Claire Joigneaux-Desplanques
BL. 2022-2023

Sur la lecture des résultats d'une régression logistique

→ Jérôme Deauviau (de la rue d'Ulm ☺), dans un article paru dans le Bulletin de méthodologie statistique (2010) :

« La régression logistique fait partie des dernières méthodes statistiques importées en sociologie en France. Son introduction dans cette discipline a occasionné deux registres de débats.

Le premier a porté **sur la légitimité même du raisonnement** (toutes choses égales par ailleurs) inhérent aux méthodes de régression multiple, débat qui est d'ailleurs antérieur en sociologie quantitative à la méthode elle-même[1].

Le second est plutôt d'ordre méthodologique puisqu'il porte sur les façons de modéliser lorsqu'on est en présence d'une variable à expliquer catégorielle.

Même si la discussion continue de porter sur le principe même de la régression logistique, il semblerait que cette méthode fasse aujourd'hui partie de la boîte à outils du sociologue quantitativiste. L'objet de cet article est de ce fait résolument méthodologique, et vise à présenter et à **discuter différentes façons de traduire les résultats d'une modélisation logit sous la forme de probabilités**.

En sociologie, on cherche en effet souvent à transformer le résultat d'une modélisation logit en probabilités ou en pourcentages pour au moins deux raisons. La première raison est liée au mode d'utilisation de la régression logistique par les sociologues. Ces derniers utilisent très souvent des variables explicatives catégorielles (sexe, PCS . . .), et lorsqu'ils sont en présence de variables numériques (salaire, âge . . .), l'usage courant veut qu'elles soient **mises en catégories**. Ce choix en est bien un, puisqu'il est tout à fait possible de laisser les variables numériques dans un modèle de régression logistique. Cette pratique de mise en catégories a bien entendu à voir avec le fait que **l'intérêt du sociologue pour les comportements l'amène à croiser plus souvent des variables catégorielles**. **La statistique du sociologue relève de la catégorie, là où celle de l'économiste relève du nombre.** »

1. Rappels sur le principe de construction des régressions logistiques =

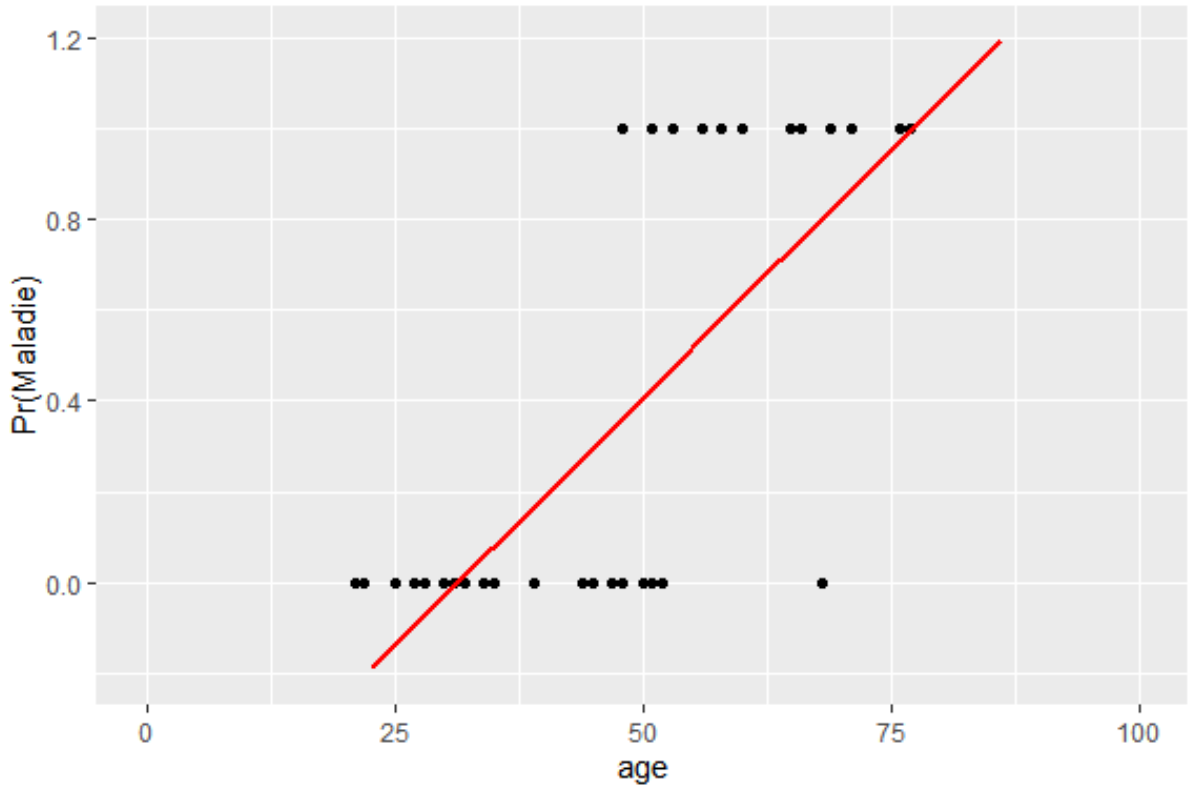
→ On cherche à distinguer l'effet propre de variables indépendantes sur une variable de type binaire (oui/non) Y, qui peut donc prendre deux valeurs 1 (oui l'individu est malade, oui il a voté à l'extrême droite, oui il est obèse, … ou 0 (non il n'est pas malade, …). On code les réponses 1 ou 0.

les k variables indépendantes sont notées ici (et souvent) : X_1, X_2, X_3, ...X_k

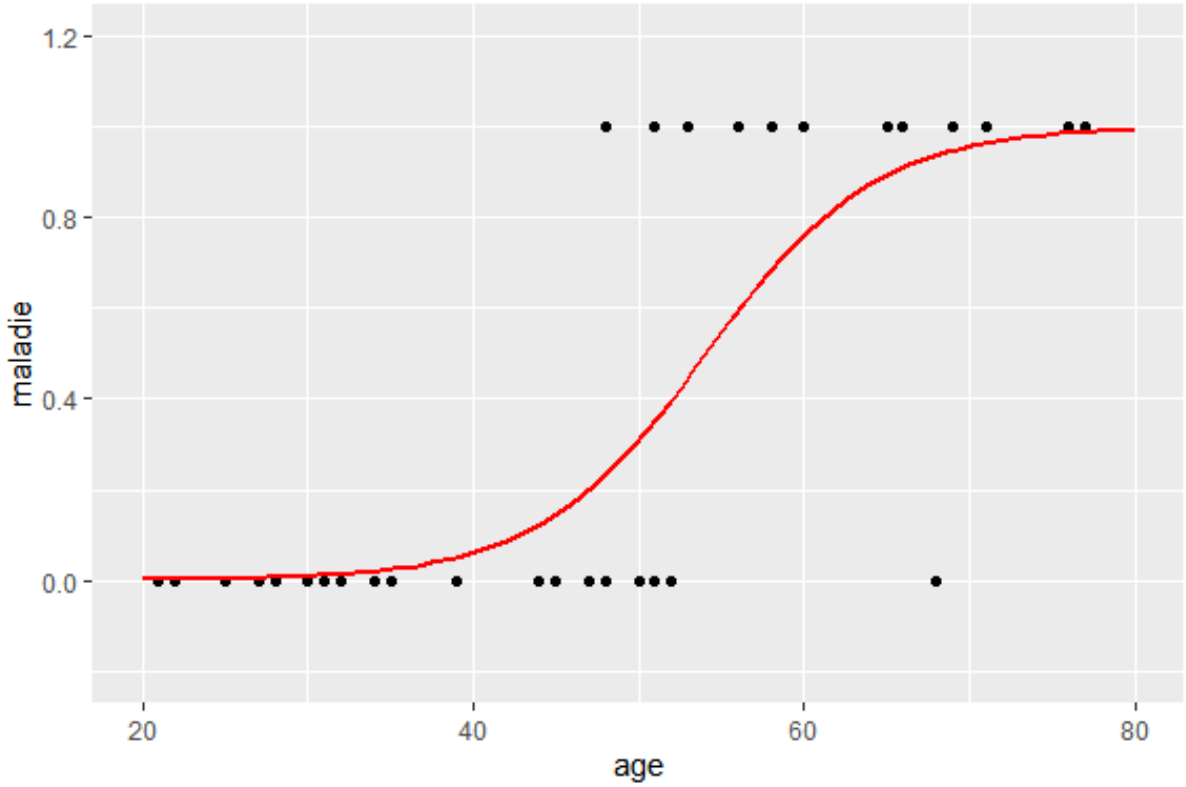
Chacune de ces variables admet différentes modalités. Par exemple pour la variable niveau de diplôme, on peut distinguer différentes modalités comme : sans diplôme/CAP-BEP/Bac/ Bac + 2/ ....
Vous pouvez d'ailleurs être attentifs à la finesse du découpage : comme toujours plus le découpage est précis et plus les catégories sont homogènes mais il renseigne aussi sur les catégories de classement jugées pertinentes par ceux qui ont construit les données et donc sur leurs représentations de l'espace social.
Comme le rappelle Dauvieau, ces variables peuvent être quantitatives ou qualitatives.
Si elles sont qualitatives elles sont catégorielles et si elles sont quantitatives on peut soit conserver une valeur numérique (parfois continue, définie sur un intervalle, par exemple l'âge est rarement négatif) soit les transformer en données catégorielles par exemple en classant els individus de l'échantillon par tranche d'âge avant d'entrer les données qui serviront à modéliser.

On cherche donc à estimer **une probabilité conditionnelle** P(Y = 1/ les valeurs prises par les différentes variables explicatives)

→ **pour modéliser cette relation entre les variables indépendantes et la probabilité pour que Y = 1, on ne cherche pas l'équation d'une droite de régression** (ou une relation linéaire entre la variable dépendante et les k variables indépendantes). En effet, dans le cas d'un modèle à deux variables, la droite de régression résumerait assez mal le nuage de points obtenus et on obtiendrait pour la variable dépendante des valeurs inférieures à 0 ou supérieures à 1, et pour une probabilité c'est un peu ennuyeux ;-).
Comme par exemple ici, où on cherche à établir une relation entre la probabilité de contracter une maladie et l'âge des patients.



Ici, on choisit de modéliser la relation entre les variables indépendantes et la variable dépendante à l'aide d'une courbe de forme sigmoïde et dont toutes les valeurs en ordonnée sont comprises entre 0 et 1.
Comme dans l'exemple qui suit où on cherche à modéliser la relation entre la probabilité d'avoir une maladie donnée et l'âge du patient.



La courbe en rouge résume mieux la forme du « nuage » de points que ne le ferait une droite.

Cette **courbe sigmoïde** est la courbe représentative d'une fonction de répartition
L'avantage d'une fonction de répartition, c'est qu'elle prend toujours des valeurs comprises entre 0 et 1. C'est logique puisque si F est une fonction de répartition associée à une loi de probabilités donnée, F(x) c'est toujours la probabilité que la variable aléatoire qui suit cette loi prenne une valeur inférieure à x. F(x) étant une probabilité, F(x) prend des valeurs comprises entre 0 et 1. CQFD.
Or ce qu'on cherche à modéliser c'est précisément p une probabilité, forcément comprise entre 0 et 1.

Dans les modèles de régression logistique, on choisit **la fonction de répartition de la loi… logistique.**
Cette fonction de répartition est définie de la façon suivante :
F(x) = P (X< x) = $\frac{e^x}{1+ e^x}$ pour tout x appartenant à l'ensemble des réels.
On va voir pourquoi ce choix conditionne la lecture qu'on peut faire des résultats d'une régression logistique

→ **On dispose d'un échantillon** avec des enquêtés pour lesquels on dispose de toute une série de données et donc les modalités que prennent, pour chacun d'entre eux, les k variables indépendantes et bien sûr la valeur prise par la variable dépendante (Y=1 ou Y=0)
On dispose donc d'un **tableau de contingence** avec autant de lignes que d'individus dans l'échantillon et autant de colonnes que de variables intégrées dans le modèle (ici k+1)
On peut construire un nuage de points à k+1 dimensions mais il n'est pas très facile à visualiser.

Chaque point représenterait un individu et les (k+1) coordonnées de ce point correspondraient aux valeurs prises chez cet individu par les différentes variables intégrées au modèle.

→ **A partir de ces données de l'échantillon,** on cherche à modéliser la relation entre la variable dépendante et les différentes variables indépendantes (une ou plusieurs variables d'intérêt et des variables de contrôle). P(Y=1/les valeurs prises par les différentes variables indépendantes)

Si on note βX pour simplifier, une combinaison linéaire de la valeur prise par les différentes variables explicatives.

$$\beta X = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k$$

Cette expression βX prend une valeur différente pour les individus dont les caractéristiques sont différentes.

On cherche à modéliser P(Y = 1/ $\boldsymbol{\beta X}$)
P (Y = 1/ βX) = probabilité que Y = 1 sachant les valeurs prises par chacune des k variables
= la probabilité d'adopter une pratique (ou d'être touché par une maladie , ou de déclarer une opinion,….) sachant toute une série de caractéristiques de l'individu.

Et pour ça on va se servir de la fonction de répartition de la loi logistique pour modéliser les liens entre logit de p (la probabilité conditionnelle qui nous intéresse) et
$$\beta X = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k$$

En effet,

Si p = $\frac{e^{\beta X}}{1+ e^{\beta X}}$ alors $\beta X = \ln\left(\frac{p}{1-p}\right)$ les deux expressions sont strictement équivalentes (et c'est très facile à démontrer par exemple en partant de la première expression

p = $\frac{e^{\beta X}}{1+ e^{\beta X}} \Rightarrow$ p (1 + $e^{\beta X}$) = $e^{\beta X} \Rightarrow$ p = $e^{\beta X}(1-p) \Rightarrow \frac{p}{1-p} = e^{\beta X} \Rightarrow \ln\left(\frac{p}{1-p}\right) = \beta X$

une fois la modélisation construite, si on connaît les caractéristiques d'un individu on connait la valeur prise par βX *pour lui , alors on connait* $\frac{e^{\beta X}}{1+ e^{\beta X}}$ et donc on peut « prédire » la probabilité que Y = 1 étant donné ses caractéristiques.

Ce qui est pratique puisque qu'une fois qu'on connaît le logit d'une probabilité p, il suffit de calculer son exponentielle pour obtenir l'odds associé à cette probabilité

En effet logit p = $\ln\left(\frac{p}{1-p}\right)$ et donc exp (logit p) = $\frac{p}{1-p}$

2. Un exemple de ce qu'on obtient

Si X_1 = revenu de la personne interrogée exprimée en millier d'euros

et X_2 = Sexe de la personne interrogée avec X_2 = 0 si l'individu est une femme et X_2= 1 si l'individu est un homme. Le sexe peut prendre deux modalités = 1 pour les hommes et 0 pour les femmes , c'est une variable « catégorielle » binaire

On cherche à déterminer l'influence de ces deux variables sur la probabilité de voter
P (Y = 1) = probabilité de voter
P (Y=0) = probabilité de s'abstenir

Ici on cherche à déterminer
P (Y=1/ βX) = P(Y = 1/ X_1, X_2) = p (on note p pour simplifier les notations)

On peut avoir comme objectif de comprendre et d'objectiver un gender gap. Dans ce cas la relation d'intérêt c'est la relation entre la probabilité d'aller voter et le sexe de la personne interrogée. Ici on a introduit une variable de contrôle, le revenu de la personne interrogée. On sait ou on croit savoir que
- la probabilité d'aller voter n'est pas totalement indépendante du niveau de revenu
- les femmes sont structurellement un revenu différent de celui des hommes
On veut donc raisonner à revenu équivalent et savoir si le fait d'être une femme dispose à aller voter.

A partir des données de l'échantillon, du nuage de points à 3 dimensions, on obtient le modèle suivant =

Logit VOTE = logit p = ln $\left(\frac{p}{1-p}\right)$
= $\beta X = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ = -1.9 + 0.062REVENU - 0.5SEXE

Avec β_0 = - 1.9. **C'est la constante qu'on voit parfois apparaître dans les tableaux avec les coefficients de la régression logistique**

Ici le revenu c'est le revenu annuel exprimé en milliers d'euros

βX peut bien prendre des valeurs très variées en fonction des revenus des personnes interrogées = la valeur minimale c'est -1.9 (une femme sans revenu) et la valeur maximale dépend du revenu maximal observé dans l'échantillon.
Le modèle prédit pour des femmes sans revenu un logit de 0.149

Pour un homme qui gagne 38 000 euros, le modèle prédit un logit
Logit p = -1.9 + 0.062x38 – 0.5 x1 = 0 et donc P/(1-P) = e^0 = 1

Pour une femme , qui gagne elle aussi 38000 euros, le modèle prédit un logit p = -1.9 + 0.062x38 – 0.5x 0 = 0.5 et donc P/(1-P) = $e^{0.5}$ = 1.5

Toutes choses égales par ailleurs, une femme a 1.5 fois plus de chances de voter (plutôt que de s'abstenir) par rapport aux hommes. C'est l'odds-ratio associé à la modalité femme (par rapport aux hommes qui servent ici de référence) d'après les résultats de cette régression,

dans cette échantillon, étant donné les variables indépendantes (variable d'intérêt et variables de contrôle) qu'on a introduites dans le modèle.

Remarque, on peut la même chose pour la variable continue qu'est le revenu.
Si le revenu augmente de 1000 euros, le logit de p augmente de 0.062 et la probabilité d'aller voter (plutôt que de s'abstenir), p/(1-p), est multiplié par expo.062 et donc par 1.063 la probabilité d'aller voter plutôt que de s'abstenir.
(remarque : il est peu probable que l'effet soit parfaitement régulier à chaque augmentation de 1000 euros du revenu. C'est d'ailleurs la raison pour laquelle, on découpe souvent des tranches de revenu plutôt.

Remarque = on aurait pu construire des **régressions emboîtées** en intégrant d'autres variables explicatives comme par exemple le diplôme : sachant que les femmes sont plus diplômées en moyenne que les hommes, on pourrait involontairement mesurer à travers la corrélation entre le sexe et la participation électorale l'effet d'une variable cachée qui serait le diplôme. Pour bien isoler le poids de la variable sexe, pour bien distinguer ce qui s'explique par des expériences différentes spécifiques aux femmes qui les disposeraient à voter plus (plus le souci de l'intérêt général, du care, normes civiques intériorisées davantage, souci de donner l'exemple aux plus jeunes, …), on pourrait penser que l'expérience différente qu'elles ont faite dans le système scolaire ou universitaire peut participer aussi à construire un rapport à la politique et à la participation politique différent. C'est cette hypothèse qu'on pourrait tester en intégrant au modèle la variable diplôme.

3. A retenir pour la lecture des données

Selon les cas, la présentation des résultats de la régression peut prendre des formes différentes et dans tous ces cas, les notes de lecture sont utiles pour éclairer votre lanterne.

→ Si les résultats sont présentés sous forme de coefficients de régression associés aux différentes modalités =

- Les coefficients mesurent l'influence d'une modalité sur le logit de la probabilité qu'on cherche à établir.

Problème : le logit n'est pas très parlant. On sait seulement qu'un logit de zéro correspond à une probabilité p de 0.5 et que plus le logit de p augmente, plus la probabilité p augmente.

- si le coefficient est négatif, la modalité diminue la probabilité d'adopter le comportement qu'on cherche à modéliser. Si le coefficient est positif c'est le contraire. Si le coefficient est nul, la modalité n'a aucun effet sur le comportement étudié.
Ici , le fait d'être un homme diminue la probabilité d'aller voter. (coefficient = -0.5) et quand le revenu augmente, la probabilité augmente aussi.
- plus la valeur ajoutée du coefficient est élevée plus la modalité modifie fortement la probabilité d'adopter le comportement étudié.
- et surtout, si vous avez une calculatrice, vous pouvez calculer précisément l'odds associé à chaque modalité, vous pouvez commenter directement sous cette forme plus parlante les résultats de la régression.

En effet, en calculant exp (coefficient), on obtient une indication plus précieuse. Ca nous permet de comparer deux odds et donc d'exprimer un odds-ratio.

→ si les résultats sont exprimés directement en odds-ratio

- Si l'odds-ratio associé à une modalité est supérieure à 1, c'est que la modalité augmente la probabilité d'adopter le comportement étudié (c'est logique parce que le coefficient est supérieur à 0 est que l'exponentielle d'un réel positif c'est toujours supérieur à 1)
- Si l'odds-ratio est inférieur à 1, c'est qu'il diminue la probabilité d'adopter le comportement étudié (c'est logique parce que le coefficient est inférieur à 0 est que l'exponentielle d'un réel négatif c'est toujours inférieur à 1).

→ si les coefficients sont exprimés en odds-ratios par rapport à un groupe de référence clairement identifié

- **Par définition, l'odds-ratio du groupe de référence est égal à 1**
- L'odds ratio associé à une modalité (être une femme par exemple) c'est en général le rapport de
  → l'odds pour un groupe d'intérêt constitué de femmes qui ont toutes les caractéristiques du groupe de référence à l'exception du fait qu'elles sont des femmes
  → l'odds du groupe de référence (un groupe d'intérêt particulier défini par des modalités particulières pour chacune des variables indépendantes et qui sert de référence pour calculer l'ensemble des odds-ratios avec tous les autres groupes d'intérêt)

Pour lire les données de façon parlante, en cas d'odds ratio inférieur à 1 on a intérêt à calculer l'inverse de l'odds-ratio.

Exemple =
avec un odds ratio associé au fait d'être jeune de 0.5 = on a intérêt à dire que, toutes choses égales par ailleurs, les chances d'adopter le comportement étudiés sont divisées par 2 chez les jeunes (plutôt que de dire qu'elles sont multipliées par 0.5)

→ bien entendu, vous devez **rester attentif aux seuils de significativité** de la même façon que pour les coefficients.

Rappel = Les seuils de significativité indiquent la probabilité que le coefficient de régression (trouvé à partir des données de l'échantillon) soit en fait nul (et donc que la modalité n'ait aucun effet). La significativité des coefficients est parfois indiquée entre parenthèse à côté du coefficient, ou dans une colonne à part. Comme il s'agit d'une probabilité elle est toujours comprise entre 0 et 1 et sauf si c'est indiqué elle n'est pas exprimée en pourcentage.

Plus ce seuil de significativité est faible et plus le coefficient qu'on a trouvé en construisant la régression est significatif, plus la mesure est robuste, plus vous pouvez en déduire que la modalité associée a bien dans la réalité l'effet qui est estimé par le modèle.

On distingue parfois des seuils de significativité =

- au seuil de 1% (noté parfois ***) = la probabilité que le coefficient soit nul est inférieure à 0.01 (1%)
- au seuil de 5% (noté parfois **) = la probabilité que le coefficient soit nul est inférieure à 0.05 (5%)
- au seuil de 10% (noté parfois *) = il y a alors moins de 10% de chance que le coefficient soit nul, la probabilité qu'il soit nul et que donc la modalité n'ait aucune influence sur le comportement étudié est inférieure à 0.1.

Annexe = un exemple d'utilisation

Comment traduire sous forme de probabilités les résultats d'une modélisation logit ?

Source = Jérôme Deauvieau
Bulletin de méthodologie statistique, 2010

Il utilise des données de l'enquête FQP 2003 pour modéliser la probabilité de devenir cadre et donc d'intégrer la PCS3 : L'échantillon sur lequel est appliqué ce modèle est composé des 1.237 individus de l'enquête FQP 2003, qui étaient professions intermédiaires du privé en 1998, et qui soit sont restés professions intermédiaires du privé (PI), soit sont devenus cadres en 2003.

$$\ln \frac{\Pr(Y = cadre)}{1 - \Pr(Y = cadre)}$$

Ce logit constitue le membre de gauche de l'équation de régression.

On trouve à droite une équation linéaire formée des variables explicatives.

$$\ln \frac{\Pr(Y = cadre)}{1 - \Pr(Y = cadre)} = \mathrm{B0} + \mathrm{B1diplome} + \mathrm{B2sexe} + \mathrm{B3age\ 2} + \mathrm{B4age\ 3}$$

Chaque variable explicative introduite dans le modèle est mise sous forme dichotomique, avec un codage en 0 ou 1 ((présence) / (absence)). Par exemple, pour le sexe, on a choisi de coder les femmes en 0 et les hommes en 1.

Pour les variables à plus de deux modalités, on forme autant de variables dichotomiques qu'il y a de modalités dans la variable et on réalise ainsi un codage disjonctif complet. Par exemple, la variable âge, qui a trois modalités, sera transformée en trois variables dichotomiques âge 1 (code 1 si âge inférieur à 35 ans, sinon 0), âge 2 (code 1 si âge compris entre 35 et 45 ans, sinon 0), âge 3 (code 1 si âge supérieur à 45 ans, sinon 0).

On remarquera qu'il manque la variable dichotomique age1 dans le modèle. L'explication est simple: pour représenter trois situations différentes, deux variables dichotomiques suffisent. Le tableau suivant présente ce raisonnement (Tableau 1). Pour représenter la variable A, qui a trois modalités, l'information contenue dans les deux premières variables (modalités 1 et 2) suffit, puisqu'il est possible de de terminer la valeur prise par la modalité 3 à partir des valeurs prises par les modalités 1 et 2.

Tableau 1. Le codage disjonctif complet

Variable	Modalité 1	Modalité 2	Modalité 3
1	1	0	0
2	0	1	0
3	0	0	1

$$\ln\frac{\Pr(Y = cadre)}{1 - \Pr(Y = cadre)} = \mathrm{B0} + \mathrm{B1diplome} + \mathrm{B2sexe} + \mathrm{B3age\ 2} + \mathrm{B4age\ 3}$$

On trouve avec les données de l'échantillon, une estimation de chacun des coefficients de régression. Ici, par exemple, la valeur estimée de B_1 c'est 0.75

$$\ln\frac{\Pr(Y = cadre)}{1 - \Pr(Y = cadre)} = -1,95 + 0,75 * \mathrm{diplôme} + 0,59 * \mathrm{sexe} - 0,29 * \mathrm{age\ 2} - 0,63 * \mathrm{age\ 3}$$

Il n'est pas nécessaire de calculer le logit de l'ensemble des situations pour extraire l'information pertinente d'un modèle logit. Chaque coefficient correspond à l'augmentation ou à la diminution du logit lorsque l'on passe du code 0 au code 1 de la variable adossée à ce coefficient.
Par exemple, **pour le sexe, le modèle indique que le logit de la probabilité de devenir cadre augmente de 0,59 lorsque l'on passe de la situation femme à la situation homme.**

Cette augmentation de 0.59 est <u>la même toutes choses égales par ailleurs</u>, c'est-à-dire, quelle que soit la configuration des autres variables introduites dans le modèle, ou, ce qui veut dire la même chose, quelles que soient les modalités pour les autres variables. En d'autres termes, si on compare le logit des hommes de moins de 35 ans, diplômés du supérieur, au logit des femmes de moins de 35 ans, diplômées du supérieur, **on trouvera un écart de 0,59**, et il en sera de même si on compare hommes et femmes chez les plus de 45 ans n'ayant pas le bac (vous pouvez vérifier en utilisant les données du tableau 2)

Tableau 2. Calcul des logit du chaque situation

Situation	logit
Pas le bac, femme, age 1	−1,95
Pas le bac, femme, age 2	−2,24
Pas le bac, femme, age 3	−2,58
Pas le bac, homme, age 1	−1,36
Pas le bac, homme, age 2	−1,64
Pas le bac, homme, age 3	−1,99
Bac, femme, age 1	−1,20
Bac, femme, age 2	−1,48
Bac, femme, age 3	−1,83
Bac, homme, age 1	−0,60
Bac, homme, age 2	−0,89
Bac, homme, age 3	−1,23

Source: INSEE, enquête FQP, 2003.

Tableau 3. Expliquer le passage à la catégorie cadre

	Coefficient	
Constante	−1,95	Test
Sexe		
Femme	ref	
Homme	0,59	(p<0,01)
Diplôme		
Inférieur au bac	ref	
Supérieur au bac	0,75	(p<0,01)
Age		
Age 1	ref	
Age 2	−0,29	(p=0,22)
Age 3	−0,63	(p<0,01)

Source: INSEE, enquête FQP, 2003.

Ensuite si on veut transformer ces coefficients, on peut aboutir à des odds ratio en utilisant une propriété de la fonction logarithme et une autre de la fonction exponentielle (elle est bijective)

$$\ln\frac{P1}{1-P1} - \ln\frac{P2}{1-P2} = 0,59,$$

Et donc, d'après les propriétés de la fonction logarithme que vous connaissez bien

$$\ln\frac{\dfrac{P1}{1-P1}}{\dfrac{P2}{1-P2}} = 0,59$$

$$\exp\left(\ln\frac{\dfrac{\mathrm{P1}}{1-\mathrm{P1}}}{\dfrac{\mathrm{P2}}{1-\mathrm{P2}}}\right) = \exp\,(0{,}59)$$

$$\text{donc}\ \frac{\dfrac{\mathrm{P1}}{1-\mathrm{P1}}}{\dfrac{\mathrm{P2}}{1-\mathrm{P2}}} = 1,80.$$

On peut donc en déduire un odds-ratio : le fait d'être un homme multiplie par 1.8, toutes choses égales par ailleurs, la probabilité de devenir cadre plutôt que de rester profession intermédiaire du privé