurs `nan`, de très forts écarts-types ou des coefficients extrêmes peuvent apparaître. Les p-valeurs du modèle non pénalisé ne sont alors pas fiables. Le classifieur régularisé peut néanmoins donner de bonnes prédictions. `MNLogit` sur les 150 observations est une analyse descriptive, tandis que la qualité prédictive sera mesurée sur les données de test réservées avant ajustement.

2.6 Qualité de la classification

```
print(classification_report(yte3, ypred3,
                            target_names=iris.target_names)
    )
print('Confusion :\n', confusion_matrix(yte3, ypred3))
```

Pour la partition `test_size=0.25, random_state=0`, *sans* `stratify`, le modèle entraîné sur les quatre mesures standardisées classe correctement 37 des 38 fleurs du test, soit 97.37%. Le tableau de confusion, dont les *lignes* sont les classes observées, est

$$\begin{pmatrix} 13 & 0 & 0 \\ 0 & 15 & 1 \\ 0 & 0 & 9 \end{pmatrix}$$ (ordre : setosa, versicolor, virginica).

Classe	Précision	Rappel	F_1	Effectif du test
Setosa	1.00	1.00	1.00	13
Versicolor	1.00	0.94	0.97	16
Virginica	0.90	1.00	0.95	9

Table 2: Évaluation du modèle multinomial sur le test : résultats reproduits avec le découpage indiqué (nombres arrondis).

Ainsi, *setosa* est parfaitement identifiée dans ce test; une *versicolor* a été classée comme *virginica*, ce qui explique le rappel $15/16 \simeq 0.94$ de la première et la précision $9/10 = 0.90$ de la seconde. Un autre découpage ou une stratification différente peut modifier ces résultats.

Précision, rappel, F_1 et `classification_report`

Pour une classe positive donnée, notons VP, FP, FN et VN les quatre cases de la matrice de confusion :

$$\text{précision} = \frac{\text{VP}}{\text{VP} + \text{FP}}, \qquad \text{rappel} = \frac{\text{VP}}{\text{VP} + \text{FN}},$$
$$F_1 = \frac{2\,\text{précision} \times \text{rappel}}{\text{précision} + \text{rappel}}, \qquad \text{exactitude} = \frac{\text{bonnes prédictions}}{n}.$$

La `macro avg` calcule une moyenne non pondérée entre classes; la `weighted avg` utilise les effectifs (`support`).

3 Maladie coronarienne (SAheart)

3.1 Problématique et description

La base étudiée est le jeu de données `SAheart` qui rassemble 462 observations issues d'une étude rétrospective portant sur des hommes d'une région d'Afrique du Sud et contient 160 cas de maladie coronarienne (`chd=1`) contre 302 témoins (`chd=0`). On cherche des *associations* entre le statut coronarien et neuf caractéristiques, non à démontrer une relation de causalité.

Le nuage `chd` en fonction de l'âge occupe uniquement les droites horizontales $y = 0$ et $y = 1$. Pour voir les superpositions, on peut ajouter un léger *jitter* vertical; ce déplacement n'altère pas les données utilisées pour estimer le modèle.

3.2 Pourquoi pas une régression linéaire ? Proportions par âge

Une régression linéaire de $Y \in \{0,1\}$ sur l'âge n'impose pas $0 \le \widehat{p}(x) \le 1$ et ne représente pas correctement la variance conditionnelle $p(x)(1-p(x))$. On étudie d'abord la *proportion observée* de malades à l'intérieur de classes d'âge. Pour une classe A_h :

$$\widehat{p}_h = \frac{\#\{i : Y_i = 1,\ \text{âge}_i \in A_h\}}{\#\{i : \text{âge}_i \in A_h\}}. \tag{16}$$

Il ne s'agit pas d'une proportion cumulée : le dénominateur change avec la classe d'âge et les fluctuations dépendent des effectifs locaux.

Variable	Signification
`chd`	Maladie coronarienne observée : 0 non, 1 oui
`age`	Âge (années)
`sbp`	Pression artérielle systolique
`tobacco`	Consommation cumulative de tabac
`ldl`	Cholestérol LDL
`adiposity`	Mesure d'adiposité
`famhist`	Antécédents familiaux : `Absent`/`Present`
`typea`	Score de comportement de type A
`obesity`	Indicateur d'obésité (indice de masse corporelle)
`alcohol`	Consommation d'alcool

Pourquoi la courbe n'atteint-elle pas nécessairement 1 ? Parce que les classes d'âge même élevées contiennent aussi des individus sans événement coronarien; la proportion observée n'est donc pas obligée d'atteindre 100%. Si la dernière classe affiche environ 0.6, cela signifie seulement qu'environ 60% des personnes *de cette classe* ont $Y = 1$. La proportion globale observée vaut $160/462 \simeq 0.346$, et non 0.6 ; de surcroît, en raison du plan cas–témoins, elle ne mesure pas la prévalence dans la population.

3.3 Ajustement `chd` ~ `age`

Le notebook indique approximativement $\widehat{\beta}_0 \simeq -3.5$ et $\widehat{\beta}_1 \simeq 0.064$; on obtient la courbe

$$\widehat{p}(a) = \frac{\exp(\widehat{\beta}_0 + \widehat{\beta}_1 a)}{1 + \exp(\widehat{\beta}_0 + \widehat{\beta}_1 a)}. \tag{17}$$

On peut tracer cette fonction sur le même graphique que les proportions empiriques de la question 2.

Pour ces 462 observations, la pente estimée est environ $\widehat{\beta}_{\text{age}} = 0.0641$ (erreur type 0.00853), soit

$$\text{OR}_{\text{par an}} \simeq e^{0.0641} = 1.066, \text{OR}_{\text{sur dix ans}} \simeq e^{0.641} = 1.898. \tag{18}$$

Chaque année supplémentaire multiplie donc la *cote observée dans le modèle* par environ 1.066, et dix années par environ 1.90. Cela ne veut pas dire qu'une personne a 90% de risque supplémentaire. L'ordonnée à l'origine décrit des log-cotes hypothétiques à âge nul et n'a pas d'interprétation clinique pertinente. Significativité et ajustement : l'âge présente une p-valeur très faible, donc un signal statistique net. Cela *ne suffit pas* à qualifier à lui seul le modèle de « bonne qualité » : il faut examiner aussi la calibration, la discrimination et le comportement hors échantillon. Ni une log-vraisemblance ni une déviance ne sont « raisonnables » en valeur absolue sans comparaison pertinente.

3.4 Lecture du tableau de significativité

Les quatre colonnes essentielles de `summary()`

`coef` : Estimation $\widehat{\beta}_j$ du paramètre sur l'échelle logit.

`std err` : Écart type estimé de l'estimateur, *erreur standard* : $\widehat{\mathrm{se}}(\widehat{\beta}_j)$; ce n'est pas l'écart type de X_j.

`z` : Stat. de Wald $z_j = \widehat{\beta}_j / \widehat{\mathrm{se}}(\widehat{\beta}_j)$ sous $H_0 : \beta_j = 0$.

`P>|z|` : p-valeur du *test bilatéral* : $2(1 - \Phi(|z_j|))$ sous l'approximation normale asymptotique.

Qu'est-ce qu'une p-valeur ?

• **Définition intuitive.** La p-valeur quantifie, *sous H_0 et relativement à une statistique de test déterminée*, le caractère extrême des observations. Une faible p-valeur constitue un argument contre H_0; une grande valeur signifie que les données n'apportent pas suffisamment d'éléments pour la rejeter, sans prouver pour autant que H_0 est vraie.

• **Définition probabiliste.** Pour un test unilatéral à droite, de statistique $T(X)$, de valeur observée t_o et dont la loi sous une hypothèse nulle simple est spécifiée :

$$p(x_o) = \mathbb{P}_{H_0}\big[T(X) \geq t_o\big]. \tag{19}$$

La statistique est aléatoire sous H_0, tandis que sa valeur observée est fixée. Pour un test bilatéral, le sens de « au moins aussi extrême » dépend du test choisi : il n'existe pas une formule universelle identique pour toutes les lois et toutes les hypothèses.

• **Lien avec le seuil de significativité.** Dans une famille de tests aux régions de rejet emboîtées, la p-valeur est le plus petit niveau α conduisant au rejet de H_0. Si $p = 0.03$, on rejette au seuil 5%, mais pas au seuil 1%. Le niveau α est une *probabilité d'erreur de première espèce* contrôlée sous l'hypothèse nulle, et non la probabilité que H_0 soit vraie après avoir vu les données.

• **Exemple : test bilatéral de Wald.** On teste $H_0 : \beta_j = 0$ contre $H_1 : \beta_j \neq 0$. Sous les conditions asymptotiques usuelles, et en l'absence de séparation :

$$Z_j = \frac{\widehat{\beta}_j}{\widehat{\mathrm{se}}(\widehat{\beta}_j)} \overset{H_0}{\sim} \mathcal{N}(0,1), \; p = 2\{1 - \Phi(|z_{\mathrm{obs}}|)\}. \tag{20}$$

Pour $z_o = 2.5$, on obtient $p \simeq 2[1 - \Phi(2.5)] \simeq 0.0124$: rejet à 5%, mais pas à 1%. Le z mesure l'écart de l'estimation à zéro en unités d'erreur standard.

• **Ce que la p-valeur ne signifie pas.**

Affirmation	Correcte ?
La p-valeur est calculée sous l'hypothèse H_0.	Oui
Une faible p-valeur fournit un argument contre H_0.	Oui
La p-valeur vaut $P(H_0 \mid \text{données})$.	Non
Une p-valeur élevée prouve que H_0 est vraie.	Non
Une faible p-valeur prouve un effet important.	Non
$p = 0.03$ entraîne un rejet au niveau 5% pour ce test.	Oui

Table 3: Interprétations correctes et incorrectes d'une p-valeur.

• **Définition plus générale : validité sous une hypothèse composite.** Pour $H_0 : \theta \in \Theta_0$, une p-valeur valide est une statistique $p(X) \in [0,1]$ satisfaisant

$$\forall \theta \in \Theta_0, \; \forall \alpha \in [0,1], \; \mathbb{P}_\theta\big[p(X) \leq \alpha\big] \leq \alpha. \tag{21}$$

Le test qui rejette pour $p(X) \leq \alpha$ a alors un niveau au plus α. Pour certains tests continus exacts sous une hypothèse nulle simple, $p(X) \sim \mathcal{U}([0,1])$ sous H_0. Les p-valeurs asymptotiques utilisées dans le tableau `summary()` ne sont valides qu'approximativement lorsque les conditions de régularité sont satisfaites.

p-valeur	Code usuel	Décision au seuil 5%
$p < 0.001$	***	rejet net de H_0
$0.001 \leq p < 0.01$	**	rejet de H_0
$0.01 \leq p < 0.05$	*	rejet de H_0
$0.05 \leq p < 0.10$	.	non-rejet à 5%
$p \geq 0.10$	aucun	non-rejet à 5%

Table 4: Seuils usuels, purement conventionnels.

Pour l'âge seul, $z \simeq 0.0641/0.00853 \simeq 7.51$ et $p \simeq 5.8 \times 10^{-14}$. On rejette $H_0 : \beta_{\mathrm{age}} = 0$ aux seuils usuels, mais cela ne suffit pas à conclure à un effet causal. Un intervalle de confiance de Wald à 95% pour le coefficient est $\widehat{\beta}_j \pm 1.96\,\widehat{\mathrm{se}}(\widehat{\beta}_j)$.

3.5 Modèle complet et comparaison des significativités

Comment lire le résumé du modèle complet ?

Sur les 462 individus, la log-vraisemblance du modèle complet est $\ell(\widehat{\beta}) \simeq -236.07$ et la déviance vaut environ $472.14 = -2\ell(\widehat{\beta})$ pour des réponses de Bernoulli non regroupées. La statistique de Pearson est proche de 452 et `Df Residuals` vaut $462 - 10 = 452$. Le pseudo-$R^2 \simeq 0.2353$ affiché par `GLM.summary()` avec les paramètres par défaut est celui de *Cox–Snell*. On ne conclut *pas* qu'il explique 23.53% de la variance, ni qu'une log-vraisemblance « élevée » suffit à qualifier l'ajustement de bon. Le nombre de paramètres et la comparaison au modèle nul sont essentiels. Enfin, pour des V.A. de Bernoulli individuelles à covariables distinctes, une déviance proche de ses degrés de liberté ne garantit pas automatiquement une bonne adéquation par un test du χ^2 asymptotique.

Tableau de référence pour le modèle complet ajusté sur les 462 individus :

Variable	$\widehat{\beta}$	Erreur type	z	p-valeur	OR	Seuil 5%
Constante	−6.1507	1.3083	−4.701	$2.58 \cdot 10^{-6}$	0.0021	Oui
`sbp`	0.0065	0.00573	1.135	0.2564	1.0065	Non
`tobacco`	0.0794	0.02660	2.984	0.00285	1.0826	Oui
`ldl`	0.1739	0.05966	2.915	0.00356	1.1899	Oui
`adip.`	0.0186	0.02929	0.635	0.5257	1.0188	Non
`famhist`	0.9254	0.22789	4.061	$4.90 \cdot 10^{-5}$	2.5228	Oui
`typea`	0.0396	0.01232	3.214	0.00131	1.0404	Oui
`obesity`	−0.0629	0.04425	−1.422	0.1551	0.9390	Non
`alcohol`	0.00012	0.00448	0.027	0.9784	1.0001	Non
`age`	0.0452	0.01213	3.728	0.000193	1.0463	Oui

Table 5: Modèle `chd sbp+tobacco+ldl+...`. Les chiffres sont arrondis.

Interprétation et précautions

À variables restantes fixées, les variables `sbp` ($p \approx 0.256$) et `obesity` ($p \approx 0.155$) ne sont *pas significatives à* 5% *dans ce modèle*, mais cela ne démontre pas l'absence de relation. La variable `obesity` est estimée avec un signe négatif mais incertain ; il est erroné d'y voir un effet protecteur. Corrélation entre `obesity` et `adiposity`, confusion, sélection de cas, ou effets de traitement sont autant de pistes pour expliquer ce phénomène et il ne faut donc pas supprimer systématiquement ce type de variable : elle peut être utile à l'ajustement même si sa p-valeur est supérieure à 0.05. L'effet estimé de l'âge passe de 0.0641 (modèle simple) à 0.0452 (modèle ajusté) ; ce changement rappelle qu'un coefficient est toujours relatif aux autres variables contenues dans le modèle.

3.6 RFE, sélection backward, AIC, BIC, déviance, résidus

3.6.1 Sélection RFE

La méthode `RFE` élimine récursivement les variables de plus faible *importance* selon l'estimateur passé en argument. Pour une régression logistique, on classe couramment les coefficients absolus ; ce n'est pas un test de significativité et cela ne repose pas sur les p-valeurs. Avec des unités différentes, une standardisation est indispensable pour comparer les coefficients. Enfin, `famhist` doit être convertie en variable indicatrice.

```
from sklearn.feature_selection import RFE
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression

X = pd.get_dummies(maladie[variables],
                   columns=['famhist'], drop_first=True,
     dtype=float)
y = maladie['chd'].astype(int)
Z = StandardScaler().fit_transform(X)  # Analyse
     descriptive seulement
sel = RFE(LogisticRegression(max_iter=3000),
          n_features_to_select=5).fit(Z, y)
print(pd.DataFrame({'variable': X.columns,
                    'retenue': sel.support_, 'rang': sel.
     ranking_}))
```

L'importance brute des coefficients n'est pas comparable entre variables non standardisées. Dans une procédure *prédictive* validée, il faut ajuster le `StandardScaler` et `RFE` uniquement sur l'échantillon d'entraînement, sinon il y a rsique de fuite d'information. Le nombre de variables sélectionnées doit être choisi ou validé, et non considéré arbitrairement comme égal à cinq.

3.6.2 Sélection backward avec AIC/BIC

Pour un modèle ayant k paramètres (*constante comprise*) :

$$\mathrm{AIC} = -2\ell(\widehat{\beta}) + 2k, \tag{22}$$

$$\mathrm{BIC} = -2\ell(\widehat{\beta}) + k\log n. \tag{23}$$

La procédure backward part du modèle complet, retire à chaque étape la variable dont l'exclusion conduit à la meilleure baisse du critère, et s'arrête lorsqu'aucune exclusion ne l'améliore. Le minimum atteint peut être local : il ne prouve pas que tous les sous-modèles ont été évalués.

Sur SAheart, une solution classique obtenue par sélection backward AIC retient les variables `tobacco`, `ldl`, `famhist`, `typea` et `age`, avec une déviance d'environ 475.69 et un AIC d'environ 487.69 ($k = 6$). Pour le modèle complet ($k = 10$), la déviance est 472.14 et l'AIC 492.14. Le BIC applique une pénalité plus forte puisque $\log(462) \approx 6.14 > 2$; il peut aboutir à un sous-modèle différent. Les procédures AIC/BIC n'ont pas le même objectif que les tests de nullité à 5%.

3.6.3 Déviance et modèle saturé

La déviance est définie par

$$D = 2\{\ell(\widehat{\beta}_{\mathrm{sat}}) - \ell(\widehat{\beta})\}. \tag{24}$$

Le *modèle saturé* ajuste un paramètre de probabilité indépendant pour chacune des observations (ou un paramètre pour chaque profil regroupé). Lorsque chaque y_i est une variable de Bernoulli individuelle, il peut reproduire $\widehat{p}_i = y_i$ et sa log-vraisemblance, au sens de la convention $0\log 0 = 0$, vaut 0. Ainsi $D = -2\ell(\widehat{\beta})$.

Modèle	Déviance	ddl résiduels	AIC
Nul (constante seule)	596.11	461	598.11
Complet (9 prédicteurs)	472.14	452	492.14
Sous-modèle AIC 5 préd.	475.69	456	487.69

Table 6: Repères numériques sur SAheart. Les ddl valent $n-k$.

Null deviance désigne la déviance de la constante seule; **residual deviance**, celle du modèle ajusté. La différence de déviances de deux modèles *emboîtés* est une statistique de rapport de vraisemblance qui suit asymptotiquement une loi du χ^2 sous H_0... si les hypothèses de régularité sont vérifiées.

Adéquation globale et résidus. Ne pas utiliser systématiquement $D \sim \chi^2_{n-k}$ pour déclarer le modèle logistique adéquat lorsque les profils des covariables sont presque tous distincts : cette approximation peut être mauvaise.

Le résidu de Pearson est $r_{P,i} = (y_i - \widehat{p}_i)/\sqrt{\widehat{p}_i(1-\widehat{p}_i)}$. Le résidu de déviance est

$$r_{D,i} = \mathrm{sgn}(y_i - \widehat{p}_i)\sqrt{2\left[y_i \log\frac{y_i}{\widehat{p}_i} + (1-y_i)\log\frac{1-y_i}{1-\widehat{p}_i}\right]}$$

Pour une réponse binaire, il est normal que les graphiques de résidus présentent deux branches; il faut rechercher les valeurs extrêmes et les observations influentes.

Identifiabilité. Le modèle doit posséder une matrice de design $\mathbb{X}$ de rang colonne plein pour identifier localement ses coefficients; mais cela ne suffit *pas* à garantir un maximum de vraisemblance fini (la séparation parfaite ou quasi parfaite peut l'empêcher). Ici, le modèle logistique complet usuel sur SAheart est estimable avec des coefficients finis et un algorithme convergent. Un contrôle complet doit vérifier le rang, les duplications de colonnes, la séparation et la stabilité numérique. La non-significativité d'une variable n'est *pas* un défaut d'identifiabilité.

3.7 Les 3 tests : Wald, score et LR

Considérons un modèle complet M_1 et un modèle réduit M_0 où r coefficients sont imposés à zéro.

Test	Statistique	Modèle nécessaire
Wald	$(R\widehat{\beta})^\top [R\widehat{V}R^\top]^{-1}(R\widehat{\beta})$	Complet (M_1)
Score / LM	$U_1(\widetilde{\beta})^\top I_1(\widetilde{\beta})^{-1} U_1(\widetilde{\beta})$	
LRT	$2[\ell(\widehat{\beta}_1) - \ell(\widehat{\beta}_0)]$	Réduit et complet

Table 7: Statistiques de test. Sous H_0, leur loi limite est en général χ^2_r.

Pour $r = 1$, le Wald est z_j^2. Dans le test du score, il faut employer *l'information efficace* et pas uniquement la variance marginale du score. Tous trois sont asymptotiquement équivalents sous H_0, mais leurs valeurs peuvent différer pour un échantillon fini.

Les *tests globaux* utilisent $H_0 : \beta_1 = \cdots = \beta_9 = 0$ et comparent le modèle complet à l'ordonnée à l'origine seule. Pour SAheart :

$$G^2_{\text{global}} = 596.11 - 472.14 \simeq 123.97,\ r = 9,\ p \ll 0.001. \tag{25}$$

Dans `Logit.summary()`, `LLR p-value` est précisément la p-valeur du rapport de vraisemblance *global* entre le modèle courant et le modèle avec constante seule. `Log-Likelihood` est $\ell(\widehat{\beta})$, `LL-Null` est $\ell(\widehat{\beta}_0)$. Dans le cas `GLM`, le LRT peut être calculé manuellement à partir de `llf` ou des déviances.

3.8 Test de Hosmer–Lemeshow et calibration

On regroupe généralement les individus en $g = 10$ classes de risque prédit (déciles). Dans la classe h, notons n_h l'effectif, $O_h = \sum_{i \in h} y_i$ le nombre observé et $E_h = \sum_{i \in h} \widehat{p}_i$ le nombre attendu. Une forme usuelle du test est

$$H_{\text{HL}} = \sum_{h=1}^{g} \frac{(O_h - E_h)^2}{E_h(1 - E_h/n_h)}, \qquad H_{\text{HL}} \overset{\text{approx.}}{\sim} \chi^2_{g-2}. \tag{26}$$

H_0 indique une adéquation des fréquences observées et attendues à l'échelle des groupes. Une p-valeur faible suggère une *mauvaise* calibration; une p-valeur élevée ne prouve pas que le modèle soit correct.

Interprétation et précautions

Le test dépend du nombre de groupes et de la gestion des *ex æquo* ; il peut manquer de puissance sur petit échantillon ou rejeter pour des écarts minimes sur grand échantillon. Il est recommandé de le compléter par un graphique de calibration, un score de Brier et, si possible, une évaluation sur données nouvelles. Ici, la calibration absolue concerne d'abord la *population échantillonnée* et non la prévalence en population générale.

3.9 Rapports de cotes, probabilités et intervalles

Exemple sur le tabac dans un modèle réduit. Si le coefficient de `tobacco` est estimé à 0.0807 (comme dans une variante du notebook), un kilogramme supplémentaire de consommation *cumulative* de tabac multiplie, à covariables restantes fixées, la *cote* d'avoir `chd=1` par $\exp(0.0807) \simeq 1.084$. Cela signifie une hausse de 8.4% de la cote, *et non* une hausse du risque relatif ou une hausse de 8.4 points de probabilité. La variable décrit une masse cumulative estimée, pas un kilogramme consommé lors de l'année étudiée.

Le coefficient des antécédents familiaux vaut environ 0.9254; à caractéristiques identiques, le passage de `Absent` à `Present` multiplie la cote par $e^{0.9254} \approx 2.52$. Le coefficient de l'âge dans le modèle complet vaut 0.0452 : dix années supplémentaires multiplient la cote par $e^{10 \times 0.0452} \approx 1.57$. La probabilité ajustée pour un individu de caractéristiques x est

$$\widehat{p}(x) = \Lambda(\widehat{\beta}_0 + \sum_j \widehat{\beta}_j x_j). \tag{27}$$

Exemple à partir du modèle complet. Considérons un individu ayant les caractéristiques suivantes :

```
age=52     sbp=160           tobacco=12
ldl=5.73   adiposity=23.11   famhist=Present
typea=49   obesity=25.30     alcohol=97.20
```

On obtient $L(x) \approx 0.906$, donc $\widehat{p} \simeq e^{0.906}/(1 + e^{0.906}) \simeq 0.712$. Il s'agit d'une *probabilité ajustée à l'échantillon*, pas d'un risque absolu utilisable tel quel pour un patient.

L'IC d'un coefficient est, approximativement, $\widehat{\beta}_j \pm 1.96\,\text{se}(\widehat{\beta}_j)$; celui de son OR s'obtient en exponentiant les deux bornes. Par exemple, pour les antécédents familiaux, $0.9254 \pm 1.96(0.2279)$ donne un IC de l'OR d'environ $[1.61; 3.94]$. Un *intervalle de confiance pour la probabilité moyenne* conditionnelle à x ne doit pas être confondu avec un intervalle de prédiction pour un individu, dont l'issue reste binaire.

3.10 Découpage entraînement/test et matrice de confusion

Pour une évaluation convenable, le modèle est appris *uniquement* sur l'échantillon d'entraînement. On peut, pour cette étape, conserver les neuf variables *prédéfinies* afin d'éviter de réutiliser les choix effectués sur toute la base. Les transformations sont elles aussi apprises sur l'entraînement.

	Prédit 0	Prédit 1
Réel 0	VN	FP
Réel 1	FN	VP

Table 8: Matrice de confusion : lignes = vérité, colonnes = prédiction. Dans `sklearn`, l'ordre des classes est ici $[0, 1]$.

La spécificité est $\text{VN}/(\text{VN} + \text{FP})$, la sensibilité est $\text{VP}/(\text{VP} + \text{FN})$, et la *balanced accuracy* est la moyenne de ces deux quantités. Le seuil de 0.5 n'est qu'un choix : en dépistage médical, les faux négatifs peuvent être particulièrement coûteux. Ce choix nécessite une justification métier.

3.11 Courbe ROC et AUC

Pour tout seuil t, le classifieur prédit 1 si $\widehat{p}(x) \geq t$. La courbe ROC représente

$$\text{FPR}(t) = \frac{\text{FP}(t)}{\text{FP}(t) + \text{VN}(t)} = 1 - \text{spécificité}(t), \tag{28}$$

$$\text{TPR}(t) = \frac{\text{VP}(t)}{\text{VP}(t) + \text{FN}(t)}. \tag{29}$$

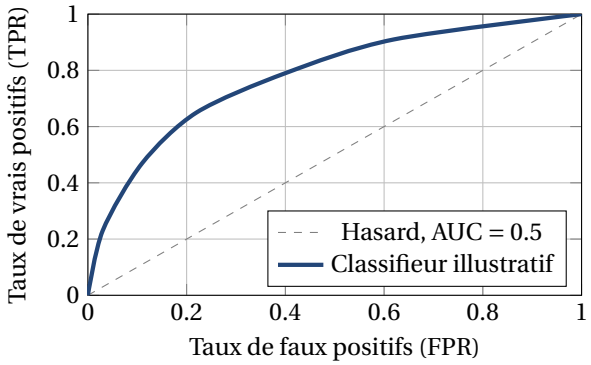


Figure 2: Schéma conceptuel ROC (non calculé sur les données SAheart).

Vérification du notebook : calculer l'AUC à partir des *probabilités* `predict_proba(...)[:,1]`, et non des classes prédites `predict(...)`. La seconde version ne mesure que le comportement du classifieur à un seuil unique et n'est pas l'aire sous la courbe ROC construite avec tous les seuils.

L'AUC est l'aire sous la courbe ROC. Elle est aussi la probabilité, sous conditions usuelles et avec la convention des *ex æquo*, qu'un individu positif tiré au hasard ait un score supérieur à celui d'un négatif. Une AUC proche de 1 indique une excellente discrimination; 0.5 correspond à l'absence de pouvoir de classement.

3.12 Bilan et conclusion

1. **Effets associés.** Sur SAheart, l'âge, le tabac, le cholestérol LDL, les antécédents familiaux et `typea` présentent des coefficients positifs et statistiquement significatifs dans le modèle complet. `sbp`, `obesity`, `adiposity` et `alcohol` ne sont pas significatifs à 5% conditionnellement aux autres variables.
2. **Complexité.** La déviance et le rapport de vraisemblance évaluent le gain d'ajustement; l'AIC et le BIC comparent ajustement et complexité; la sélection automatique peut être instable.
3. **Prédiction.** Le rapport de classification, la matrice de confusion et l'AUC mesurent des aspects de discrimination; les courbes de calibration et les scores probabilistes complètent le diagnostic.
4. **Validité.** Les résultats sont associatifs, potentiellement sensibles aux traitements, à la colinéarité, à la forme linéaire du logit, à l'échantillonnage cas–témoins et au choix du seuil.

Tableau de synthèse

Question	Outil / quantité	Ce qu'il permet (et ne permet pas)
Un coefficient est-il nul ?	Wald z / p-valeur	Association conditionnelle, pas causalité
Un groupe de variables aide-t-il ?	LRT, score, Wald conjoint	Test de contraintes emboîtées
Quel modèle est plus parcimonieux ?	AIC, BIC	Comparaison pénalisée
Le modèle décrit-il les fréquences ?	Résidus, HL, calibration	Adéquation / calibration, avec limites
Classe-t-on bien les individus ?	Confusion, sensibilité, précision, ROC/AUC	Discrimination dépendant du seuil ou non
Les probabilités sont-elles fiables ?	Courbe de calibration, Brier	Qualité probabiliste sur la population évaluée
L'estimation est-elle possible ?	Rang de X, séparation, convergence	Identifiabilité et existence d'un maximum fini