INTRODUCTION

A LA REGRESSION LOGISTIQUE

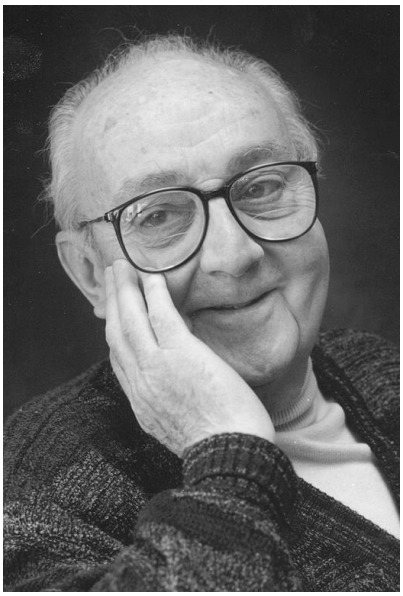
Support de formation (version 1 au 20-04-2022)

26 AVRIL 2022

Marc Thévenin – Arno Muller

« All models are wrong, but some are useful »

Georges Box (1919-2013)*



Pour un résumé des discussions autour de cette fameuse phrase :
https://en.wikipedia.org/wiki/All_models_are_wrong

* Accessoirement, était également le beau-fils de Ronald Fisher

Table des matières

INTRODUCTION

De la mesure d'une variable binaire à une mesure d'association 5

Au-delà d'un simple tableau croisé…. la régression 7

Le modèle logistique : un peu d'histoire 10

Les domaines d'utilisation régression logistique 11

Plan de la formation 12

Bibliographie (accessible web) 12

CONCEPTS

Probabilité et Odds 14

Probabilité 14

Odds 15

Rapport de probabilités et Odds ratio 16

Le modèle linéaire 17

Rappels 17

Le modèle de probabilité linéaire 20

LE MODELE LOGISTIQUE

Approche descriptive et approche causale/explicative 23

Approche descriptive 23

Approche causale/explicative 26

Estimation du modèle 27

INTERPRETATION DES RESULTATS

Introduction et rappels sur l'inférence 31

Interprétation des résultats dans une régression multiple 35

Variable quantitative 36

Variable discrète/catégorielle 37

Interactions 40

Retour à l'échelle des probabilités: les effets marginaux 48

TP: Et l'ajout de l'IMC ça change quoi? 52
Test et indicateurs reposant sur la vraisemblance 53
Classification : concordances et courbe de ROC 59

PROGRAMMATION

R 65
Estimation et lecture du modèle 65
Autres mise en forme de l'output 72
jtools 72
Gtsummary 76
Qualité du modèle 78
Introduction de pondérations avec estimation robuste 80
STATA 82
Estimation du modèle avec logit 82
Analyse de la qualité du modèle 88
Introduction de poids de sondage (estimation robuste) 91

INTRODUCTION

De la mesure d'une variable binaire à une mesure d'association

En statistique, on peut distinguer deux grands types de variables : celles de type quantitatives dont la valeur est directement mesurable, et celles de type discrètes, non directement mesurables. Ces variables sont souvent appelées variables catégorielles ou qualitatives. A la suite d'une opération de regroupement, une variable quantitative peut-être discrétisée.

Durant cette formation, nous allons nous intéresser au cas limite d'une variable de type discrète dont l'information quelle donne repose sur un signal ou à l'appartenance à un groupe. Nous nous limiterons à des situations ou états dits *binaires* ou *dichotomique*.

Quelques exemples de problématique :

- Etre en emploi ou non
- Avoir ou non un ou plusieurs enfant
- Faire ou avoir fait des études supérieures
- Etre propriétaire ou non de son logement (ou tout autre statut d'occupation)
- Avoir ou non migré dans une autre région ou dans un autre pays au cours d'une période
- Avoir ou non fait l'objet de violence
- Avoir ou non le sentiment d'avoir été discriminé dans l'emploi ou dans un autre domaine en raison d'un des 25 critères définis par la loi.

L'expression commune à ces situations binaires est l'usage plus ou moins explicite d'un **oui/non**. Il s'agit donc d'un signal de type $vrai/faux$. On parle souvent de *variable booléenne* dans le jargon informatique. Il nous faut donc traduire cette information en valeurs qui nous permettra, à minima, de compter… et donc de pratiquer la statistique. **La traduction numérique de ces situations est tout simplement le couple de valeur** $\{0,1\}$.

Ce qui suit est assez important, et préfigure beaucoup d'éléments de la formation. Prenons une première variable avec des observations totalement fictives, par exemple sur fait de "suivre ou d'avoir suivi des études supérieures" [variable y].

- On a 10 observations: $y_i = (non, oui, oui, oui, oui, non, non, oui, non, oui)$ qu'on peut donc traduire numériquement par $y_i = \{0,1,1,1,1,0,0,1,0,1\}$.
- On en tire facilement un nombre de personnes du supérieur $\sum y_i = 6$ et sa proportion $\frac{6}{10} = 0.6$.
- On en déduit directement le nombre et/ou la proportion de personnes qui n'ont pas suivies d'études dans le supérieur, qui s'élève à 0.4.

Prenons maintenant une nouvelle variable caractérisant de manière toute aussi fictive une autre caractéristique des individus (x). Cette variable peut prendre 3 états *A*, *B* et *C*.

Pour les 10 observations: $x_i = (A, B, B, C, A, A, B, B, C, A)$. Comment traduire les 3 états de cette "variable" en valeurs numériques ? On pourrait présenter cette information avec un codage numérique, par exemple 0 pour A, 1 pour B et 2 pour C. Cela ne change rien, on ne saura toujours pas compter le nombre de personnes dans chaque état.

Pour une information binaire ou dichotomique, la transformation $\{0,1\}$ va de soi. Si on a plus de deux affectations possible, on peut également transformer l'information d'origine en plusieurs variables, appelées *indicatrices*, *dummy* ou *variable muette*, une pour chaque état, avec comme valeurs numériques $(0,1)$:

- Etre ou non A (x_1): 1 si oui 0 sinon.
  - $x_{1,i} = \{1,0,0,0,1,1,0,0,0,1\}$ et $\sum x_{1,i} = 4$ avec une proportion associée de 0.4.
- Etre ou non B (x_2): 1 si oui 0 sinon.
  - $x_{2,i} = \{0,1,1,0,0,0,1,1,0,0\}$ et $\sum x_{2,i} = 4$ avec une proportion associée de 0.4.
- Etre ou non C (x_3): 1 si oui 0 sinon.
  - $x_{3,i} = \{0,0,0,1,0,0,0,0,1,0\}$ et $\sum x_{3,i} = 2$ avec une proportion associée de 0.2.

On est allé jusqu'à construire les 3 variables indicatrices associées à *x*, mais ce n'était pas nécessaire. Connaissant l'effectif total de l'échantillon et la répartition dans deux modalités, la troisième est directement obtenue: Par exemple $\sum x_{3,i} = 10 - 4 - 4 = 2$ De la même manière, pour la variable y nous n'avions pas eu besoin d'une variable complémentaire indiquant le fait de ne pas avoir fait d'étude supérieure pour récupérer son total . Toute l'information était contenu dans la première variable indicatrice ($10 - 6 = 4$).

On introduit une notion de **degré de liberté**, centrale en statistique, ici simplement utilisé pour mesurer la fréquence d'un évènement discret.

Question : on veut maintenant connaître la **répartition conditionnelle** du tableau suivant, soit pour chaque modalité de *x*, le nombre personnes qui suivent où ont suivi des études supérieures. On connait les *répartitions marginales* (sous-totaux ou marges) des variables y et (x_1, x_2, x_3), mais cette information est-elle suffisante pour connaitre précisément la répartition des effectifs dans les deux dimensions croisées?

	$y = 0$	$y = 1$	**Total y**
$x_1 = 1$	-	-	**4**
$x_2 = 1$	-	-	**4**
$x_3 = 1$	-	-	**2**
Total x	**4**	**6**	**10**

Réponse : Non

Même question, mais on sait que pour x_1==1 2 personnes n'ont pas suivi d'études supérieures (y=0). L'information est-elle suffisante ?

	$y=0$	$y=1$	Total y
$x_1=1$	2	-	**4**
$x_2=1$	-	-	**4**
$x_3=1$	-	-	**2**
Total x	**4**	**6**	**10**

Réponse : Non

Toujours la même question, mais on sait que pour x_3==1, 1 personne a suivi des études supérieures (y=1). L'information est-elle suffisante ?

	$y=0$	$y=1$	Total y
$x_1=1$	2	-	**4**
$x_2=1$	-	-	**4**
$x_3=1$	-	1	**2**
Total x	**4**	**6**	**10**

Réponse: oui
On peut retrouver par simples différences l'ensemble des fréquences conditionnelles

	$y=0$	$y=1$	Total y
$x_1=1$	2	2	**4**
$x_2=1$	1	3	**4**
$x_3=1$	1	1	**2**
Total x	**4**	**6**	**10**

On retrouve bien le calcul du nombre de degré de liberté d'un test d'indépendance du Khi2: Si p est le nombre d'indicatrices en ligne du tableau et k le nombre de colonne, le nombre de degrés de liberté est égal à $(k-1)\times(p-1)$. Soit le produit du nombre d'informations nécessaires pour les lignes et les colonnes.

En multipliant les effectifs du tableau par 10, on peut réaliser un test du khi2 à 2 degrés de liberté ($p=0.04$), comparer les proportions de personnes qui ont suivies des études supérieures pour les trois niveaux de la variable *x* ($p1=0.5$, $p2=0.75$ et $p3=0.75$). Par la régression linéaire ou une analyse de la variance, la probabilité que les observations contredisent l'hypothèse d'égalité des 3 proportions est inférieure à 5% ($p=0.04$).

Au-delà d'un simple tableau croisé…. la régression

D'autres facteurs peuvent avoir une influence sur le fait d'avoir suivi des études supérieures : la génération, la csp des parents, l'origine ou la nationalité, la zône de résidence…. Comment rendre compte de tous ces facteurs sachant que croisés entre eux ils peuvent être corrélés et donc comporter des effets dits de structure ?

Lorsqu'on souhaite prédire des probabilités et/ou analyser leurs écarts, l'approche par la régression multiple est une technique qui permet d'intégrer plusieurs facteurs, de tenir compte de leur structure croisée. Le modèle logistique, qui s'appuie sur une loi de probabilité est devenu la méthode la plus répandue pour réaliser ce type d'analyse.

Le cas typique et extrême du populaire « paradoxe de Simpson »

- Attention il ne s'agit pas vraiment d'un paradoxe, juste que le sens de la conclusion va s'inverser en raison de l'omission d'un facteur d'analyse.

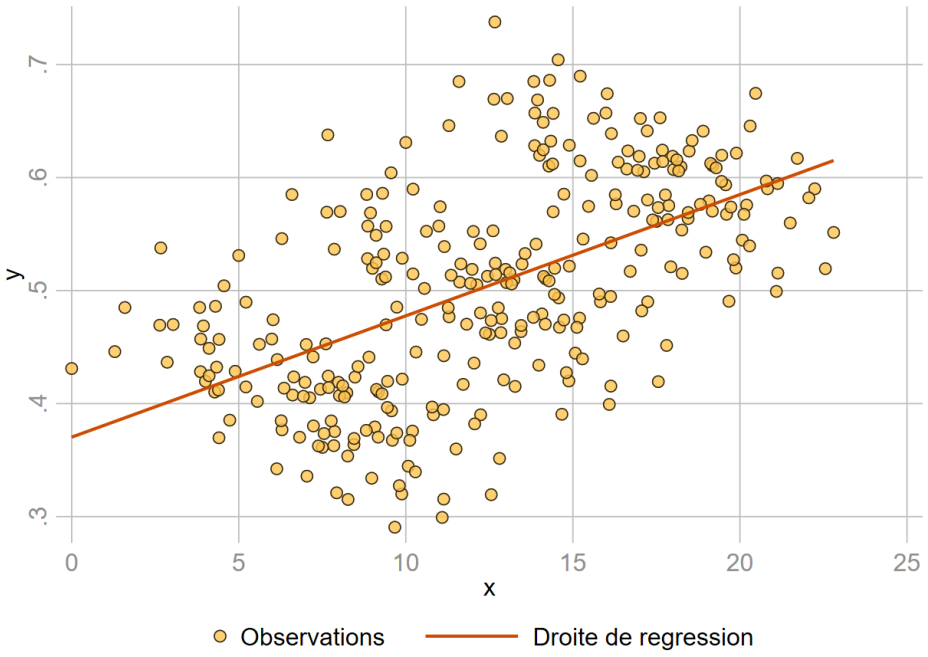
- Traduit des effets de sélection ou de biais d'omission, dans le cas particulier d'une variable dite « confondante » (corrélée directement à la variable qu'on analyse et à une variable supposée l'expliquer).

- Un des exemples les plus connus : taux d'admission entre les filles et les garçons à l'université de Berkeley.

- Les filles présentent un taux de réussite moyen pour l'admission à l'université inférieur à celui des garçons (une trentaine de % contre une quarantaine de mémoire). L'université a été accusée de discriminer les filles.
- Cependant par filière/département de l'université, les % d'admission pour les filles n'est jamais inférieur à celui des garçons et, à quelques exceptions, toujours supérieur à celui des garçons. D'où le « paradoxe ».
- Pourquoi ? Les filles postulent plus que les garçons dans les filières les plus sélectives qui présentent donc, en moyenne des taux de réussite plus faible. Et inversement pour les garçons. Le résultat global/moyen en défaveur des filles est donc le résultat d'une sélection différenciée du choix de la filière. Il n'y aurait donc pas ce « paradoxe » si la sélectivité entre filières était uniforme. Petite remarque : les filières étaient sélectives parce qu'elles étaient dotées différemment par l'université, et ne pouvaient accepter que peu de candidat.e.s… plutôt dans les sciences sociales d'ailleurs…
- Si on faisait un modèle en intégrant à la fois le sexe et la filière, les chances de succès ne sont plus défavorables aux filles. La variable filière est dite « confondante », la variable sexe prise en compte seule est dite « endogène », et elle ne devient « exogène » que lorsqu'on lui associe le type de filière.

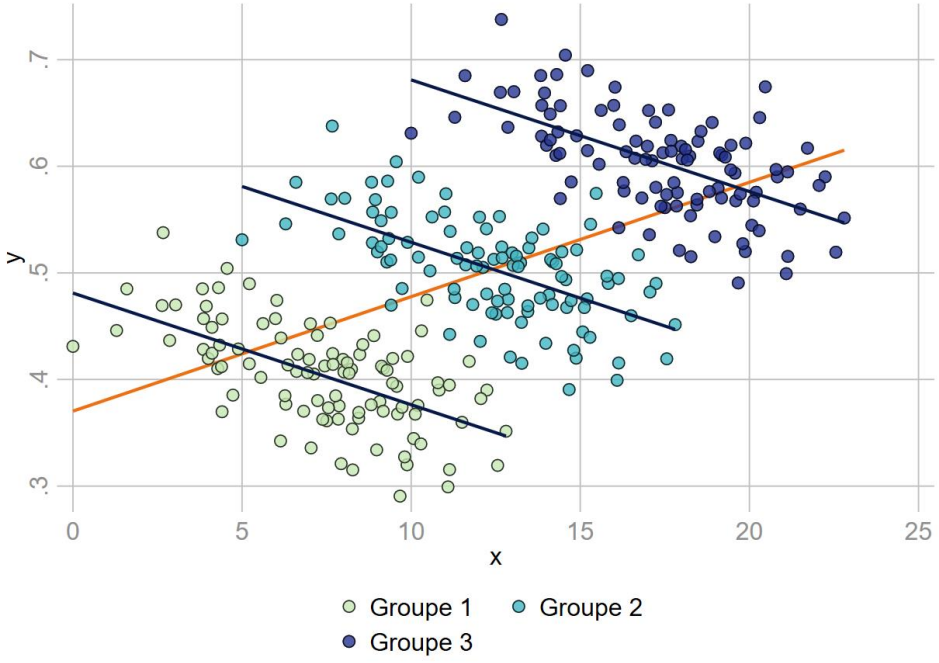
Une illustration visuelle (classique, on la retrouve partout). On va regarder la corrélation entre deux variables continue y et x, plus pratique pour une représentation graphique. Les valeurs ont été simulées.

On regarde dans un premier temps la corrélation entre les variables y et x. Il ne fait aucun doute qu'elle est positive.

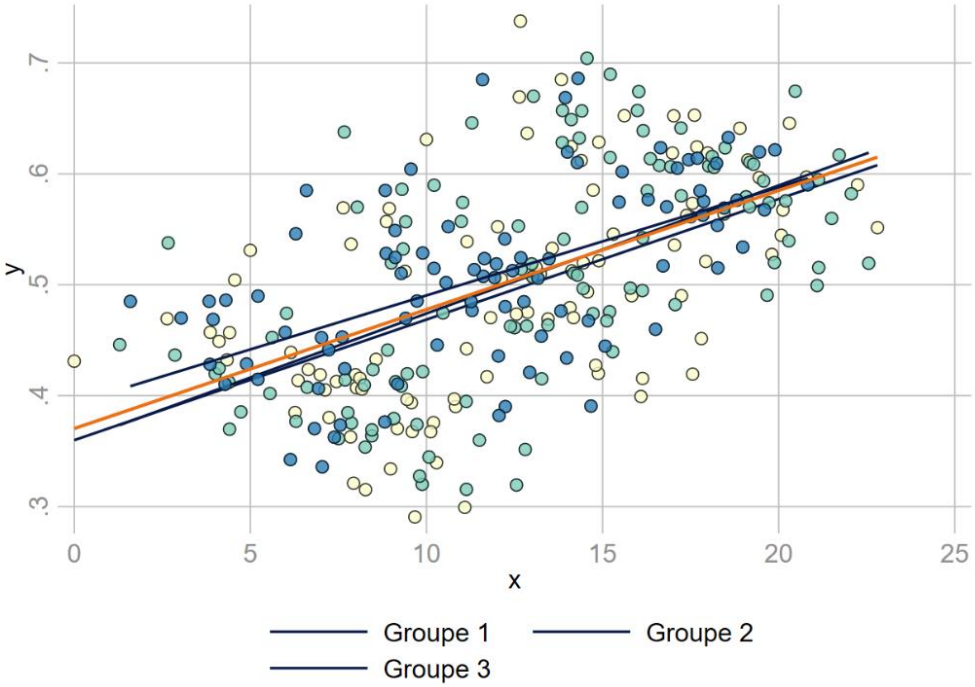


On introduit une nouvelle variable dans le graphique, de type catégorielle, qui donne l'appartenance à 3 groupes.

Cela change du tout au tout pour la variable x. Dans chaque groupe, la liaison entre y et x est négative, le lien positif entre les deux variables numériques poolées n'était qu'apparente.



Si l'appartenance à un des 3 groupes étaient parfaitement aléatoire, le lien positif entre x et y serait pertinent, x serait exogène, et il n'y aurait aucun intérêt à introduire le facteur groupe dans l'analyse :



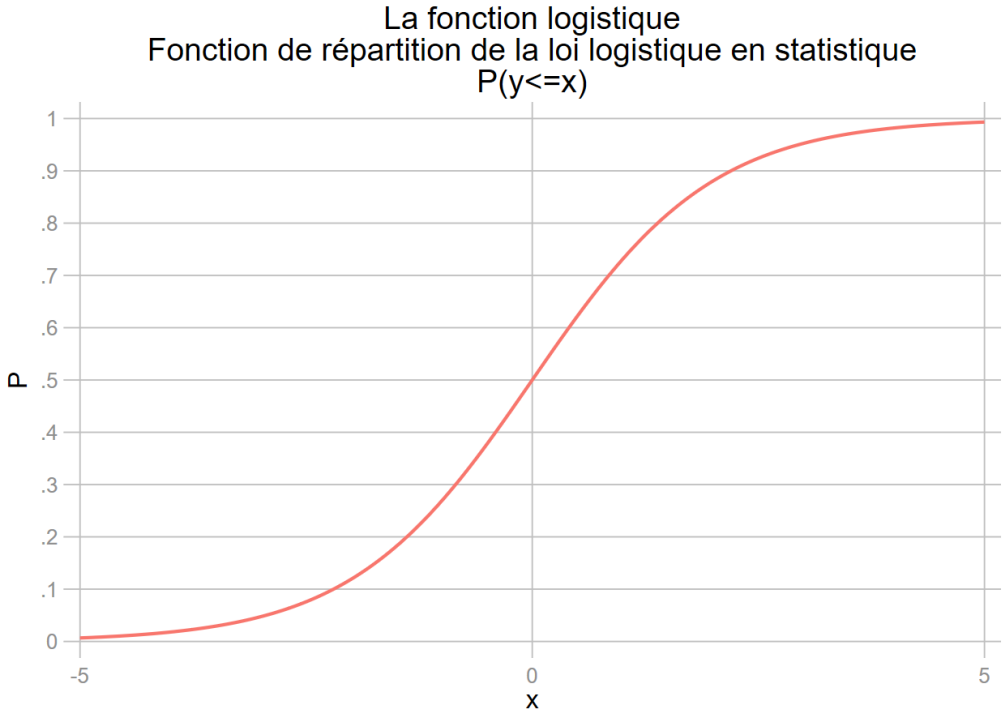
Le modèle logistique : un peu d'histoire

- A la fin du 19ème siècle début du 20ème : pour les variables quantitatives on connait la régression linéaire et la méthode des moindres carrés ordinaires. Pour les variables discrètes/qualitatives le test dit d'indépendance du Khi2 est proposé par K.Pearson.
- Dans les années 20, Ronald Fisher développe la méthode du maximum de vraisemblance, qui démultiplie la boite à outils des statisticien.ne.s: calcul d'estimateurs et analyse de leurs propriétés, tests …
- Années 30 : en s'appuyant sur la méthode du maximum de vraisemblance, le modèle appelé **probit** est proposé pour estimer une probabilité conditionnelle à partir d'une technique de régression. Ce modèle se voit rapidement popularisé en biométrie et en psychométrie. Problème : si on peut estimer les probabilités conditionnelles, les paramètres estimés par le modèle ne sont pas directement interprétables en termes d'écarts - absolus ou relatifs - de probabilité. Ce modèle reste encore très utilisé, en particulier par les économistes.
- A l'origine dans une analyse des dynamiques des populations, au 19ème siècle, un mathématicien belge PF.Verhulst trouve une fonction de type sigmoïde (phase d'accélération => phase constante => phase de décélération) qu'il nomme **fonction logistique**. Cette fonction sera redécouverte par deux statisticiens au début du 20ème siècle dans un cadre probabiliste. Dans les années 40, J.Berkson défend l'utilisation de

cette forme fonctionnelle pour estimer des probabilités conditionnelles. La fonction de répartition de la loi normale (modèle probit) est remplacée par celle de la loi logistique standard. Son avantage est direct, les paramètres estimés sont directement interprétables et la méthode pourra être généralisée à des situations non binaire/dichotomique (plus de 2 classes d'appartenance).

La formulation de la fonction logistique standard est égale à :

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



Les domaines d'utilisation régression logistique

- Dans tous les champs scientifiques : biomédecine, épidémiologie, psychométrie, et dans les sciences sociales en générale…
- Depuis ses débuts dans les techniques dites d'intelligence artificielle (deep learning, réseau de neurones…).
- Peut-être étonnant au premier abord, dans les sciences criminalistes appelées également *forensique*. Dans les expertises judiciaires ce qui est appelée la charge ou le poids de la preuve (*weight of evidence*) est mesuré par une transformation logarithmique d'un rapport de probabilités…ce qu'on va appeler par la suite un *logit*.

Plan de la formation

- **Définition des concepts de base**
  - Probabilité, Odds, et Odds Ratios.
  - Quelques rappels sur la régression linéaire, les problèmes posés par la régression linéaire dans l'analyse d'une variable discrète.
- **Identification du modèle logistique**
  - Approche descriptive
  - Approche causale
  - Estimation des paramètres (utile pour l'analyse de la qualité du modèle)
- **Interprétation des résultats**
  - « Reference coding » versus « effect/deviation coding »
  - Rappels sur la significativité des résultats
  - Odds Ratio : cas d'une variable continue et d'une variable catégorielle/discrète
  - Croisement et interactions
  - Retour à l'échelle des probabilité s: l'effet marginal
- **Analyse de la qualité du modèle**
  - Indicateurs et tests basés sur la vraisemblance
  - Classification : concordance et courbe de ROC

Bibliographie (accessible web)

Cedric Afsa (Insee)

Incontournable (a inspiré ce document) : https://www.insee.fr/fr/statistiques/2022139
La seconde partie est exclusivement consacrée à une application

Laurent Rouvière (Université Rennes 2)

Matheux mais clair - avec applications R – Généralisation aux modèles polytomiques – bien plus de points abordés qu'ici

Document : https://lrouviere.github.io/doc_cours/poly_logistique.pdf
Slides : https://perso.univ-rennes2.fr/system/files/users/rouviere_l/chapitre2_glm.pdf

Les grands débats

C.Moods

Article encore débattu – très technique – aborde la question des comparaisons et des interprétations causales (accès selon l'institution)

https://academic.oup.com/esr/article-abstract/26/1/67/540767

ASA et p-value

Déclaration (« statement ») de l'ASA* sur l'interprétation et l'utilisation des « p-value »… **à lire absolument**

https://www.tandfonline.com/doi/full/10.1080/00031305.2016.1154108

En milieu de page : « ASA Statement on Statistical Significance and *P*-Values (Ronald L. Wasserstein). Le dossier permet également d'accéder à suite de commentaires de statisticien.n.es

* *American Statisticians Association*.

CONCEPTS

Probabilité et Odds

Probabilité

Soit une expérience discrète : lancer d'une pièce ou d'un dé pour les cas d'école, ou tout autre évènement comme ceux listés dans l'introduction (être propiétaire ou non, …). Si a est une réalisation de cet évènement ou expérience, pour les cas d'école: $a_i = (pile, face)$ dans le cas d'un lancer de pièce ou $a_i = \{1,2,3,4,5,6\}$ dans le cas d'un lancer de dé.

Propriétés :

- $0 \leq P(y = a_i) \leq 1$
- $\sum P(y = a_i) = 1$

On l'a déjà vu en introduction, on en tire donc la propriété que pour tout $j \neq i$,

$$p(y = a_j) = 1 - \sum p(y = a_{i \neq j}).$$

Remarque : les cas limites, probabilité égale à 0 ou à 1, sont en théorie acceptables mais poseront des problèmes dans les modèles (situation de séparabilité parfaite).

Dans le cas binaire simplifié qui nous intéresse $a = \{0,1\}$ et donc $p(y = 0) = 1 - p(y = 1)$.

- On peut appeler $a = 0$ évènement complémentaire.
- Ce processus suit une loi dite de Bernouilli qui peut être représenter par la formule suivante (fonction de masse):

$$\boldsymbol{p(y = a) = p^a \times (1 - p^{1-a})}$$ qui prend donc la valeur p si $a = 1$ et la valeur $1 - p$ si $a = 0$

Calcul

Si on s'intéresse maintenant à son calcul, celui de la probabilité est souvent comme le rapport entre le **nombre de cas favorables (oui) sur le nombre de cas possibles (oui + non)**. Lancer de dé/pièce sont des expériences aléatoires classiques dont la théorie, la probabilité a priori, est connue. Sortie de ces situations, ce sont les informations sur la population en général qui donne la théorie, mais elles doivent être disponibles, ou alors on utilisera les informations d'un échantillon pour en inférer la théorie (la réalité).

De manière pratique, l'estimation de la probabilité est calculée, comme dans la partie introductive, comme la moyenne de n réalisations indépendantes mesurées par une indicatrice $y = (0,1)$:

- $p(y = 1) = \frac{\sum y_i}{n}$.
- On en parlera par la suite, cet estimateur est dit du maximum de vraisemblance.

Odds

- La notion d'**Odds** (avec un **s**) est connue depuis le 16ème siècle. On la traduit en français généralement par *chance*.
- Sa mesure, si on la compare à celle de la probabilité, est le **nombre de cas favorables (oui) sur le nombre de cas défavorables (non)**:
  - $\boldsymbol{odds(y=1) = \frac{\sum y_i}{n - \sum y_i}}$
- On peut déduire directement sa valeur de celle des probabilités :
  - $\boldsymbol{odds(y=1) = \frac{p(y=1)}{1-p(y=1)} = \frac{p(y=1)}{p(y=0)}}$

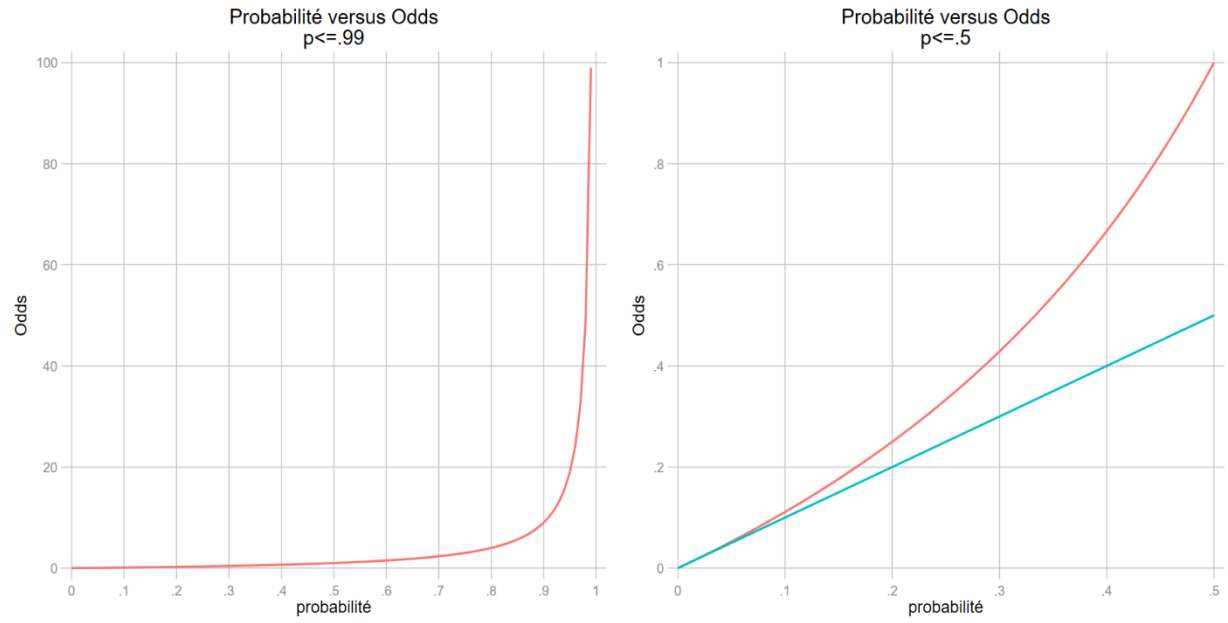
- Propriété: $0 \leq odds(y = a_i) < \infty$ [remarque: les cas limites correspondent à des probabilités 0 ou 1]

Exemple de valeurs comparées entre probabilités et Odds

- A une probabilité de 0.9 correspond un Odds de 9
- A une probabilité de 0.8 correspond un Odds de 4
- A une probabilité de 0.5 correspond un Odds de 1 (on a donc une chance de tomber sur pile ou face)
- A une probabilité de 0.2 correspond un Odds de 0.25
- A une probabilité de 0.1 correspond un Odds de 0.11
- A une probabilité de 0.05 correspond un Odds de 0.05

Question1: pour un lancer de dé à 6 face quelle est la probabilité et l'odds de tomber sur la face 2? Même question pour tomber sur une autre face que 2.

Question2: concernant la relation entre probabilité et Odds, que remarquez-vous?



Application: on fait une première analyse en terme de probabilité et d'odds sur le risque d'hypertension avec comme facteur associé celui du sexe.

sex	hypertension 0	1	Total
Male	2,611	2,304	4,915
Female	3,364	2,072	5,436
Total	5,975	4,376	10,351

- **Probabilité**
  - p(y=1) = 4376/10351 = 0.423
  - p(y=1 si homme) = 2304/4915 = 0.469
  - p(y=1 si femme) = 2072/5436 = 0.381
- **Odds**
  - odds(y=1) = 4376/5975 = 0.731
  - odds(y=1 si homme) = 2304/2611 = 0.884
  - odds(y=1 si femme) = 2072/3364 = 0.616

Rapport de probabilités et Odds ratio

Il n'y a pas de concept supplémentaire à ajouter, on va comparer pour deux niveaux d'un facteur les probabilités et Odds (Odds Ratio).

Dans un commentaire, il est important de ne confondre ces deux rapports en évitant de lire un **Odds Ratio** avec la formulation "*la probabilité est multipliée par α entre le groupe A et le groupe B*".

Question : dans quelle situation cela reste néanmoins possible (rapport de probabilité approché par un Odds Ratio)?

Si on reprend l'application précédente, le calcul du Ratio des Odds (OR) entre les hommes et les femmes est égal à:

- $\boldsymbol{OR_{hf} = \frac{odds(y=1 si homme)}{odds(y=1 si femme)} = \frac{0.884}{0.616} = 1.43}$
- Les hommes ont donc 1.43 chances de plus d'être atteint d'hypertension que les femmes (ou +43%)

Si on calcule le rapport de probabilité on obtient 1.23 (+23%). Cela confirme bien que lorsqu'on s'éloigne d'une situation "peu fréquente", les rapports d'Odds et de probabilités ne doivent pas être confondue.

On peut calculer un intervalle de confiance pour un Odds Ratio (sa variance étant égale à la somme des inverses des fréquences conditionnelles). Pour un test de différence avec la situation où les Odds de deux groupes ne diffèrent pas (OR=1), on peut faire un test d'indépendance du khi2. Nous verrons plus loin l'utilité d'une régression linéaire (égalité des probabilité) ou logistique (égalité des Odds) pour faire ce genre d'opération. L'approche par la régression va s'avérer, en plus, particulièrement utilisée lorsque des pondérations doivent être posées sur l'analyse.

Le modèle linéaire

Rappels

On va se mettre dans la situation la plus simple, pédagogique, avec deux variables directement mesurables (quantitatives). Nous éviterons tant que possible d'utiliser les notions de variables expliquées et explicatives, d'autant qu'on se limitera ici dans le cadre d'un modèle à une seule covariable (un seul x).

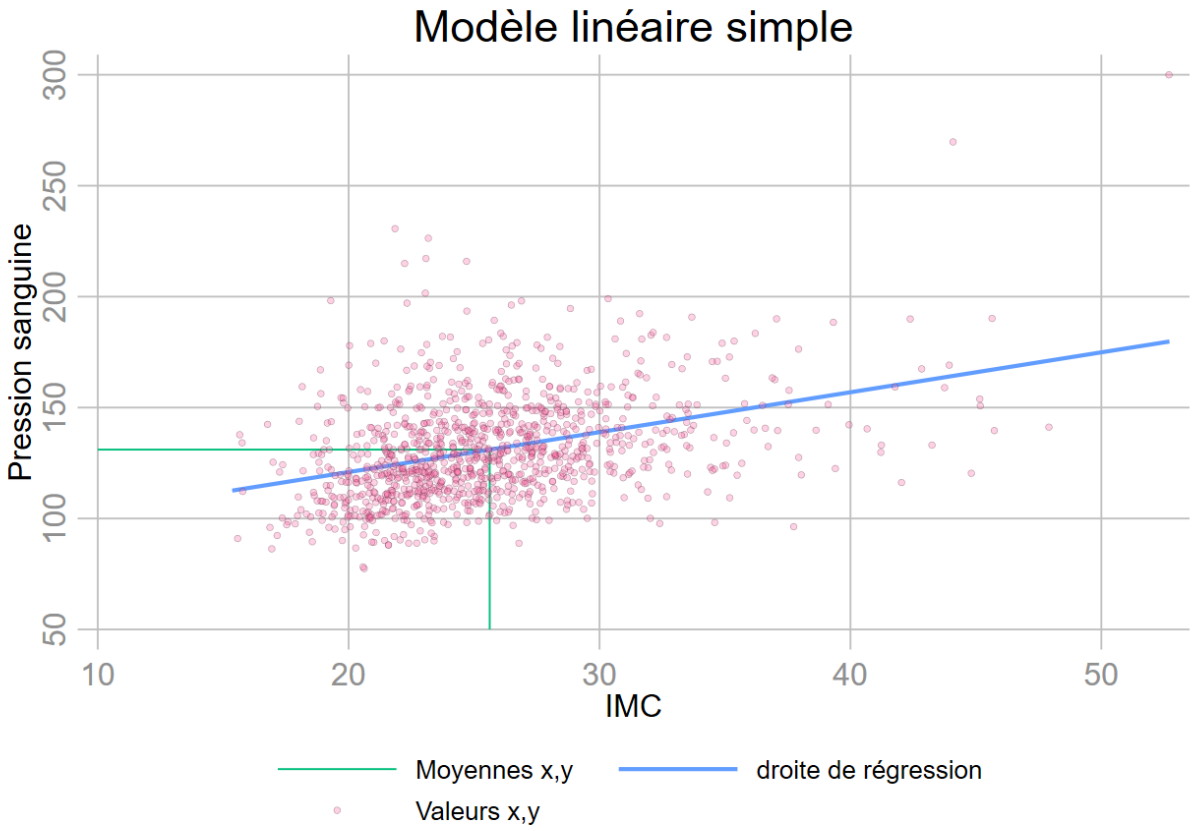
- L'équation du modèle donne l'espérance conditionnelle de y par rapport à x et s'écrit $\boldsymbol{E(y|x) = a + bx}$.
- Si y_i est une observation de y: $\boldsymbol{y_i = E(y_i|x_i) + u_i = a + bx_i + u_i}$.
  - u_i est l'erreur - il en faut - ou résidu issue du modèle: $\boldsymbol{u_i = y_i - E(y_i|x_i)}$.
  - La technique des moindres carrés ordinaires (MCO ou OLS en anglais), est la méthode de calcul privilégiée, et visera a trouver une solution pour b, la "pente" de la droite de régression, telle que la valeur de ces u_i soit au final minimale.
  - b dans cette forme strictement additive d'un modèle linéaire est régulièrement appelé **effet marginal**: si x varie de $+1$ alors y variera de b. Formellement il s'agit de la dérivée première de y par rapport à x.

Exemple On reprend l'analyse de l'hypertension, mais en analysant la variable d'origine qui mesure directement la pression artérielle. Le fait d'être en hypertension exprime donc le fait que la pression artérielle dépasse un certain seuil (il s'agit donc d'un regroupement). On regardera comment l'espérance du niveau de pression artérielle, conditionnellement au niveau d'IMC, « s'écarte » de l'espérance non conditionnelle, soit la pression artérielle moyenne observée dans l'échantillon.

Pour faciliter les représentations graphiques on a tiré aléatoirement un sous échantillon de 10% de l'échantillon d'origine, soit environ 1000 personnes.

```
Moyenne de y (pression artérielle): 130.9
Moyenne de x (imc): 25.5

------------------------------------------------------------------------------
    pression |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bmi |      1.6569     0.0437    37.89   0.000       1.5712      1.7426
       _cons |     88.5686     1.1373    77.88   0.000      86.3393     90.7978
------------------------------------------------------------------------------
```

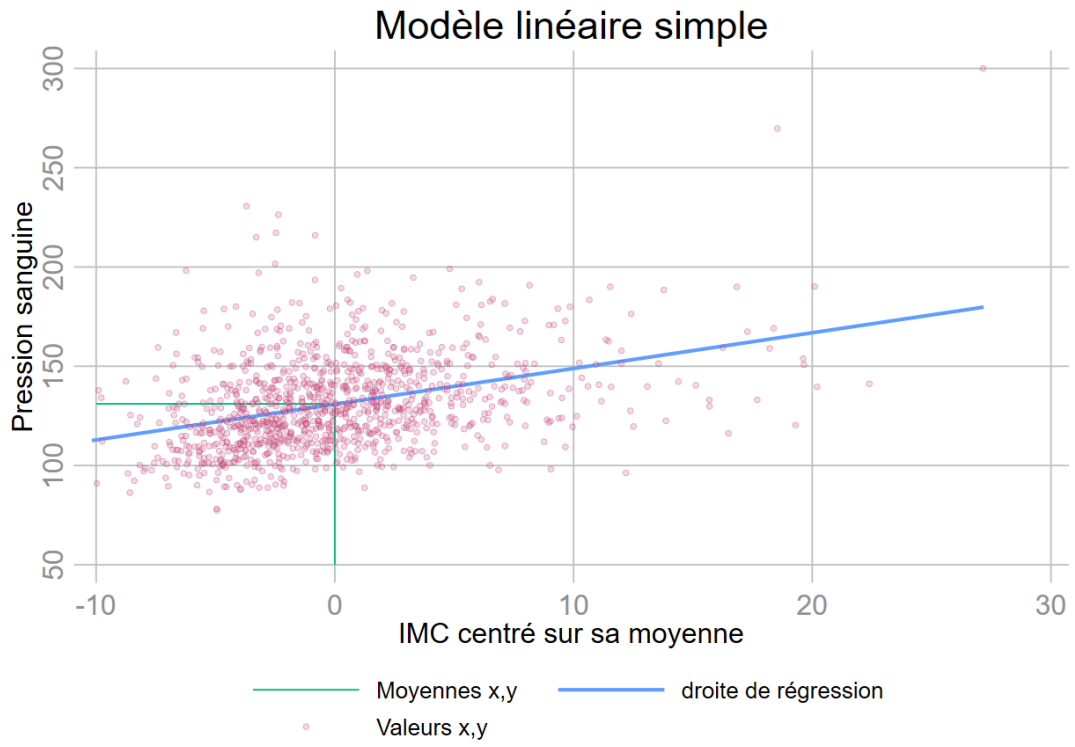


On observe donc une corrélation positive entre la pression artérielle et l'IMC. Lorsque l'IMC augmente d'une unité, la pression artérielle augmente de 1.66 unités. On remarque qu'au niveau de l'IMC moyen (25.5), la valeur prédite de la pression artérielle est égale à sa moyenne : 130.9 = 88.57 + 1.66.

Attention à l'interprétation de la constante qui peut avoir peu ou pas de réalité, ce qui est le cas ici. En effet, dans cette spécification de x, simplement sa valeur observée, elle donne le niveau de pression pour un IMC=0, soit une personne de poids nul. On peut changer la spécification de l'indice de masse corporelle dans le modèle en prenant pour chaque observation, sa différence avec l'IMC moyen. On centre donc la variable. En termes de lecture, le gain est immédiat pour la lecture de la constante, qui est maintenant la pression artérielle moyenne. L'effet marginal b reste identique car on a simplement translaté les valeurs de l'IMC par rapport à sa moyenne.

```
Moyenne de y (pression artérielle): 130.9
Moyenne de la valeur centrée" de x (imc - imc moyen): 0


    bpsystol |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     diffimc |      1.6569     0.0437    37.89   0.000       1.5712      1.7426
       _cons |   130.8817     0.2149   608.96   0.000     130.4604    131.3029
------------------------------------------------------------------------------
```



Quelques propriétés

- L'espérance, des termes d'erreur est nulle: $\boldsymbol{E(u) = 0}$.
- Les erreurs sont de variance constante, elles ne dépendent pas de x. C'est la fameuse hypothèse d'**homoscédasticité**.
  - Pour scédasticité Wiki dit => vient du grec ancien skedastikos: « dissipatif », « dispersif ». Visiblement pas d'autres usages qu'en Statistique, on ne blamera donc pas les profs de français du collègue ou du lycée. Scédasticité = variance des erreurs.
  - L'hypothèse est testable et si problème, on peut le corriger.
  - Cas typique : l'introduction de pondérations engendre automatiquement la non homoscédasticité des erreurs. On règle le problème avec une estimation robuste de la variance des estimateurs, après normalisation des pondérations (somme des poids = taille de l'échantillon).
- Une autre propriété à vérifier est la non colinéarité des covariables introduites dans le modèle : une variable ne doit pas être une combinaison linéaire d'autres variables.
  - On parle ici de colinéarité ou de multicolinéarité très forte. Dans ce cas les estimateurs ne sont plus de variance minimale, le modèle pourrait devenir impossible à estimer. Cette propriété est facilement testable.
  - Remarque : la notion de colinéarité entre variable discrètes ne va pas de soit.

Propriété additionnelle et fondamentale pour un modèle à visée explicative ou causale

- **Exogénéité**: Les erreurs ne doivent pas être corrélés aux covariables : $\boldsymbol{E(u|x) = 0}$. C'est de cette hypothèse qu'on tire la notion de **variable exogène**, et c'est également lorsque cette propriété est respectée qu'il est possible d'utiliser une expression du type **toute chose égale par ailleurs**.

Le modèle de probabilité linéaire

Le modèle linéaire classique est une méthode centrée sur des déviations conditionnelles par rapport à un moyenne (l'espérance). On l'a vu, la moyenne est un estimateur d'une probabilité, normée sur l'intervalle allant de 0 à 1 (plutôt $[0^+, 1^-]$). Il n'est donc pas a priori farfelu d'analyser une réponse discrète/binaire à partir du modèle linéaire standard.

Nous allons, par l'exemple, estimé un modèle linéaire pour évaluer le risque d'être en hypertension. On va également lister les critiques faites au modèle à probabilité linéaire, puis rapidement comparer, sous forme visuelle, les résultats avec celle d'un modèle purement probabiliste comme le modèle logistique.

Visuellement :

- Comme y ne prend que deux valeurs il n'y a plus vraiment de nuage de points, les 2 valeurs sont seulement dispersées sur les valeurs de l'IMC.
- On remarques que certaines valeurs prédites de y sortent des bornes, à savoir qu'on obtient une probabilité prédite >1 pour des valeurs élevées de l'IMC.
- La valeur de la constante si on prend les valeurs brutes de l'IMC n'a de nouveau pas de sens.

Avec les valeurs de l'IMC

```
------------------------------------------------------------------------------
      highbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bmi |      0.0329     0.0030    11.14   0.000       0.0271      0.0387
       _cons |     -0.4147     0.0770    -5.39   0.000      -0.5658     -0.2636
------------------------------------------------------------------------------
```

Avec les valeurs centrées de l'IMC (voir niveau de la constante avec un IMC=0)

```
      highbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     diffimc |      0.0329     0.0030    11.14   0.000       0.0271      0.0387
       _cons |      0.4255     0.0143    29.74   0.000       0.3975      0.4536
------------------------------------------------------------------------------
```

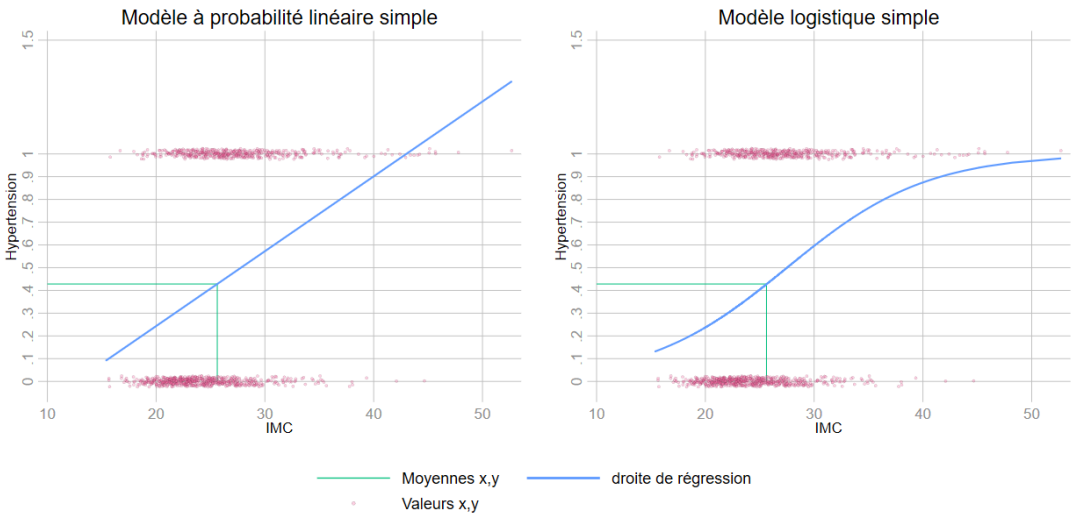
Remarque : avec des valeurs centrées de l'IMC, la probabilité moyenne (la constante) diffère légèrement de celle du début de la section (0.423) car on a sélectionné aléatoirement 10% de l'échantillon d'origine.

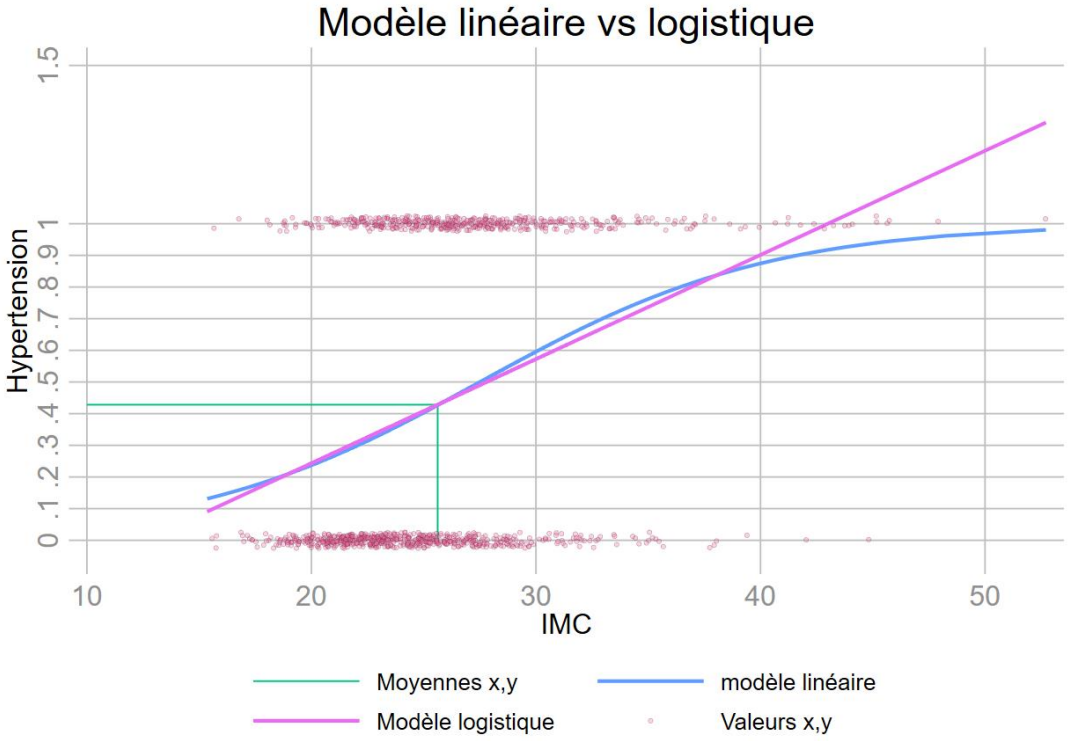
En terme d'interprétation, un augmentation d'un point de l'IMC augmente la probabilité d'être en hypertension de 0.03 points (soit +3 points de pourcentage)

Limites et critiques faites au modèle à probabilités linéaires :

- Les espérances conditionnelles (valeurs prédites) peuvent sortir des bornes de définition de la probabilité. C'est vrai, on le voit sans l'exemple.
  - L'importance de ce problème dépend généralement de la rareté de l'évènement étudié (ou sa très grande fréquence). Dans l'exemple, avec en moyenne 40% de personnes en situation d'hypertension seulement une dizaine d'observations sont concernées, et on pourrait tronquer leur valeur prédite à 1 (ce qui est généralement fait).
  - On peut souligner que pour une variable quantitative bornée (par exemple le salaire), le modèle linéaire ne contraint pas non plus les espérances conditionnelles à respecter les bornes de définitions de la variable, et il n'est pas rare que la prédiction sorte également des bornes.
- Le modèle est par définition hétéroscédastique. C'est vrai. Mais on a la méthode pour résoudre ce problème (estimation robuste).
- Les erreurs ne sont pas distribuées selon une loi normale, car le terme d'erreur ne peut prendre que 2 valeurs pour chaque observation en raison du caractère binaire de la variable de réponse. C'est vrai. Mais pour les modèles probabilistes type logistique ce n'est pas mieux (attention point compliqué en raison de la présence d'une variable latente sous-jacente).
- Les modèles probabilistes repose sur des formes fonctionnelles (loi de probabilité), ce qui n'est pas le cas avec le modèle linéaire standard. C'est vrai, et c'est surement le point le plus problématique. Mais comme on peut le voir avec les graphiques qui suivent, la distribution des valeurs prédites par les modèles sont très proches si on compare le modèle linéaire standard et un modèle probabiliste qui utilise une fonction sigmoïde, quasiment linéaire sur un large intervalle, au moins entre 30% et 70%.
- Sa non généralisation : le modèle à probabilité linéaire n'est pas généralisable aux variables qualitatives qui ont plus de 2 modalités. Le modèle logistique, par exemple, s'applique avec des hypothèses identiques à une variable qualitative dont les modalités ne sont pas ordonnées (modèle logistique multinomial) ou ordonnées (modèle logistique ordinal). Le modèle logistique binaire ou dichotomique, n'est en fait qu'un cas particulier de ces deux types de modèles.

Au final, malgré des certains problèmes, le modèle à probabilité linéaire est une méthode qui peut être envisagée lorsque l'évènement étudié n'est pas rare, et ce situe par exemple dans une fourchette allant de 40% à 60%.


Modèle à probabilité linéaire simple
Modèle logistique simple
Hypertension
IMC
Moyennes x,y
droite de régression
Valeurs x,y


Modèle linéaire vs logistique
Hypertension
IMC
Moyennes x,y
modèle linéaire
Modèle logistique
Valeurs x,y

LE MODÈLE LOGISTIQUE

Notation de la combinaison linéaire

Comme on va se situer dans le cadre d'un modèle de regression de type multiple, souvent appelé *multivarié*, avec plus d'une covariable, on va simplifier dès à présent les notations. En repartant d'une expression de type modèle linéaire:

$$y = b_0 \times 1 + b_1 \times x_1 + b_2 \times x_2 + \ldots + b_k \times x_k = b_0 + \sum_{j=1}^{k} x_j = xb_j$$

Remarque: pour la constante x_0 est toujours égale à 1.

Approche descriptive et approche causale/explicative

Approche descriptive

Objectifs

- Montrer que la fonction logistique est le reflet la définition d'une simple proportion (d'où approche descriptive).
- Décrire rapidement ses propriétés et les quantités associées qui seront estimés par le modèle : le *logit* et *contraste logistique*.

Pour montrer que la fonction logistique, on va dire sa formule, est au final l'expression la plus simple d'une proportion, on va repartir du modèle linéaire. On cherche à identifier les facteurs facilitant l'appartenance à deux groupes A et B.

Dans la logique du modèle à probabilité linéaire, on peut estimer les probabilités conditionnelles d'appartenance à chaque groupe ainsi :

- probabilité pour $g1$: $p(A|x) = xb_A$
- probabilité pour $g2$: $p(A|x) = xb_B$

Problème soulevé pour le modèle à probabilité linéaire : rien n'assure que la probabilité conditionnelle estimée soit comprise entre 0 et 1.

Solution (intermédiaire) : on va s'assurer que la probabilité soit non négative en posant l'exponentielle :

- probabilité pour A: $p(A|x) = e^{xb_A}$
- probabilité pour B: $p(B|x) = e^{xb_B}$

On résout certes le problème de la négativité de la probabilité obtenue, mais avec cette formulation elle risque d'être très souvent supérieure à 1.

Solution (presque finale): on norme simplement les quantités obtenues, positives, par leur somme. On obtient alors strictement l'expression d'une proportion :

- **probabilité pour A: $\boldsymbol{p(A|x) = \frac{e^{xb_A}}{e^{xb_B}+e^{xb_A}}}$**
- **probabilité pour B: $\boldsymbol{p(B|x) = \frac{e^{xb_B}}{e^{xb_A}+e^{xb_B}}}$**

On voit bien que $p(A|x) + p(B|x) = 1$, nos probabilités sont parfaitement calculées.

Presque finale…car Trop de paramètres à estimer

Approche simple/intuitive : sachant que la somme des probabilités est égale à 1, il n'est pas nécessaire d'estimer simultanément un jeu de paramètres pour les appartenances aux deux groupes. Si on s'intéresse prioritairement au groupe A par exemple, et si on pose que tous les paramètre pour le second sont égaux à 0:

$$\boldsymbol{p(A|x) = \frac{e^{xb_A}}{e^{0} + e^{xb_A}} = \frac{e^{xb_A}}{1 + e^{xb_A}}}$$

On retrouve donc la formulation de la fonction de répartition de la loi logistique:

$$p(A|x) = \frac{e^{xb_A}}{1 + e^{xb_A}} = \frac{1}{1 + e^{-xb_A}} = \frac{1}{1 + e^{-x_A}}$$

Et tout simplement pour le second groupe:

$$p(B|x) = 1 - \frac{e^{xb_A}}{1 + e^{xb_A}}$$

Suridentification:
On peut montrer, c'est l'approche standard, que si n'on impose pas de contrainte aux paramètres d'un des deux groupes, il n'y a pas de solution unique.
Soit λ un nombre non nul. Si on multiplie la quantité normée plus haut par $\frac{e^{\lambda}}{e^{\lambda}}$ (=1):

- $\frac{e^{\lambda}}{e^{\lambda}} \times \frac{e^{xb_A}}{e^{xb_B}+e^{xb_A}} = \frac{e^{x(b_A+\lambda)}}{e^{x(b_B+\lambda)}+e^{x(b_1+\lambda)}}$
- Les paramètres b_A et $b_A + \lambda$ sont tous les deux des solutions, et il y a donc une infinité.
- On doit donc poser une contrainte sur un des deux ensembles de paramètres, au choix celui pour le groupe A ou le groupe B, qui sera alors un groupe de référence.
- Si on pose $b'_A = b_A - b_B = b$ et donc $b'_B = b_B - b_B = 0$, la solution est unique (en multiplient par le rapport de λ sur lui même, on modifie la probabilité obtenue.

On retrouve de nouveau la formulation d'une probabilité d'affection avec la loi logistique :

$$p(A|x) = \frac{1}{1 + e^{-xb}}$$

Petit spoil : c'est exactement ce que qu'on fait avec les covariables discrètes, binaire ou non, en choisissant par exemple une modalité de référence. Ce type de contrainte est obligatoire pour estimer un jeu unique de paramètres associées à une variable non quantitative.

Propriétés

Odds et logit :

L'Odds associé à la probabilité: $\frac{p(A|x)}{1-p(A|x)} = e^{xb}$. Ce résultat est obtenu par simple calcul.

On va définir le **logit** comme le logarithme de l'Odds :

$$\boldsymbol{logit(A|x) = log\left(\frac{p(A|x)}{1-p(A|x)}\right) = xb}$$

L'intérêt du logit ne réside pas dans son interprétation, mais sur sa forme linéarisée. Le logit est donc égal à la valeur de la combinaison linéaire obtenu par le modèle. C'est sur cette forme logarithmique que les paramètres du modèle vont être estimés.

Odds ratio et contraste logistique :

Une fois définie la notion de logit, est très facile de déduire l'Odds Ratio qui sera obtenu.

On va comparer avec un Odds ratio, les chances d'appartenance à A pour deux valeurs d'une variable x binaire $\{0,1\}$. Dans le cadre d'une régression multiple, on introduit une covariable supplémentaire z dont le paramètre estimé sera égal à c.

On peut écrire: $logit(A|x,z) = a + bx + cz$.

Si on fait varier seulement x:

- logit pour $x = 1$: $logit(A|x = 1, z) = a + b + cz$
- logit pour $x = 0$: $logit(A|x = 0, z) = a + cz$
- Si on calcule la différence entre les deux logit :

  $$\boldsymbol{logit(A|x = 1, z) - logit(B|x = 0, z) = b.}$$

  Le terme constant et celui de l'autre variable sont éliminés par différence. Attention, ce n'est pas cette élimination qui justifie l'expression *"toute chose égale par ailleurs"*.

- En appliquant l'exponentielle, on retrouve le rapport des Odds :

  $$\boldsymbol{OR(A)_{(x=1\ versus\ x=0)} = e^{b}}$$

On va définir **le contraste logistique** comme la différence entre deux logits. C'est ce contraste qui sera estimé par le modèle (sauf pour le terme constant qui sera un logit). Les Odds ratios, directement interprétables, sont tout simplement l'exponentiel du contraste logistique.

Approche causale/explicative

Warnings:

- Bien plus complexe mais point essentiel. Au niveau mathématique, on restera sur l'essentiel. On va juste introduire les notions de modèle structurel, de variable latente, de biais d'endogénéité.
- Au final on estime toujours le même modèle, avec juste un petit hiatus sur la constante. C'est sur la validation de son pouvoir causal/explicatif qu'on va ajouter des hypothèses supplémentaires.
- Par extension, cette identification justifie régulièrement la modélisation des variables polytomiques de types ordinales, en particulier lorsqu'on analyse des variables basées sur des perceptions ou des opinions, par nature fortement dépendantes à des traits latents.
- Un cas particulier de cette identification est la théorie des choix discrets, qui ne sera pas abordée directement.

Identification du modèle

On repart d'un modèle linéaire (hé oui) à ceci près que la variable d'intérêt n'est pas directement observable. Elle est liée au phénomène qu'on analyse, et peut traduire des attitudes, préférences, « capacités ». La variable qu'on analyse est une traduction, ou plutôt une manifestation de cette variable latente.

On suppose que « derrière » la variable discrète/binaire qu'on analyse, il existe une variable latente Y^* strictement continue. C.Afsa dans son document prend l'exemple du niveau d'un l'élève comme la traduction latente de la réussite ou non à un examen.
La valeur du trait latent peut être relié à une série de covariables, celles de notre modèle logistique par exemple. Pour une observation particulière :

$$Y_i{}^* = b_0{}' + xb_i{}' + u_i$$

- Il s'agit du modèle dit « structurel ».
- On fait l'hypothèse de Y^* suit une distribution symétrique, par exemple normale (gaussienne) ou logistique.
- C'est également (et surtout) le cas du terme résiduel u_i qui contient toutes les informations que le modèle ne peut pas saisir. C'est le point le plus important de cette approche : aucune covariable doit être corrélée avec une ou plusieurs autres qui non pas été introduites. Soit par ce qu'on en a oublié, soit parce qu'elles ne sont pas observées (absente des données), soit parce qu'elles ne peuvent pas être observées. **C'est le fameux biais d'endogénéité**. On doit donc s'assurer du caractère parfaitement aléatoire et non sélectif du terme d'erreur. Si on ne peut pas, il faut proscrire dans le commentaire les expressions de type « toutes choses égale par ailleurs ».

On a pris soin de séparer ici le terme constant, dont l'expression sera un peu différente que précédemment.

Comment s'articule $y = \{0,1\}$ et Y^*?

Soit s_0 une valeur inconnue considérée comme un seuil, lui même inconnu. Le dépassement de ce seuil par la variable $Y*$ activera le signal mesuré par la variable binaire Y.

- $y = 1$ si $Y^* = {b_0}' + x_i b' + u_i > s_0$
- $y = 0$ si $Y^* = {b_0}' + x_i b' + u_i < s_0$

Comme on suppose que $Y*$ n'est pas directement observable, on doit alors utiliser la variable binaire y

$$P(y_i = 1|x) = P({Y_i}^* > s_0 \mid x) = P({b_0}' + x_i b' + u_i > s_o)$$

Après manipulation de l'équation en posant $b_0 = {b_0}' - s_0$:

$$P(y_i = 1|x) = P(u_i < x_i b)$$

On postule que les erreurs/résidus suivent une loi logistique, représentée par sa fonction de répartition $P(u_i < x_i b) = F(x_i b)$.

Au final :

$$\boldsymbol{P(y_i = 1|x) = P(u_i < x_i b) = F(x_i b)}$$

Remarques:

- Encore une fois, attention aux commentaires de type explicatif/causal et à l'utilisation de l'expression « toute chose égale par ailleurs ».
- C'est sur cette approche qu'est fondé le cadre général du modèle linéaire généralisé, brièvement décrit plus bas.
- L'extension à un type de modèle à variable ordinales modifie l'interprétation de la constante voire, selon les logiciels, à des changements de signes des paramètres estimés

Estimation du modèle

Juste quelques informations sur la méthode d'estimation. Important car des indicateurs de performances du modèle et des tests statistiques en sont déduit : test du maximum de vraisemblance, « pseudo R2 », critères d'information type BIC/SC et AIC, tests de Wald sur les paramètres estimés.

La vraisemblance

- Le modèle logistique est estimé par la méthode du maximum de vraisemblance.
- La vraisemblance est formellement un produit de probabilités, la probabilité d'observer l'échantillon.
- La vraisemblance est construite à partir d'un processus qui génère les données observées : cela peut être une densité (loi normale) ou une fonction de masse (loi discrète de type Bernouilli).
- Ces processus ont un ou plusieurs paramètres a priori inconnus, dont on cherche des estimateurs à partir des observations :
  - Loi normale: moyenne et écart type

  - loi binomiale/Bernouilli : probabilité
- Ces paramètres sont estimés par la maximisation de la vraisemblance :
  - Moyenne d'une loi normale: $\sum \frac{y_i}{n}$
  - Probabilité d'une loi de bernouilli: $\sum \frac{y_i}{n}$

De la vraisemblance d'un processus de Bernouilli à l'estimation d'un modèle logistique

- On observe $y = \{0,1\}$ dont on cherche à prédire l'occurence via des probabilités.
- On a n obserations.
- La vraisemblance s'écrit $\boldsymbol{L(y_1, y2, \ldots, p) = L(.) = \prod p^{y_i} \times (1-p)^{1-y_i}}$
- La solution de l'estimation non conditionnelle par le maximum de vraisemblance conduit à ce que $p = \sum \frac{y_i}{n}$
  - Pour des raisons pratiques on a transformé la vraisemblance par la log-vraisemblance (le produit devient une somme)
  - On dérive la log-vraisemblance par rapport à p et cherché une solution tel que $logL(.) = 0$.

Remarque/question: après quelques manipulations la vraisemblance s'écrit:

$$\boldsymbol{L(.) = \prod (1-p) \times e^{y_i \times log\left(\frac{p}{1-p}\right)}}$$

Que remarque t-on?

Fonction de lien logit et modèle linéaire généralisé

- Pour une estimation avec au moins une covariable : $log\left(\frac{p}{1-p}\right) = xb$, le logit est mesuré par une combinaison linéaire de covariables ou prédicteurs. On parle de fonction de lien.
- La particularité du logit, est que la fonction de lien est dite *naturelle* ou *canonique*. C'est ce qui permet d'interpréter directement les paramètres estimés.
- Une fonction de lien appliquée à la vraisemblance permet d'obtenir une estimation de type **modèle linéaire généralisé** ou **glm** (dans la réalité c'est un peu/beaucoup plus complexe).

Estimation

- Pour estimer le modèle logistique, on remplace le paramètre inconnu p par la fonction de répartition de la loi logistique. Le vecteur des b de la combinaison linéaires deviennent les paramètres à estimer du modèle (on observe y et x).
- La vraisemblance est transformée par son logarithme, et après réécriture, on calcule les dérivées partielles de la log vraisemblance pour chaque paramètre b. On obtient une fonction dite de *score*.
- Comme on cherche une solution de type maximum, on cherche b tel que l'équation de score soit égale à 0. Malheureusement il n'est pas possible de trouver une solution numérique directe (les équations sont trop complexes).

- On utilise un algorithme d'optimisation pour obtenir les solutions, généralement l'algorithme de Newton-Raphson, proche de la logique des moindres carrés ordinaires (avec un facteur de repondération). Cet algorithme approche la solution par itérations successives en utilisant le vecteur des b (le gradient) et la matrice des dérivées secondes des fonctions de scores (matrices des variances covariances) estimés à chaque étape.
- Une fois la solution « stable » - les paramètres ne varient plus d'une itération à une autre- on obtient les résultats du modèle.

Remarque: calcul de la vraisemblance d'un modèle

Une fois le modèle estimé, on peut calculer sa vraisemblance à partir des probabilités obtenues. On va comprendre rapidement que la méthode consiste à minimiser l'écart entre ce qui est observé et ce qui est prédit. A partir d'un exemple totalement fictif avec 5 observations (donc non estimable dans la réalité):

- y sont les valeurs observées de la variable qu'on analyse.
- p_0 et p_1 sont les probabilités estimées par deux modèles (on va dire logistique) par la méthode du maximum de vraisemblance.
  - p_0 est issu d'un modèle sans covariable (non conditionnel), p_0 est donc égal à la simple moyenne de y.
  - p_1 sont issus d'un modèle avec des covariables.
- cl_0 et cl_1 sont les contributions à la vraisemblance de chaque observation pour les deux modèles. Si $y = 1$ elle est égale à p, si $y = 0$ elle est égale à $1 - p$.
- L_0 et L_1 sont les vraisemblances pour les deux modèles.

i	y	p0	p1	cl0	cl1
1	1	.6	.7	.6	.7
2	0	.6	.4	.4	.6
3	1	.6	.8	.6	.8
4	1	.6	.5	.6	.5
5	0	.6	.1	.4	.9

Pour le modèle non conditionnel (seulement la constante): la probabilité prédite est identique et égale à la moyenne de y. En appliquant la valeur de y et $p_0 = 0.6$ à la fonction de masse de Bernouilli on obtient donc seulement deux valeurs pour la contribution à la vraisemblance:

- $cl0 = 0.6 = p_0$ si $y = 1$

- $cl0 = 0.4 = 1 - p_0$ si $y = 0$.

La vraisemblance est égale au produit de ces contributions: $L0 = 0.034$.

Pour le modèle avec covariables : Selon les valeurs observées par les covariables, la contribution à la vraisemblance varie. Elle est très forte pour la dernière observation où $y =$

0 et $p = 0.1$, avec une contribution de 0.9. La vraisemblance est égale au produit de ces contributions: $L1 = 0.151$.

Avec cette illustration fictive, on voit que le l'introduction de covariables apporte une information supplémentaire par rapport à la mesure calculée sur la moyenne brute (modèle dit « vide »). Par analogie avec le modèle linéaire on pourrait regarder la somme des écarts au carré entre les valeurs observées et les valeurs prédites. Elle est bien évidemment plus faible pour le modèle que pour l'espérance non conditionnelle de y.

Il est d'usage de calculer et d'utiliser le logarithme de la vraisemblance, c'est cette quantité qui est reportée par les logiciels (ou la déviance qui lui est associée) et celle utilisée dans le test de leur rapport. Comme il s'agit d'un produit de probabilité elle devient très vite très proche de 0 lorsque la taille de l'échantillon augmente. Attention comme il s'agit d'une probabilité, son logarithme va être négatif. Plus la log-vraisemblance est proche de 0 et donc plus elle augmente. Dans l'illustration, la log vraisemblance de l'espérance non conditionnelle est égale à -3.38 et celle pour le modèle avec covariables est égale à -1.89.

Remarques:

Autres fonctions de lien

Pour estimer le modèle logistique, on a utilisé la fonction de répartition de la loi logistique. Deux alternatives existent, mais les paramètres obtenus ne sont pas directement interprétables.

- Le **modèle probit**: on prend la fonction de répartition de la loi normale. Très utilisé en économie, il permet entre autres d'estimer des modèles plus complexes comme le probit bivarié (solution pour un biais de causalité vu précédemment), grâce à la propriété multidimentionnelle de la loi normale.
- Le modèle **complémentaire log-log**: rarement utilisé et uniquement en analyse des durées en temps discret en raison de son lien avec la fonction de survie estimée à l'aide d'un modèle de Cox.

Modélisation directe de proportions

C'est à savoir, le modèle logistique permet de modéliser non pas seulement des variables binaire $\{0,1\}$, mais toutes variables dont les valeurs sont comprises entre 0 et 1, soit des proportions. On pourrait penser à une analyse de part de certaines dépenses dans les ménages (logement, transport…). Cette possibilité est donnée par les propriétés du modèle linéaire généralisé dont un facteur de dispersion a été contraint (=1) pour une réponse binaire. Tous les logiciels permettent ce genre d'estimations (`glm(…)` avec R, `proc genmod` avec Sas, `glm` avec Stata).

Pour s'en convaincre, on peut également penser à l'introduction de pondérations dans le modèle : si les valeurs ne sont pas modifiées pour y=0, pour y=1 la valeur qui sera utilisée sera celle de la pondération [toujours normalisé les pondérations svp].

INTERPRÉTATION

La section sera principalement développée par l'exemple (risque d'hypertension)

Dans un premier temps, on va présenter les données et les variables dans le détail, estimer un premier modèle, commenter les résultats selon le type de variable, faire un bref rappel sur le test de Wald sur les paramètres.

Introduction et rappels sur l'inférence

Données et variables

La base est issue du programme d'études NHANES (https://www.cdc.gov/nchs/nhanes/about_nhanes.htm). Il s'agit ici de la seconde vague réalisée dans la deuxième moitié des années 70 (donc c'est un peu ancien)

Nombre d'observations : 10351

Variables avec quelques commentaires :

- **highbp**: variable d'intérêt, sous forme d'indicatrice ${0,1}. 1 si la personne ne présente de l'hypertension, 0 sinon
- **female**: sous forme d'indicatrice, être une femme (1) ou non (0). Une version categorielle est également présente. Il s'agit de la variable **sexe** avec les modalités "female" et "male".
- **black**: sous forme d'indicatrice, permet de repérer les personnes afro-américaines dans l'échantillon. 1 si "black", 0 sinon.
- **age**: variable quantitative (valeurs mesurées) qui indique l'âge de la personne. Sa moyenne est de 47 ans avec un minimum de 20 ans et un maximum de 74 ans
- **region**: variable catégorielle à 4 modalités qui indique la région d'habitation de la personne. "NE" pour une personne qui habite dans le Nord-Est, "MW" pour le Middle West, "S" pour le Sud et "W" pour l'Ouest.
- **hsize**: taille du ménage. Elle est en format numérique, avec 5 valeurs de 1 à 5. Les tailles supérieures ou égales à 5 ont été regroupées. Elle sera traitée comme une variable catégorielle.
- **rural**: sous forme d'indicatrice, vivre dans une zône rurale.
- **bmi**: variable quantitative, plutôt continue, qui mesure l'indice de masse corporelle. Sa valeur moyenne est de 25.5, avec un minimum de 12.4 et un maximum de 61.1.

On va commencer par un petit rappel, sous forme d'exercice. On va juste regarder l'association entre le risque d'hypertension et le sexe.

```
           |   1 if bpsystol >=
           | 140|bpdiast >= 90, 0
 1=female, |       otherwise
    0=male |         0          1 |     Total
-----------+----------------------+----------
         0 |     2,611      2,304 |     4,915
         1 |     3,364      2,072 |     5,436
-----------+----------------------+----------
     Total |     5,975      4,376 |    10,351
```

Calculer les Odds associés à l'hypertension pour les femmes et les hommes ainsi que leur rapport (femme versus homme) à partir de ce tableau croisé.

On estime un modèle logistique simple avec les mêmes variables. Le premier tableau affiche les résultats sous forme de logit - celui des hommes - et de contraste logistique soit la différence entre le logit des femmes et celui des hommes. Le second est présenté sous forme d'Odds et d'Odds ratio. Calculer les "chances" (l'Odds) d'hypertension pour les femmes.

Logit et contraste logistique

```
Logistic regression                               Number of obs     =     10,351
                                                  LR chi2(1)        =      81.22
                                                  Prob > chi2       =     0.0000
Log likelihood = -7010.1552                       Pseudo R2         =     0.0058

------------------------------------------------------------------------------
      highbp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |    -0.3595     0.0400    -9.00   0.000      -0.4379     -0.2812
       _cons |    -0.1251     0.0286    -4.38   0.000      -0.1811     -0.0691
------------------------------------------------------------------------------
```

Odds et odds ratio

```
Logistic regression                               Number of obs     =     10,351
                                                  LR chi2(1)        =      81.22
                                                  Prob > chi2       =     0.0000
Log likelihood = -7010.1552                       Pseudo R2         =     0.0058

------------------------------------------------------------------------------
      highbp | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |     0.6980     0.0279    -9.00   0.000       0.6454      0.7549
       _cons |     0.8824     0.0252    -4.38   0.000       0.8343      0.9333
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

A partir du tableau croisé, on réalise un simple test d'indépendance du Khi2, dont le résultat est le suivant :

```
          observed frequency
          expected frequency

-----------------------------
          |  1 if bpsystol >=
          | 140|bpdiast >= 90,
1=female, |    0 otherwise
0=male    |        0         1
----------+------------------
        0 |     2611      2304
          | 2837.129  2077.871
          |
        1 |     3364      2072
          | 3137.871  2298.129
-----------------------------

          Pearson chi2(1) =  81.1787   Pr = 0.000
```

En comparant l'output de la régression, de visu seulement, remarquez-vous des éléments proches ou concordants ??

Pause sur inférence des paramètres estimés et le test de Wald

Propriété de l'estimateur :

- Les paramètres estimés par la méthode du maximum de vraisemblance, sont des variables aléatoires. Par exemple le contraste logistique estimé est sujet à un certain degré d'incertitude, ou plutôt reflète les données qui ont été collectées (en statistique on parle d'estimateur ponctuel).
- Propriété : le contraste logistique suit une loi normale (si n suffisamment grand), dont la moyenne est la valeur obtenue par l'estimation. Une variation standardisée a été également estimée, appelée erreur type (σ). Si on nomme $\hat{b}$ l'estimation obtenue de ce contraste logistique, la valeur de ce contraste suit alors une loi normale: $N\left(\hat{b}, \sigma(\hat{b})\right)$.
- Exemple: pour le risque d'hypertension, on peut écrire que le contraste logistique b suit une loi normale N(-0.3595,0.0400).

De cette propriété, il est possible de construire deux choses :

- Des intervalles dits « de confiance ».
- Des tests visant à comparer l'estimation obtenue au modèle à une autre valeur, typiquement tester si cette valeur est différente de 0 (ie Odds Ratio différent de 1). Cette valeur peut être un autre estimateur du modèle, ou tout autre combinaison de paramètres estimés par le modèle. Le résultat de cette comparaison est obtenu par le test de Wald.

Test de Wald

Si on restreint le test à une comparaison entre le contraste estimé et 0 (ou OR par rapport à 1), la normalité de l'estimateur permet de standardiser l'écart avec la valeur de comparaison.

Version standard (Khi2)

- $W = \left(\frac{\hat{b}-0}{\sigma(\hat{b},0)}\right)^2 = \frac{\hat{b}^2}{variance(\hat{b})}$.
- W suit une loi du khi2 à 1 degré de liberté.
- La probabilité obtenue donne un seuil de conformité à $\hat{b} = 0$.
- Il s'agit d'un test bilatéral, on ne teste pas le signe de l'écart entre $\hat{b}$ et 0 mais son existence.
- Cette version (d'origine) du test est celle affichée par le logiciel SAS.

Version sur la loi normale (la plus utilisée : R, Stata, Python)

- $W = \left(\frac{\hat{b}-0}{\sigma(\hat{b},0)}\right) = \frac{\hat{b}}{\sigma(\hat{b})}$.
- W suit une loi normale $N(0,1)$.
- Il s'agit également d'un test bilatéral et la probabilité donne également un seuil de conformité pour $\hat{b} = 0$

Remarques : il n'y a pas de différence entre ces deux versions qui reposent sur les mêmes hypothèses. La version standard (Khi2) est jugée plus conservatrice sur un petit échantillon, elle favorise l'hypothèse nulle.

Reprenons l'output de la régression simple :

```
Logistic regression                                 Number of obs     =     10,351
                                                    LR chi2(1)        =      81.22
                                                    Prob > chi2       =     0.0000
Log likelihood = -7010.1552                         Pseudo R2         =     0.0058

------------------------------------------------------------------------------
      highbp |       Coef.   Std. Err.       z     P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |    -0.3595     0.0400    -9.00   0.000      -0.4379     -0.2812
       _cons |    -0.1251     0.0286    -4.38   0.000      -0.1811     -0.0691
------------------------------------------------------------------------------
```

On calcule donc facilement z: $\frac{-0.3595}{0.0400} = -9$. Si on prend son carré on obtient la statistique de test de la version standard (81). En comparant au test d'indépendance du khi2 sur le tableau croisé ?

Généralisation du test (version sur la loi "normale")

- Si on souhaite comparer la valeur de $\hat{b}$ avec une valeur quelconque c, la statistique de test est donnée par:

$$\boldsymbol{W = \frac{\hat{b} - c}{\sigma(\hat{b})}}$$

- Si on souhaite comparer deux contrastes estimés par le modèle, disons $\widehat{b_1}$ et $\widehat{b_2}$:

$$W = \frac{\widehat{b_1} - \widehat{b_2}}{\sigma(\widehat{b_1}, \widehat{b_2})}$$

Interprétation des résultats dans une régression multiple

On va maintenant partir d'un modèle plus enrichi

Paramètres estimés : logit et contrastes logistiques

highbp	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
sex						
Male	0.0000	(base)				
Female	-0.4263	0.0436	-9.79	0.000	-0.5117	-0.3410
race2						
Other	0.0000	(base)				
Black	0.5269	0.0736	7.16	0.000	0.3827	0.6711
age	0.0502	0.0015	33.16	0.000	0.0473	0.0532
hsize						
1	-0.2839	0.0770	-3.69	0.000	-0.4348	-0.1331
2	-0.1534	0.0666	-2.30	0.021	-0.2838	-0.0229
3	0.0000	(base)				
4	-0.0916	0.0783	-1.17	0.242	-0.2451	0.0618
5	-0.0071	0.0761	-0.09	0.926	-0.1563	0.1421
region						
NE	0.0000	(base)				
MW	-0.1378	0.0639	-2.16	0.031	-0.2631	-0.0125
S	-0.1421	0.0647	-2.20	0.028	-0.2689	-0.0152
W	-0.0469	0.0644	-0.73	0.467	-0.1732	0.0794
rural	-0.0193	0.0466	-0.41	0.679	-0.1105	0.0720
_cons	-2.3863	0.0989	-24.12	0.000	-2.5801	-2.1924

Odds et odds ratio

```
Logistic regression                               Number of obs     =     10,351
                                                  LR chi2(11)       =    1640.97
                                                  Prob > chi2       =     0.0000
Log likelihood =   -6230.28                       Pseudo R2         =     0.1164

------------------------------------------------------------------------------
      highbp | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |
        Male |     1.0000  (base)
      Female |     0.6529     0.0284    -9.79   0.000       0.5995      0.7111
             |
       race2 |
       Other |     1.0000  (base)
       Black |     1.6937     0.1246     7.16   0.000       1.4662      1.9564
             |
         age |     1.0515     0.0016    33.16   0.000       1.0484      1.0547
             |
       hsize |
           1 |     0.7528     0.0580    -3.69   0.000       0.6474      0.8754
           2 |     0.8578     0.0571    -2.30   0.021       0.7529      0.9773
           3 |     1.0000  (base)
           4 |     0.9124     0.0714    -1.17   0.242       0.7826      1.0637
           5 |     0.9929     0.0756    -0.09   0.926       0.8553      1.1526
             |
      region |
          NE |     1.0000  (base)
          MW |     0.8713     0.0557    -2.16   0.031       0.7687      0.9876
           S |     0.8676     0.0562    -2.20   0.028       0.7642      0.9849
           W |     0.9542     0.0615    -0.73   0.467       0.8410      1.0826
             |
       rural |     0.9809     0.0457    -0.41   0.679       0.8954      1.0747
       _cons |     0.0920     0.0091   -24.12   0.000       0.0758      0.1117
------------------------------------------------------------------------------
```

Variable quantitative

On va regarder la variable **age**

Pour une variable quantitative, l'Odds ratio mesure la variation des chances/risque lorsque la variable s'accroît d'une unité (dans l'exemple +1 an).
lorsque l'âge augmente d'un an, les chances/risque d'être en hypertension sont multiplié par 1.05.

- Cette variation est constante et linéaire : que l'on passe de 24 à 25 ans ou de 60 à 61, le risque est toujours multiplié par 1.05. Il s'agit en fait d'un effet multiplicatif moyen. Il possible de tester cette linéarité (voir en fin de section).

Remarque/rappel :

- Toujours faire attention à l'interprétation du terme constant (logit ou odds) avec ce type de variables. Ici la constante est calculé un risque d'hypertension à la naissance. Cette mesure n'est pas impossible, mais les données sont collectées pour des individus âgés d'au moins 20 ans.

- Effet d'échelle : il n'y a ici qu'une seule variable quantitative dans le modèle, les autres étant discrètes donc une échelle identique (valeur égales à 0 ou 1). Il faut se méfier des comparaisons, voire même éviter de comparer des paramètres estimés sur les échelles différentes. Il y a bien des techniques de standardisation, mais elle rend les paramètres plus difficilement interprétable (exemple : Odds ratio lu sur 1point de variation de l'écart-type de la variable).

Variable discrète/catégorielle

On s'intéressera à la variable **Région d'habitation**

Rappels:

- Pour l'estimation, les variables sont nécessairement transformées en indicatrices.
- Dans la section portant sur l'identification descriptive du modèle logistique, on a du poser une contrainte sur la variable dépendante pour éviter d'avoir une infinité de solutions (et donc ne pas pouvoir estimer le modèle).
  - Le même problème se pose pour les covariables. Cela peut apparaître trivial, car le principe du contraste logistique / odds ratio implique une comparasion, mais il n'est techniquement pas possible d'estimer le modèle si on impose pas une contrainte en amont aux variables.
  - Formellement on applique une contrainte aux paramètres, concrètement cette contrainte passe par le codage de la variable.
  - La traduction la plus répandue pour les variables discrètes, et généralement implémentée par défaut, est le codage dit à la référence.
  - Il en existe d'autres : l'*effect* ou le *deviation coding* (voir exemple plus bas), le codage adjacent pour les variables de type ordinale.

Codage à la référence versus effect coding

Codage à la référence

Si une variable discrète à k modalités, on va utiliser $k-1$ indicatrices. Pour celle qui est omise, les observations prennent donc toujours la valeur 0 dans les autres catégories. On pourrait également recoder l'indicatrice de la modalité en référence telle que toutes les observations soient égales à 0.

En terme d'interprétation pour un modèle logistique, le contraste logistique donne un écart ajusté, entre l'Odds d'une modalité d'une variable catégorielle et celui de la catégorie qui a été omise (ou toujours égale à 0).

Extrait du tableau de régression pour la variable *hsize*

```
** Contraste logistique (autres variables non reportées)
------------------------------------------------------------------------------
      highbp | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      region |
         NE  |     0.0000  (base)
         MW  |    -0.1378     0.0639    -2.16   0.031      -0.2631     -0.0125
          S  |    -0.1421     0.0647    -2.20   0.028      -0.2689     -0.0152
          W  |    -0.0469     0.0644    -0.73   0.467      -0.1732      0.0794
------------------------------------------------------------------------------

** Odds Ratio (autres variables non reportées)
------------------------------------------------------------------------------
      highbp | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      region |
         NE  |     1.0000  (base)
         MW  |     0.8713     0.0557    -2.16   0.031       0.7687      0.9876
          S  |     0.8676     0.0562    -2.20   0.028       0.7642      0.9849
          W  |     0.9542     0.0615    -0.73   0.467       0.8410      1.0826
------------------------------------------------------------------------------
```

L'indicatrice omise est le Nord-Est des Etats-Unis. *Une personne vivant dans le Middle West a donc des chances/risques d'être en hypertension 0.87 fois plus faible qu'une personne vivant dans le Nord-Est*.

N'importe quel contraste logistique et Odds ratio peut être calculé rapidement à partir d'une spécification particulière du modèle :

- Si on met en référence une personne vivant dans le Middle West, le contraste logistique entre une personne vivant dans le Nord-Est et une personne vivant dans le Middle West est simplement $-(-0.1378) = +0.1378$, et l'Odds Ratio $\frac{1}{0.1378} = 1.15$. Une personne vivant dans le Nord-Est a 1.15 fois plus de chance d'être en hypertension qu'une personne vivant dans le middle West. Le résultat est bien évidemment identique au précédent, seul le signe du contraste logistique a été modifié. Le résultat du test de Wald est donc lui-même strictement identique.

- On veut obtenir le contraste logistique ou l'Odds ratio entre deux modalités qui n'était pas à la référence. Par exemple on veut regarder l'Odds ratio entre une personne vivant dans l'Ouest et une personne vivant dans le Middle-West. Cette dernière modalité devient la référence.

  - ***Contraste***: $(W\ vs\ NE) - (MW vs\ NE) = (-0.0469) - (-0.1378) = +0.09$. L'exponentielle permet d'obtenir directement l'Odds ratio associé (1.094).
  - ***Odds ratio***: $\frac{OR(W\ vs\ NE)}{OR(MW vs\ NE)} = \frac{0.9542}{0.8713} = 1.094$. Une personne vivant dans l'Ouest a 1.1 fois plus de chances/risques en plus d'être en hypertension qu'une

personne vivant dans le Middle-West. Mais qu'en est-il de l'inférence (ou significativité de l'écart) ?

- ***On ne peut pas obtenir directement l'inférence sur l'écart avec les informations données par la paramétrisation d'origine (Nord-Est en référence)***, car on a besoin des covariances des paramètres des deux modalités comparées. On peut facilement les calculer à la main avec cette matrice, mais le plus rapide est de directement réestimer le modèle par simple changement de référence, ici le Middle-West.

```
Odds Ratio (autres variables non reportées)
------------------------------------------------------------------------------
      highbp | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      region |
         NE  |     1.1477     0.0734     2.16   0.031       1.0126      1.3009
         MW  |     1.0000  (base)
          S  |     0.9957     0.0593    -0.07   0.943       0.8860      1.1190
          W  |     1.0952     0.0660     1.51   0.131       0.9732      1.2324
-------------+----------------------------------------------------------------
```

Important: le choix de la référence est en principe neutre du point de vu de l'estimation, et est plutôt guidé par un impératif d'interprétation des résultats. Dans l'exemple, on choisit de mettre un focus sur la région Nord-Est, où le risque d'hypertension est comparé à celui des 4 autres régions.

Il faut néanmoins se méfier de la taille des effectifs, et il est fortement déconseillé de mettre en référence un groupe dont la taille est très faible (quelques dizaine d'observations au plus) ou si la différence d'effectifs avec les autres groupes est trop marquée.

Effect coding (variable binaire)

Très brièvement car son utilisation est particulièrement rare. Cela reste en revanche, la paramétrisation par défaut sur SAS.

- Proximité avec l'analyse de la variance
- On ne compare pas les logits ou Odds entre deux modalités d'une variable qualitative. On estime l'écart entre les logits des modalités par rapport à la moyenne des logits de ces modalités. On transforme toujours les variables en indicatrices, mais les valeurs sont différentes : par exemple -1 et 1 avec deux modalités dans la variable d'origine. Si la variable a plus de deux modalités : une avec la valeur de 1, une avec la valeur de -1 et les autres en 0.
- La valeur de ce contraste sera simplement la valeur du contraste estimé avec un codage à la référence divisé par deux. L'Odds ratio obtenu compare donc l'Odds d'une modalité à un Odds moyen, ou plutôt une moyenne d'Odds.
- Pour une spécification standard d'un modèle, ce codage ne présente pas d'avantage particulier par rapport au précédent. En présence d'interaction (croisement de

variables) il peut néanmoins faciliter la lecture de certains paramètres. Mais il reste toujours très rarement utilisé même dans ce cadre.

Interactions

Important:

- La prise en compte ou non d'interaction ne relève pas d'un quelconque biais statistique. On interroge seulement la stricte additivité du modèle, qui peut générer des approximations.
- Introduire des interactions, ou plus généralement des effets non linéaires dans le modèle, doit répondre à une question. Il faut rester parcimonieux, en particulier s'il n'y a pas d'utilité à complexifié un modèle sur des variables strictement de contrôle.

Quelques questions que l'on peut se poser sur le modèle précédent :

- Le risque d'hypertension est plus élevé pour les hommes que pour les femmes ($\frac{1}{0.6529} = 1.50$). De même ce risque est plus élevé pour les personnes afro américaine (1.69). Mais on voudrait savoir si ces deux informations tiennent la route une fois combinée : par simple additivité du modèle d'origine, les femmes afro-américaines doivent avoir un risque d'hypertension nettement moins élevé que les hommes afro-américain, ce qui signifie que la situation comparée entre femmes et hommes est identique quelle que soit leur origine. On peut donc tester la validité de ce résultat avec une interaction.
- On peut introduire une interaction entre une variable quantitative et une variable qualitative : le risque d'hypertension augmente avec l'âge et il est plus élevé pour les hommes que pour les femmes. Mais est-ce que l'augmentation - constante - du risque d'hypertension est la même pour les hommes que pour les femmes.
- Un cas particulier est la présence d'effets dit « quadratique ». Par exemple pour l'âge l'augmentation du risque d'hypertension qui a été estimé est constant quelle que soit la tranche d'âge, mais est-ce-que cela est vérifié ?

On présentera dans le détail que la situation d'une interaction ou d'un croisement entre deux variables qualitatives binaires, avec ensuite quelques résultats pour des croisements qui impliquant une variable quantitative.

Interaction entre deux variables qualitatives (binaire)

Extrait du modèle de départ

Contrastes logistiques

```
------------------------------------------------------------------------------
      highbp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |
        Male |     0.0000  (base)
      Female |    -0.4263     0.0436    -9.79   0.000      -0.5117     -0.3410
             |
       race2 |
       Other |     0.0000  (base)
       Black |     0.5269     0.0736     7.16   0.000       0.3827      0.6711
```

Odds ratio

```
------------------------------------------------------------------------------
      highbp | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |
        Male |     1.0000  (base)
      Female |     0.6529     0.0284    -9.79   0.000       0.5995      0.7111
             |
       race2 |
       Other |     1.0000  (base)
       Black |     1.6937     0.1246     7.16   0.000       1.4662      1.9564
```

Résumé rapide : on observe donc que le risque d'hypertension est moins chez les femmes et pour les personnes qui ne sont pas afro-américaines. On doit donc s'attendre à ce que dans la sous population afro-américaine, le risque d'hypertension soit plus faible pour les femmes que pour les hommes, avec éventuellement un Odds ratio avoisinant la valeur de 0.65. Qu'en est-il vraiment ?

Trois façons de procéder :

- On génère une nouvelle variable qui croisent les informations des deux variables d'origine.
- On introduit un *terme d'interaction.*
- On stratifie sur une des deux variables et on compare les résultats pour la variable introduite dans le modèle.

Création d'une variable croisée

Les effectifs croisés des deux variables sont les suivants :

```
           |  1 if race=black, 0
   1=male, |       otherwise
  2=female |         0          1 |     Total
-----------+----------------------+----------
      Male |     4,415        500 |     4,915
    Female |     4,850        586 |     5,436
-----------+----------------------+----------
     Total |     9,265      1,086 |    10,351
```

On génère la variable *blackfem* qui a donc 4 modalités :

```
        blackfem |      Freq.     Percent        Cum.
-----------------+-----------------------------------
black=0 female=0 |      4,415       42.65       42.65
black=0 female=1 |      4,850       46.86       89.51
black=1 female=0 |        500        4.83       94.34
black=1 female=1 |        586        5.66      100.00
-----------------+-----------------------------------
           Total |     10,351      100.00
```

Si on veut comparer le risque d'hypertension entre les femmes et les hommes d'origine afro américaine, on peut poser la catégorie de référence au niveau des hommes afro américain. Sous forme d'Odds ratios :

```
------------------------------------------------------------------------------
           highbp | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------------+-----------------------------------------------------------
         blackfem |
black=0 female=0  |     0.7131     0.0752    -3.21   0.001       0.5800      0.8769
black=0 female=1  |     0.4484     0.0471    -7.63   0.000       0.3650      0.5509
black=1 female=0  |     1.0000  (base)
black=1 female=1  |     0.8916     0.1181    -0.87   0.386       0.6877      1.1559
------------------------------------------------------------------------------
```

Alors que le risque d'hypertension était nettement moindre pour les femmes dans l'ensemble de la population, ce n'est pas le cas chez les femmes afro-américaine, dont le risque n'est diminué que de 11% avec les hommes (OR=0.89) avec un écart assez nettement non significatif.

On peut retrouver assez facilement l'Odds Ratio pour les femmes non afro américaine: $OR(non\ afro)_{f/h} = \frac{0.4484}{0.7131} = 0.63$

Si on fait le rapport entre les deux Odds Ratio, on mesure cet excédent de risque chez les femmes afro-américaine: $\frac{0.89}{0.63} = 1.42$. Il s'agira du terme d'interaction qui est estimé par la seconde façon de procéder.

Si on calcule les Odds ratio Afro-américain versus non Afro-américain pour les hommes et les femmes, ainsi que le rapport.

- Pour les homme: $\frac{1}{0.7131} = 1.40$ [+30%]
- Pour les femmes: $\frac{0.8916}{0.4484} = 1.99$ [+89%]
- Si on fait le rapport entre les deux OR, on retrouve bien 1.42.

Terme d'interaction

La façon la plus standard pour introduire une interaction, est d'estimer directement le rapport des OR précédent (ou la différence de leur logarithme).

Contrastes

```
------------------------------------------------------------------------------
      highbp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |
           0 |     0.0000  (base)
           1 |    -0.4639     0.0461   -10.06   0.000      -0.5544     -0.3735
             |
       black |
           0 |     0.0000  (base)
           1 |     0.3381     0.1055     3.21   0.001       0.1313      0.5448
             |
female#black |
         1 1 |     0.3492     0.1403     2.49   0.013       0.0743      0.6241
------------------------------------------------------------------------------
```

Rapports d'odds

```
------------------------------------------------------------------------------
      highbp | Odds Ratio* Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |
           0 |     1.0000  (base)
           1 |     0.6288     0.0290   -10.06   0.000       0.5744      0.6883
             |
       black |
           0 |     1.0000  (base)
           1 |     1.4022     0.1479     3.21   0.001       1.1404      1.7243
             |
female#black |
         1 1 |     1.4180     0.1989     2.49   0.013       1.0771      1.8667
------------------------------------------------------------------------------
* Le terme d'interaction n'est pas un OR, mais un rapport d'OR. Attention au libél
lé de la colonne
```

On retrouve les Odds ratio calculés avec la première méthode. Il faut donc bien interpréter les OR qui est affiché.

- L'OR de 0.63 est l'OR Femme versus homme pour les personnes non afro-américaine.
- L'OR de 1.40 est l'OR Afro-américain versus non afro américain pour les hommes.
- Les deux OR affichés doivent donc être lu pour la catégorie en référence de l'autre variable.

  - Femme versus homme : ce sont les non afro-américain en référence
  - Afro-américain versus non afro-américain: ce sont les hommes en référence
- Le terme d'interaction selon son signe ajoute ou retranche du risque qui permet de calculer les OR pour les modalités qui ne sont pas en référence :
  - femme versus homme chez afro-américain: $1.4180 \times 0.6288 = 0.89$
  - Afro-américain versus non afro-américain chez femmes: $1.4180 \times 1.4022 = 1.99$

Stratification des modèles

- Technique simple, couramment utilisé non pas pour introduire une seule interaction mais pour en tester un nombre plus important. Elle n'est donc pas utile si seulement un seul croisement est introduit. Attention à la taille des effectifs de la variable de stratification et donc à la comparaison des paramètres. Dans l'exemple, il est préférable de faire un modèle pour les hommes et un modèle pour les femmes plutôt qu'un modèle sur l'origine : l'échantillon est composé de 52% de femmes, et de 10% de personnes afro-américaines

	Homme	Femme
highbp		
black	1.393**	2.169***
	(3.19)	(7.28)
age	1.035***	1.072***
	(17.22)	(27.84)
1.hsize	0.755**	0.817*
	(-2.79)	(-2.28)
2.hsize	1	1
	(.)	(.)
3.hsize	1.009	1.374**
	(0.10)	(3.21)
4.hsize	1.007	1.119
	(0.07)	(1.00)
5.hsize	1.049	1.336**
	(0.51)	(2.72)
1.region	1.183	1.117
	(1.91)	(1.17)
2.region	1	1
	(.)	(.)
3.region	0.954	1.029
	(-0.56)	(0.33)
4.region	1.121	1.084

```
                   (1.38)          (0.91)

rural               0.987           0.988
                   (-0.20)         (-0.17)
--------------------------------------------
N                    4915            5436
--------------------------------------------
Exponentiated coefficients; t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001
```

On retrouve malgré des valeurs différentes, les mêmes effets que précédemment pour la variable black. Ce type de stratégie permet de comparer, d'un coup d'œil, si d'autres effets d'interaction sont présents dans le modèle selon le sexe de la personne : le risque augmente t-il plus pour les femmes selon l'âge par rapport aux hommes, peut-être des éléments de différences selon la taille du ménage mais plus difficile à investiguer lorsque sa composition n'est pas connue (monoparentalité…).

Remarques sur les interactions avec variables quantitatives

- La première approche n'est plus possible, on serait amené à estimer autant d'Odds ratio que de valeur d'une variable quantitative.
- Si on croise une variable qualitative et une variable quantitative on peut soit introduire directement le terme d'interaction, soit stratifier sur la variable qualitative. Pour le terme d'interaction, attention à l'Odds ratio affiché pour la variable qualitative, il sera estimé à 0 de la variable quantitative. On peut donc conseiller une version centrée sur la moyenne
- Si on croise deux variables quantitatives on peut seulement utiliser l'approche par le terme d'interaction.

Variable qualitative croisée avec une quantitative

Exemple avec le sexe et l'âge

Odds ratio et rapport d'Odds ratio (extrait)

```
--------------------------------------------

                     Age          Age centré
--------------------------------------------
highbp
0.female               1               1
1.female           0.136           0.603
                  (0.000)         (0.000)

age                1.037           1.037
                  (0.000)         (0.000)

female#agee        1.032
                  (0.000)
```

- Si on prend simplement la variable âge, l'OR pour les femmes a été estimé à un âge=0. Par construction, il est très faible (0.136)

- Si on prend la variable âge centrée sur sa moyenne, l'OR est estimé sur l'âge moyen de l'échantillon qui est de 47 ans, soit 0.60. Cette approche est préférable niveau lecture comme il n'y a pas d'information pour les personnes de moins de 20 ans.
- Pour les hommes, le risque d'être en hypertension augmente avec l'âge. Il est multiplié chaque année de 1.037.
- Le terme d'interaction indique ici un excès de risque d'hypertension pour les femmes par rapport au hommes lié au vieillissement. Ce risque est ici multiplié de 1.032. On peut également dire que le risque d'hypertension, va converger aux âges élevés entre hommes et femmes.

Odds ratios femme versus homme à 47 ans et à 60 ans :

Supposons un homme et une femme qui, en dehors de l'âge ont les mêmes caractéristiques, celles de la constante, dont l'Odds est fixé arbitrairement à 0.5 (logit= -0.69). On aura besoin des paramètres estimé sur l'échelle des logits et contraste logistique pour calculer ces Odds Ratios.

```
--------------------------
Age centré: échelle log
--------------------------
highbp
0.female               0
1.female          -0.505
                  (0.000)

age               0.0359
                  (0.000)

 female#age       0.0312
                  (0.000)
```

A 47 ans

- Pour un homme de 47 ans (âge centré =0), l'Odds d'être en hypertension est égal à :
  - $e^{(-0.69)} = 0.5$
- Pour une femme de 47 ans (âge centré =0), l'Odds d'être en hypertension est égal à :
  - $e^{(-0.69-0.50)} = 0.3$
- Donc à 47 ans, l'OR femme/homme du risque d'hypertension est tout simplement donné par le paramètre estimé soit 0.603

A 60 ans

- Pour un homme de 60 ans (âge centré =13), l'Odds d'être en hypertension est égal à :
  - $e^{(-0.69+0.0359\times13)} = 0.79$
- Pour une femme de 60 ans (âge centré =13), l'Odds d'être en hypertension est égal à :
  - $e^{(-0.69-0.505+0.0359\times13+0.0312*13)} = 0.72$
- Donc à 60 ans, l'OR femme/homme du risque d'hypertension est égal à :
  - $\frac{0.72}{0.79} = 0.91$

Le calcul de ces valeurs a été données à titre d'exemple, bien évidemment il faut à minima donner leur intervalle de confiance ou un résultat de test qui tiennent compte des effectifs

aux différents âges (effectifs à vérifier en amont). Si on enrichit l'information avec ce croisement on postule néanmoins un terme d'interaction constant selon l'âge, ce qui est une hypothèse plutôt forte.

On pourrait donc discrétiser la variable âge en créant quelques groupes d'âge et introduire une interaction entre le sexe (2 modalités) et une variable âge de type ordinale avec 3-4 tranches.

Remarques sur les effets quadratiques

- Une variable quantitative est introduite dans un modèle sous forme de fonction. Par défaut on stipule une stricte linéarité ou proportionnalité en la variable dépendante et cette variable: $f(x) = x$.
- N'importe qu'elle transformation peut être testée et introduite dans le modèle : carré de x, son logarithme….
- Un cas particulier, appelé *effet quadratique*, consiste à appliquer une interaction d'une variable quantitative sur elle-même. En se limitant, et ce n'est pas conseillé d'aller sans bonne raison au-delà, à un croisement simple (effet d'ordre2) consiste à écrire pour un modèle logistique : $log\left(\frac{p}{1-p}\right) = a + b_1x + b_2x^2 + \cdots$. Le signe de $b2$ va indiquer un renforcement ou une modération de l'OR lorsque x augmente.

Juste une rapide illustration avec l'âge (version centrée).

Dans le modèle d'origine, l'Odds ratio lié à l'âge était de 1.05: une différence d'un an multipliait le risque d'hypertension de 1.05, quelque que soit la tranche d'âge considéré. On va introduire un effet *effet quadratique* sur la variable, dans sa version centrée par rapport à la moyenne.

```
Contraste et double contraste logistique
------------------------------------------------------------------------------
       highbp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
          sex |
      Female  |    -0.4275     0.0436    -9.81   0.000      -0.5128     -0.3421
        black |     0.5327     0.0738     7.22   0.000       0.3881      0.6772

         age  |     0.0497     0.0015    32.35   0.000       0.0466      0.0527
        age^2 |    -0.0003     0.0001    -2.99   0.003      -0.0005     -0.0001
--------------+---------------------------------------------------------------
Odds ratio et rapport d'Odds Ratios
------------------------------------------------------------------------------
       highbp | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
          sex |
      Female  |     0.6522     0.0284    -9.81   0.000       0.5988      0.7103
        black |     1.7035     0.1256     7.22   0.000       1.4742      1.9684

         age  |     1.0509     0.0016    32.35   0.000       1.0478      1.0541
        age^2 |     0.9997     0.0001    -2.99   0.003       0.9995      0.9999
--------------+---------------------------------------------------------------
```

- Au niveau de l'âge moyen de l'échantillon, une augmentation d'un multiplie le risque de d'hypertension de 1.05
- L'effet quadratique ou d'interaction de l'âge avec lui-même indique que cet accroissement tend à se modérer ave
- On va juste donner les résultats, la méthode calcul est plus complexe et il n'est pas forcément utile de la retenir. De nouveau on considère des caractéristiques identiques pour les personnes dont on compare l'âge.
  - Entre 20 et 21 ans : l'Odds ratio est de 1.07
  - Entre 47 et 48 ans : l'Odds ratio est de 1.05
  - Entre 60 et 61 ans : l'Odds ratio est de 1.04

On observe donc un léger, mais présenté significatif, tassement de l'accroissement du risque d'hypertension selon l'âge.

Alternative : on peut créer une version ordinale de la variable quantitative. En choisissant bien les seuils de définition de chaque intervalle (ici des groupes d'âges), on est également en mesure de capter ces effets non linéaires Ils seront constants à l'intérieur de chaque intervalle.

Retour à l'échelle des probabilités: les effets marginaux

- Odds et Odds Ratio sont connus depuis très longtemps, mais pas de tous les publics, contrairement à la probabilité.
- Avec la régression linéaire, on peut lire les résultats en terme d'effet marginal: lorsque x augmente de 1 y augmente/baisse de b.
- Pour une probabilité, on peut faire un modèle à probabilité linéaire, mais il comporte des limites, fait l'objet de débats et n'est pas très largement diffusé.
- Avec un modèle logistique, sur l'échelle des Odds on lit les résultats sur une échelle multiplicative. Ce n'est pas toujours aisé. Sur l'échelle des logits, on a bien un effet marginal (le contraste logistique), mais il n'est pas directement interprétable.
- On ne peut pas passer directement sur une lecture sur l'échelle des probabilités avec la fonction logistique (pas mieux avec la version probit), les termes des autres covariables ne peuvent pas être annulé par simple soustraction.

Très diffusé en économie, une méthode existe pour lire par exemple les résultats d'un modèle logistique en terme d'écart en point de probabilité. Sa principale limite repose sur le fait qu'il n'y a pas une technique unifiée de calcul. Néanmoins celle appelé "Average Marginal Effect" est plutôt consensuelle et repose sur une technique de calcul simple et intuitive.

Petite remarque sur les logiciels: elle est depuis de nombreuses années implémentée dans Stata (commande **`margins`**). Un wrapper a été implémenté il y a quelques années dans R (librairie du même nom **`margins`**). Sur ce point Sas est plutôt en retrait. Il a toujours été possible de programmer manuellement leur calcul avec ce logiciel, mais l'opération est plutôt fastidieuse. Cedric Afsa propose dans son document de travail une macro Sas mais il reste,

tout du moins en France, une forte tradition d'utilisation exclusive des Odds Ratio dans les commentaires lorsque le logiciel utilisé est SAS (encore fortement utilisé en sociologie par exemple).

Principe de l'AME pour une variable discrète

- Avec le modèle logistique on affecte aux observations des probabilités conditionnelles selon la relation $p(y = 1 \mid x) = \frac{1}{1+e^{-xb}}$
- La moyenne de ces probabilités prédites est très proche de la probabilité non conditionnelle (fréquence de l'évènement analysé).
- Une variable qualitative, avec un codage à la référence, est représentée dans le modèle par une série de variables indicatrices $0,1$.
- Pour une caractéristique précise et pour chaque observation, on peut contrefactualiser la probabilité prédite : un homme devient une femme, une femme devient un homme (c'est plutôt amusant), un jeune devient un vieux et inversement …
- Comme au niveau de chaque observation on ne fait varier qu'une caractéristique, la contribution à la probabilité prédite reste constante pour les autres caractéristiques (variables quantitatives) car on se situe au niveau de chaque individu. La contrefactualisation permet donc de neutraliser, au niveau de chaque observation, l'effet des autres caractéristiques. C'est très malin !
- Pour chaque observation on fait donc varier une caractéristique binaire - 1 devient 0 ou 0 devient 1 - et on fait la différence des deux probabilités prédites obtenues, dont une est celle estimée par le modèle. L'effet marginal est donc calculé sur chaque individu.
- L'AME, effet marginal moyen, est alors simplement la moyenne de ces différences. Seule l'inférence de ce nouvel indicateur est plus complexe à obtenir (méthode Delta).
- L'AME permet alors de lire les résultats du modèle en termes d'écart en points de probabilité (en points de % si multiplié par 100).

AME pour les femmes avec une régression dont on a juste ajouté la variable black

Résultats de la régression

```
logit et contraste
------------------------------------------------------------------------------
      highbp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |    -0.3627     0.0400    -9.06   0.000      -0.4412     -0.2843
       black |     0.3491     0.0645     5.41   0.000       0.2226      0.4755
       _cons |    -0.1607     0.0294    -5.47   0.000      -0.2182     -0.1031
------------------------------------------------------------------------------

Odds et Odds Ratios
------------------------------------------------------------------------------
      highbp | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |     0.6958     0.0278    -9.06   0.000       0.6433      0.7526
       black |     1.4177     0.0915     5.41   0.000       1.2493      1.6089
       _cons |     0.8516     0.0250    -5.47   0.000       0.8039      0.9020
------------------------------------------------------------------------------
```

A partir de ce modèle on peut prédire les probabilités d'être en hypertension. Comme on a seulement 4 situations, les individus se voient prédire une des 4 probabilités possibles. La moyenne des probabilité prédite est égale à 0.423 (identique à la fréquence non conditionnelle).

- Femme non afro-américaine : p=0.37
- Femme afro-américaine : p=0.46
- Homme non afro-américain : p=0.46
- Homme afro-américain : p=0.55

A partir de la contrefactualisation, on va recalculer pour chaque observation une probabilité prédite en faisant seulement passer les femmes en homme. Sur la variable indicatrice utilisée par le modèle, les femmes passent donc en 0.

- On calcule pour chaque femme la nouvelle probabilité prédite (on ne modifie pas le modèle) avec ce changement de modalité. Si la femme n'est pas afro américaine elle se voit donc affecter la probabilité de 0.46, et si elle est afro-américaine la probabilité de 0.55.
- Pour chaque femme, on fait la différence entre ces deux probabilités. La contribution au calcul de la probabilité pour la variable *black* étant identique, elle se trouve donc neutralisée. Ici cette différence ne peut prendre que deux valeurs : -0.087 pour une non afro américaine et -0.090 pour une afro-américaine.
- On calcule la moyenne de ces différences pour obtenir cet AME.

Au final on trouve un AME de **-0.088pt**. Par rapport aux hommes, la probabilité d'être en hypertension est donc réduite de 0.088 points en moyenne (ou -8.8pt de pourcentage). On peut faire la même chose pour les hommes, et l'AME sera simplement égal à +0.088pt.

AME pour toutes les variables qualitatives du modèle. Celui relatif à l'âge est reproduit pour information, mais l'application de la technique est débattue.

Average Marginal Effect

	Delta-method				
	dy/dx	Std. Err.	z	P>|z|	[95% Conf. Interval]
sex					
Female	-0.0889	0.0090	-9.85	0.000	-0.1066 -0.0712
black	0.1092	0.0151	7.22	0.000	0.0796 0.1389
age	0.0104	0.0002	42.09	0.000	0.0099 0.0109
hsize					
2	0.0265	0.0132	2.00	0.045	0.0006 0.0525
3	0.0582	0.0157	3.71	0.000	0.0275 0.0889
4	0.0392	0.0166	2.36	0.018	0.0067 0.0717
5	0.0567	0.0161	3.52	0.000	0.0251 0.0883
region					
MW	-0.0286	0.0133	-2.15	0.031	-0.0547 -0.0026
S	-0.0295	0.0134	-2.20	0.028	-0.0559 -0.0032
W	-0.0098	0.0134	-0.73	0.467	-0.0361 0.0166

```
             |
     1.rural |    -0.0040     0.0096    -0.41   0.679      -0.0229      0.0149
------------------------------------------------------------------------------


Odds Ratios
------------------------------------------------------------------------------
      highbp | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |
        Male |     1.0000  (base)
      Female |     0.6529     0.0284    -9.79   0.000       0.5995      0.7111
             |
       black |     1.6937     0.1246     7.16   0.000       1.4662      1.9564
             |
         age |     1.0515     0.0016    33.16   0.000       1.0484      1.0547
             |
       hsize |
           1 |     1.0000  (base)
           2 |     1.1395     0.0744     2.00   0.045       1.0026      1.2950
           3 |     1.3283     0.1023     3.69   0.000       1.1423      1.5447
           4 |     1.2120     0.0990     2.35   0.019       1.0327      1.4225
           5 |     1.3189     0.1045     3.49   0.000       1.1292      1.5406
             |
      region |
          NE |     1.0000  (base)
          MW |     0.8713     0.0557    -2.16   0.031       0.7687      0.9876
           S |     0.8676     0.0562    -2.20   0.028       0.7642      0.9849
           W |     0.9542     0.0615    -0.73   0.467       0.8410      1.0826
             |
       rural |
           0 |     1.0000  (base)
           1 |     0.9809     0.0457    -0.41   0.679       0.8954      1.0747
             |
       _cons |     0.0692     0.0076   -24.20   0.000       0.0558      0.0860
------------------------------------------------------------------------------
```

On voit très clairement que les conclusions sur modèle sont identiques, la lecture plus intuitive : écart absolu au lieu d'un effet multiplicatif, concepts mieux partagé. En prenant la variable *black*: on passe d'une lecture type: que les chances/risques d'être en hypertension sont multipliés par 1.70 pour les personnes afro-américaines, à une lecture type: la probabilité d'être en hypertension est, en moyenne, de 0.11 point plus élevées pour les afro-américain.e.s.

Limites :

- On ne peut pas calculer directement cet indicateur avec un terme d'interaction dans le modèle. Elles sont introduites indirectement par croisement des AME.

- Bien que très populaire, il y a d'autres approches pour calculer ces effets marginaux. Contrairement aux lectures directes avec les OR, la démarche n'est donc pas unifiée.

- Pour les variables quantitatives plusieurs techniques existent, celle de l'AME fait l'objet de débats. Pour des variables avec une mesure bien identifié (comme l'âge) elle reste une

solution tout à fait utilisable, c'est moins le cas pour des variables quantitatives représentant par exemple des scores ou un indicateur construit (ou l'IMC dans la base de données).

Pour une discussion sur les effets marginaux, on peut se reporter au texte de Cedric Afsa page 28 (https://www.insee.fr/fr/statistiques/2022139).

Solutions logiciels

- **Stata**: le calcul des effets marginaux est depuis de nombreuses années implémentée dans Stata (commande `margins`).
- **R**: Un wrapper de la commande Stata a été implémenté il y a quelques années, avec une librairie du même nom `margins`.
- **Sas**: sur ce point plutôt en retrait. Il a toujours été possible de programmer manuellement le calcul des effets marginaux, mais l'opération est plutôt fastidieuse. Cedric Afsa, propose dans son document de travail une macro Sas mais il reste, tout du moins en France, une forte tradition d'utilisation exclusive des Odds Ratio dans les commentaires lorsque le logiciel utilisé est SAS (encore fortement utilisé en sociologie par exemple).

TP: Et l'ajout de l'IMC ça change quoi?

Introduire l'Indice de Masse Corporelle comme facteur de risque

QUALITÉ DU MODELE

- Boites à outil, rapide, qui va être réduite aux :
  - Tests et indicateurs construits sur la vraisemblance :
    - Test du rapport des vraisemblances (modèle imbriquées)
    - Pseudos-R2 (avec un s il y en moult versions pas tous basés sur la vraisemblance d'ailleurs)
    - Vraisemblances pénalisées (AIC et BIC), leur comparaison dans le cas non imbriqué
  - Indicateurs de classification : matrice de confusion et AUC (courbe de ROC)

Ne sera pas abordé (peut-être dans version ultérieure) :

- Le test d'Hosmer et Lemeshow: construction intéressante mais peu stable, plutôt en voie d'abandon
- Test de multicolinéarité et observations influentes.
-

Test et indicateurs reposant sur la vraisemblance

Rapports des vraisemblances et test

Rappel : la vraisemblance est formellement la probabilité d'observer l'échantillon, et est directement calculable après estimation du modèle. Implicitement, elle induit une mesure globale de l'erreur du modèle. Cette mesure de l'erreur est calculée via la **déviance** (reportée dans l'output de R par défaut).

Par analogie, la déviance est l'équivalent de la somme des carrés pour la régression linéaire. La vraisemblance du modèle est comparée à celle d'un modèle dit « saturée », tout simplement celui où $p(y = 1 \mid x) = y$. La vraisemblance du modèle saturé est donc égale à 1. Comme pour la vraisemblance, la deviance est exprimée en logarithme.

$$\boldsymbol{Deviance} = -\mathbf{2log}(\boldsymbol{L})$$

Test du rapport de vraisemblances

- **Seulement dans des situations imbriquées**: on compare un modèle à $(k + p)$ paramètres à un modèle à k paramètre(s).
  Si k égal à 0, le modèle de référence est le modèle sans covariable.

- On calcule le carré du rapport de ces deux vraisemblances, dont la valeur suit un khi2 à p degré(s) de liberté. De nouveau on passe par les logarithmes donc par la log-vraisemblance.

- Si N est suffisamment grand, un test du rapport des vraisemblances est asymptotiquement égal à un test d'indépendance du khi2.

Statistique du test:

$$\boldsymbol{LR = 2 \times (l_{k+p} - l_p)}$$

- LR suis un khi2 à p degré(s) de liberté.
- Au cas où : $H0: l_{k+p} = l_p$.

- Toujours positif car les log-vraisemblance sont négatives
- Remarque : pour la vraisemblance comme pour la déviance, leur nombre de degrés de liberté est égal au nombre d'observations - 1 plus le nombre de paramètres estimés (voir output par défaut de R). Dans le cadre d'un test de comparaison de deux vraisemblances imbriquées, le nombre d'observations est toujours identiques dans les deux modèles

Exemple : régression simple comparant le risque d'hypertension entre femmes et hommes versus le modèle sans covariable.

Modèle sans covariable

```
Log likelihood = -7050.7655

------------------------------------------------------------------------------
      highbp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |     -0.311      0.020   -15.65   0.000       -0.350      -0.272
------------------------------------------------------------------------------
```

Modèle conditionnel Femme versus homme

```
Log likelihood = -7010.1552

------------------------------------------------------------------------------
      highbp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |     -0.360      0.040    -9.00   0.000       -0.438      -0.281
       _cons |     -0.125      0.029    -4.38   0.000       -0.181      -0.069
------------------------------------------------------------------------------
```

- Calcul de la valeur de la région critique du test: $LR = -2 \times 40.6 = 81.22$
- Nombre de degré de liberté = 1
- La probabilité associée à la valeur critique est inférieure à 0.000

On vérifie que la statistique du test est (quasiment) identique à celle d'un test d'indépendance du khi2.

```
          observed frequency
          expected frequency

------------------------------
1 if      |
bpsystol  |
>=        |
140|bpdia |
st >= 90, |
0         |  1=male, 2=female
otherwise |     Male    Female
----------+-------------------
        0 |     2611      3364
          | 2837.129  3137.871
          |
        1 |     2304      2072
          | 2077.871  2298.129
------------------------------

          Pearson chi2(1) =  81.1787   Pr = 0.000
 likelihood-ratio chi2(1) =  81.2206   Pr = 0.000
```

On voit également que dans l'output du modèle $z^2 = 81$. On est pas loin non plus.

On veut comparer les modèles avec et sans la variable *region* :

- Dans le modèle, le nombre de paramètres associés à la variable région est égal à 3 .
- La log vraisemblance du modèle sans la variable région est égal à -5831.7123.
- La log vraisemblance du modèle avec la variable région est égal à -5827.8325.
- La statistique du test LR est donc égale 2*3.88 = 7.76.
- A 3 degrés de liberté, la probabilité d'être dans la région critique du test est égale à 0.0512.

Le tableau qui suit est l'extrait de l'output de la régression pour la variable *region* seulement.

```
------------------------------------------------------------------------------
      highbp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
.
.
.
      region |
         MW  |     -0.148      0.067    -2.23   0.026       -0.279      -0.018
          S  |     -0.141      0.066    -2.12   0.034       -0.271      -0.011
          W  |     -0.035      0.067    -0.52   0.601       -0.166       0.096
.
.
.

-------------+----------------------------------------------------------------

# test multiple : b1 – b2 – b3=0
-------------+----------------------------------
   Variables |       df      Khi2      Prob>chi2
-------------+----------------------------------
       region|        3    7.75716      0.051
-------------+----------------------------------
```

Que remarquez-vous ?

Pseudo R2

Très bref, j'admets ne pas les regarder (ou plutôt de ne plus les regarder).

- De nombreuses versions, principalement basée sur la vraisemblance
- Comme la vraisemblance augmente mécaniquement avec le nombre de paramètres, des versions pénalisent la mesure
- Interprétations pas simple, tout du moins difficilement assimilable au R^2 de la régression linéaire (% de la variance expliquée).
- Version Standard (Mac-Fadden): **Pseudo $R^2 = 1 - \frac{L_m}{L_0}$**, avec L_m la vraisemblance du modèle estimé et L_0 la vraisemblance du modèle non conditionnel. Au numérateur la vraisemblance du modèle approche la somme des erreurs au carré, différence entre y_i et p_i la probabilité prédite pour chaque observation i. Au dénominateur la vraisemblance du modèle non conditionnel approche la somme des carrés (différence entre y_i et $\bar{y}$).
- Mac Fadden a proposé un pseudo-R2 pénalisé (on déduit du numérateur le nombre de degré de liberté utilisé dns le modèle). Mais il peut être négatif.
- Il y aurait un certain consensus sur l'utilisation de la version de ***Nagelkerke***.

Vraisemblance pénalisée

- La vraisemblance augmente mécaniquement avec le nombre de paramètres ajoutés dans un modèle.
- Les tests du rapport de vraisemblance ne sont possibles qu'avec des modèles imbriqués (ajouts de paramètres) sur un même échantillon.
- Pour tester des modèles non imbriqués on peut utiliser des vraisemblances pénalisées par le nombre de paramètres et la taille de l'échantillon.
- Deux principaux indicateurs, appelés **critères**. Il s'agit de l'AIC et du BIC. Pour les comparaisons de modèles c'est principalement l'AIC qui est utilisé, il ne pénalise que le nombre de paramètres, mais moins fortement que le BIC. Pour le BIC, on pénalise en plus un plus faible nombre d'observations.
- Il n'y a pas de test associé, on retient seulement des seuils de différences entre deux modèles, une **préférence** (je ne parle pas de sélection) ira vers le **modèle dont la valeur du critère est la plus faible**.

Si k est le nombre de paramètres du modèle et N le nombre d'observations

- **Critère d'Akaike (AIC)**: $2 \times \big(k - log(L)\big)$
- **Critère d'Schwarz (BIC)**: $-2 \times log(L) + k \times \log(N)$

Ces deux critères reposent explicitement sur la déviance. C'est donc bien sur une valeur plus faible qu'une règle de décision peut porter.

Leur intérêt en pratique :

- Comparer deux modèles non imbriqués, par exemple avec un même nombre de degrés de liberté
- Comparer deux spécifications différentes d'une même covariable ou d'une même dimension explicative. Par exemple, même si cela n'est pas toujours conseillé, regarder la différence entre une version ordinale regroupée d'une variable quantitative avec sa version d'origine, à laquelle on peut ajouter ou non un effet quadratique.

Seuils de comparaison *(conventionnel)*

- Différence entre 0 et 2 : gain faible
- Différence entre 3 et 6 : gain positif
- Différence entre 7 et 10 : gain fort
- Différence supérieure à 10 : gain très fort

Application : on va comparer la spécification du modèle avec la variable brute pour l'âge (1 paramètres) et sa version regroupée assez arbitrairement (4 groupes d'âge: moins de 30 ans, 30-49, 50-59, 60 et plus).

Variable âge d'origine (1ddl) [extrait]

```
------------------------------------------------------------------------------
      highbp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
.
.
         age |      0.047      0.002    29.63   0.000        0.044       0.050
.
.
-------------+----------------------------------------------------------------


Critères d'information
-----------------------------------------------------------------------------
       Model |          N   ll(null)  ll(model)      df        AIC        BIC
-------------+---------------------------------------------------------------
           . |     10,351  -7050.765  -5827.832      12   11679.66    11766.6
-----------------------------------------------------------------------------
```

Variable âge avec regroupement [extrait]

```
------------------------------------------------------------------------------
      highbp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
.
.
        cage |
          2  |      0.740      0.073    10.11   0.000        0.597       0.884
          3  |      1.580      0.084    18.86   0.000        1.415       1.744
          4  |      1.903      0.071    26.64   0.000        1.763       2.043
.
.
-------------+----------------------------------------------------------------

-----------------------------------------------------------------------------
       Model |          N   ll(null)  ll(model)      df        AIC        BIC
-------------+---------------------------------------------------------------
           . |     10,351  -7050.765  -5875.681      14   11779.36   11880.79
-----------------------------------------------------------------------------
```

Pour l'AIC et comme pour le BIC, la préférence va aller nettement vers la version d'origine de la variable (AIC= 11679 versus AIC=11779 pour les groupes d'âge). Il n'y a rien d'étonnant à cela, à l'intérieur de chaque groupe d'âge, la contribution du facteur âge au risque d'hypertension reste constant alors qu'elle continue d'augmenter dans la version d'origine de la variable.

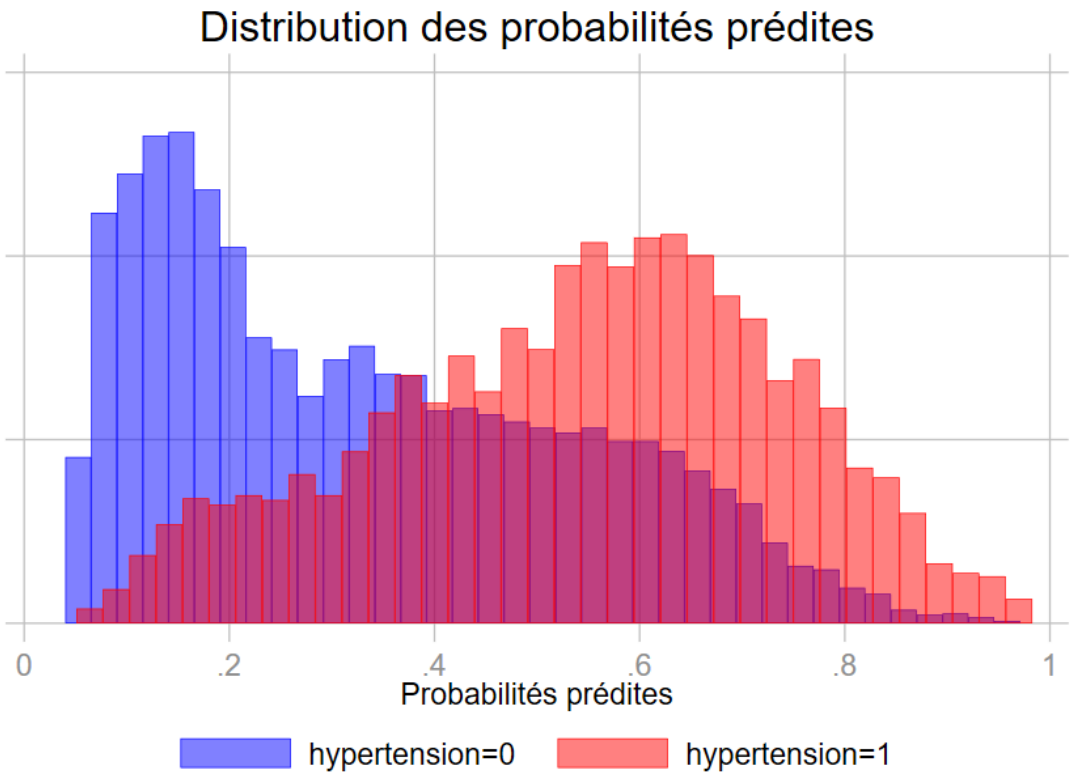
Classification : concordances et courbe de ROC

Partons d'un graphique simple. On peut recommander de le réaliser systématiquement après avoir estimé un modèle logistique (Hosmer et Lemeschow). Il s'agit tout simplement de représenter la distribution des probabilités prédites par le modèle selon la réponse observée, dans l'exemple selon que les personnes soient ou non en situation d'hypertension.

Modèle

```
Logistic regression                                     Number of obs     =     10,351
                                                        LR chi2(13)       =    2350.17
                                                        Prob > chi2       =     0.0000
Log likelihood = -5875.6815                             Pseudo R2         =     0.1667

------------------------------------------------------------------------------
      highbp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |     -0.477      0.045   -10.55   0.000       -0.566      -0.389
       black |      0.389      0.075     5.18   0.000        0.242       0.536
             |
        cage |
           2 |      0.740      0.073    10.11   0.000        0.597       0.884
           3 |      1.580      0.084    18.86   0.000        1.415       1.744
           4 |      1.903      0.071    26.64   0.000        1.763       2.043
             |
      region |
          MW |     -0.133      0.066    -2.00   0.045       -0.263      -0.003
           S |     -0.126      0.066    -1.90   0.057       -0.255       0.004
           W |     -0.024      0.067    -0.36   0.718       -0.155       0.107
             |
       hsize |
           2 |      0.040      0.067     0.59   0.553       -0.092       0.172
           3 |      0.080      0.080     1.00   0.317       -0.077       0.238
           4 |     -0.050      0.087    -0.58   0.561       -0.220       0.120
           5 |     -0.002      0.085    -0.03   0.978       -0.170       0.165
             |
         bmi |      0.135      0.005    26.40   0.000        0.125       0.145
       _cons |     -4.695      0.156   -30.12   0.000       -5.001      -4.390
------------------------------------------------------------------------------
```



On peut aborder la classification du couple observations-prédictions de deux manières (qui se recoupent au final).

Paires concordantes et discordantes

On compare la probabilité prédite par le modèle de toutes les paires d'observations dont le niveau de réponse est différent :

- Une paire d'observations $y_i = 1$ et $y_j = 0$ sera concordante si $p(y_i) > p(y_j)$
- Une paire d'observations $y_i = 1$ et $y_j = 0$ sera discordante si $p(y_i) < p(y_j)$
- Un % de paires concordantes supérieur ou égal à 70% serait plutôt bon signe sur la qualité du modèle (je n'ai toujours pas trouvé de règle de décision précise sur ce type d'indicateur)
- Limites :
  - Sensible à la rareté de l'évènement. Lorsque l'évènement est en deçà de 10% ou 20%, le % de concordance dépasse facilement les 80% (voir plus bas : très peu de « faux négatifs »).
  - Classification invariante aux différences de probabilités prédites : on évalue de la même manière une différence de +.5pt et une différence de +.001pt.
  - Par définition une probabilité prédite n'est que le reflet de toutes les valeurs prises par les covariables introduites. Des « jumeaux » parfaits ont donc des probabilités prédites identiques. Quelle règle doit-on appliquer pour les paires d'observations dont la valeur de y diffère

(situation courante avec des covariables seulement catégorielle, disparaît souvent avec de la présence de variables quantitatives).

Matrices de confusion et courbes de ROC

- On fixe un seuil pour la probabilité prédite et on compte pour les deux niveaux de la réponse le nombre d'observations dont la probabilité prédite inférieur ou supérieure de ce seuil. On construit un tableau appelé **matrice de confusion**.
  - Par défaut ce seuil est souvent fixé à 0.5 (on le retient pour la suite)
  - Si $y = 1$ et $p(y)$>.5 l'observation est appelée ***vrai positif***
  - Si $y = 1$ et $p(y)$<.5 l'observation est appelée ***faux positif***
  - Si $y = 0$ et $p(y)$<.5 l'observation est appelée ***vrai négatif***
  - Si $y = 0$ et $p(y)$>.5 l'observation est appelée ***faux négatif***
  - La somme ou le pourcentage de **vrais positifs** et ***négatifs*** donne l'importance des observations bien classées.

Pour le risque d'hypertension, la matrice de confusion au seuil de 50% est donnée par le tableau suivant.

```
              -------- True --------
Classified |        Y=1         Y=0   |      Total
-----------+--------------------------+-----------
    p>.5   |      2721          1427  |       4148
    p<.5   |      1655          4548  |       6203
-----------+--------------------------+-----------
   Total   |      4376          5975  |      10351
```

- Le ***% de vrais positifs*** est calculé en rapportant au nombre total d'observations où p>.5 : %VP=2721/4148=>*65%*.
- Le ***% de vrais négatifs*** est calculé en rapportant au nombre total d'observations où p<.5 : %VN=4548/6203=>*73%*.

On a donc utilisé ici les lignes du tableau. En rapportant les valeurs sur la diagonale au nombre total d'observations, on obtient un nombre/pourcentage d'observations « **bien classés** »: (2721+4548)/10351 => *70.2%*

On a calculé les % de vrais positifs et négatifs du point de vue de la probabilité prédite. On peut appliquer le même type de raisonnement du point de vue de la valeur observée. On obtient alors les notions de **sensibilité** et **spécificité**, assez médiatisé il y a deux ans avec la mise en place des tests PCR lors de la première vague de l'épidémie COVID. Elles sont préférées aux deux premiers pourcentage car on les rapporte à ce qui est observé et non à ce qui est prédit.

- La ***sensibilité*** (à seuil donnée) rapporte le nombre de *vrais positifs* au nombre de personne dont la réponse est 1, dans notre exemple les personnes en hypertension. Le dénominateur représente donc la somme du nombre de vrais positifs et et de faux négatifs : Sensibilité (%) = 2721/4376=> 62.2%.

- La ***spécificité*** (à seuil donnée) rapporte le nombre de *vrais négatifs* au nombre de personne dont la réponse est 0, dans notre exemple les personnes qui ne sont pas en hypertension. Le dénominateur représente donc la somme du nombre de faux positifs et de vrais négatifs : Spécificité (%) = 4548/5975=> 76.2%. Comme on change seulement les marges, le % de "bien classés" est identique.

Limite : on a raisonné sur un seuil conventionnel, et toutes les mesures précédentes vont varier avec ce seuil. On retrouve comme pour la concordance des effets de voisinage : on évalue identiquement des probabilités prédites proches ou éloignée du seuil.

Solution : On calcule la sensibilité et la spécificité en faisant varier le seuil de 0 à 1 et on les tracer sur un graphique.

Cette courbe est appelée **courbe de ROC**, dont l'origine pendant la seconde guerre mondiale afin de tester l'efficacité des radars et sonars pour détecter les bruits de fond (faux positifs) : oiseaux, bancs de sardines…).

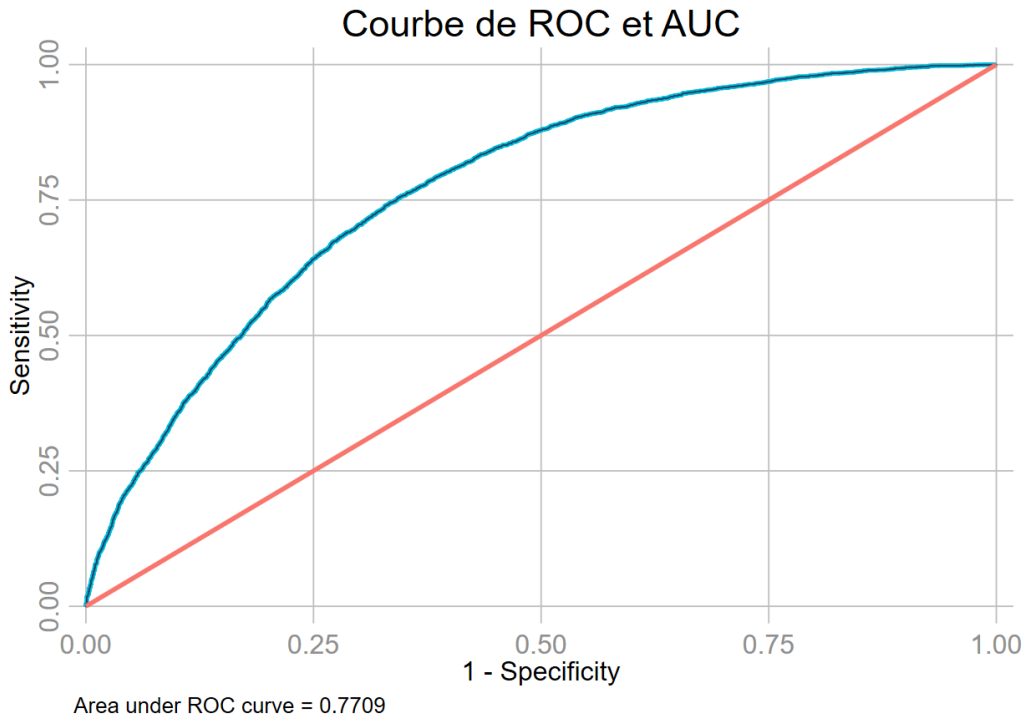
- Pour faciliter la lecture du graphique on peut représenter (1-spécificité = faux positif sur nombre de y=0) ou inverser l'ordre de l'axe des abscisses.
- Un « bon modèle » est un modèle qui présente à tous les seuils un niveau élevé de sensibilité et un niveau faible de spécificité.
- La diagonale représente la situation où la réponse observée est affectée de manière totalement aléatoire : c'est le modèle sans covariable.

  - On peut calculer l'aire entre la courbe de ROC et la diagonale, appelée **AUC** (*Area Under Curve*). Elle mesure la capacité « discriminante » du modèle, c'est-à-dire à quel point les caractéristiques introduites dans le modèle vont affecter la réponse.
  - Sa valeur minimale est de 0.5 et correspond à une courbe qui se confond avec la diagonale. Il s'agit de la valeur pour de l'AUC pour le modèle vide (constante seulement)
  - Sa valeur maximale est de 1 et correspond à des prédiction parfaite (modèle saturée vu plus haut dans le calcul de la déviance)
  - Une valeur de 0.7 peut être considéré comme seuil minimal à atteindre pour un modèle en sciences sociales.

Valeur de la sensibilité aux deux seuils extrêmes :

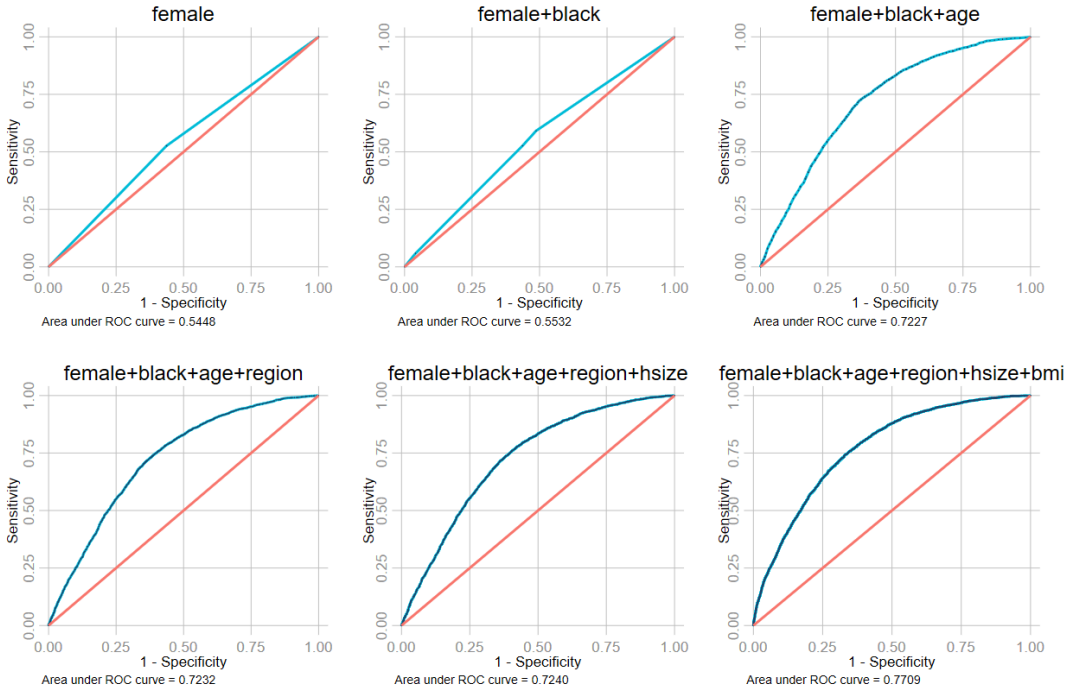
- Si le seuil est égal à 0 : par définition sensibilité=0 et spécificité=1 (1 - spécificité =0). Le % d'observations bien classées est égal à la proportion d'observations tel que y=0 (57.7% pour l'exemple sur l'hypertension).
- Si le seuil est égal à 1 : par définition sensibilité=1 et spécificité=0 (1 - spécificité =1). Le % d'observation bien classé est égal à la proportion d'observation tel que y=1 (42.3% pour l'exemple sur l'hypertension.
- Les courbes commencent et finissent donc sur les mêmes valeurs.

Application pour le risque d'hypertension : pour le modèle final (avec l'IMC) l'aire sous la courbe de ROC est égale à 0.771. C'est plutôt correct.

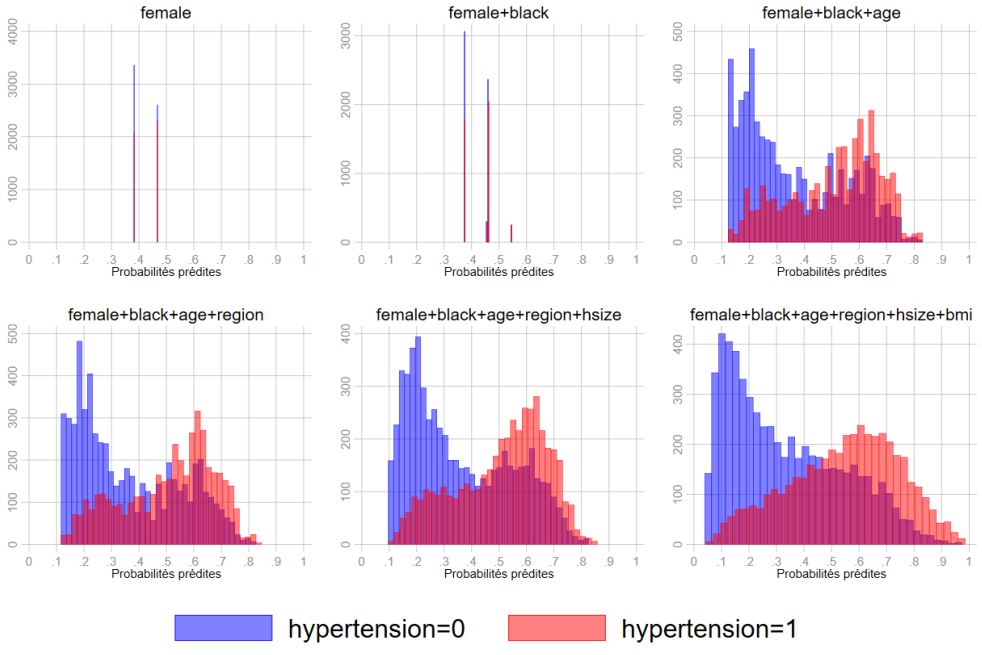


Le graphique suivant montre la transformation de la courbe de ROC et les valeurs successives de l'AUC par introduction successives des variables. On a également ajouté un graphique donnant la transformation des histogrammes de la distribution des probabilités prédites (attentions aux échelles sur l'axe des ordonnées qui se modifient)

Courbe de ROC et AUC



Distribution des probabilités prédites



…. to be continued avec la régression polytomique

APPLICATION LOGICIELS

Sas et Python à venir…..

R

Ouverture de la base

```
hypertension <- read.csv("//ined.fr/Partages/SMS/5- Utilisateurs/Préparation COURS
/2022- MT AM/logit/bases/hypertension.csv")

df = hypertension
```

Estimation et lecture du modèle

Estimation

Procédure standard avec une fonction **glm** générique (la fonction lm estimera une regression linéaire par les moindres carrés ordinaires).

- On indique la combinaison linéaire
- On précise le type d'estimation, ici la densité ou plutôt la fonction de masse de bernouilli (`family=binomial`). Par défaut R utilise le lien logistique, il n'est donc pas nécessaire de le préciser.
- On indique classiquement la source des données
- Il y a mieux (voir plus bas), on affiche l'output avec les paramètres en colonne et quelques éléments d'inférence avec la fonction **`summary()`**

Par rapport au support de formation on part du modèle final avec les variables *female*, *black*, *age*, *region*, *hsize* et *bmi*.

```
fit = glm(highbp ~ female + black + age + region + hsize + bmi, family=bino
mial, data=df)
summary(fit)

##
## Call:
## glm(formula = highbp ~ female + black + age + region + hsize +
##     bmi, family = binomial, data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7041  -0.9023  -0.4797   0.9843   2.4929
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.981818   0.172263 -34.725  < 2e-16 ***
## female      -0.483227   0.045297 -10.668  < 2e-16 ***
## black        0.386141   0.075124   5.140 2.75e-07 ***
## age          0.046854   0.001529  30.640  < 2e-16 ***
## regionNE     0.152511   0.066512   2.293   0.0218 *
## regionS      0.011232   0.061484   0.183   0.8550
## regionW      0.113434   0.062746   1.808   0.0706 .
## hsize        0.015598   0.018744   0.832   0.4053
## bmi          0.135256   0.005120  26.416  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 14102  on 10350  degrees of freedom
## Residual deviance: 11659  on 10342  degrees of freedom
## AIC: 11677
##
## Number of Fisher Scoring iterations: 4
```

Lecture de l'output par défaut

- Les éléments liés à la déviance sont décrits plus bas. Néanmoins pour l'output par défaut, la mesure de la déviance du modèle dit *NULL* permet de récupérer le nombre d'observations qui ont été utilisées, et donc mesurer l'importance des observations exclues en raison des valeurs manquantes). Il s'agit du nombre de degrés de liberté du « NULL MODEL » + 1. Ici le nombre d'observations est égal à 10351 (10350 degrés de liberté sur les observations + 1)

- Paramètres estimés : échelle logarithmique, soit le logit pour la constante et les contrastes logistisques [Estimate]

- L'erreur sur le paramètre [Std.Error]
- La statistique z. R calcule le test de Wald pour **Estimate=0** sur la loi normale [z value]. z est égal au rapport du paramètre et de l'erreur type (-33.887 = -5.799/0.1711).

- La probabilité de ne pas être à l'extérieur de la région critique est donnée par Pr(>|z|). Selon l'humeur, la probabilité est affichée ou non sous format scientifique. Ce n'est pas top. Des librairies permettent d'afficher des outputs sans ce type de report (voir plus bas).

- L'output affiche également des seuils critiques via les incontournables *. La valeur des seuils est indiquée en dessous du tableau.

  Petite remarque : il est intéressant de constater que R maintient cette information, alors que le report de ses seuils est de plus en plus critiqués, certaines revues commençant à les interdire). Voir la declaration de l'ASA (https://www.tandfonline.com/doi/full/10.1080/00031305.2016.1154108): point 3 "***Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold***".

Variables qualitatives

Format caractère

Si la variable n'est pas directement une indicatrice, la solution par défaut est l'ordre alphabétique pour une variable caractère : par exemple ici MW pour région. Pour modifier la référence, il faut la passer en format facteur (indicatrices).

Format numérique

Pour la variable *hsize*, comme l'information est numérique, elle est traitée comme une variable quantitative. Pour la traiter en variable qualitative avec un niveau de référence, soit utiliser factor(nom_variable) dans la formule, soit la transformer en facteur en amont. Cela permettra de changer le niveau de la référence.

```
fit = glm(highbp ~ female + black + age + region + factor(hsize) + bmi, fam
ily=binomial, data=df)
summary(fit)

##
## Call:
## glm(formula = highbp ~ female + black + age + region + factor(hsize) +
##     bmi, family = binomial, data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7034  -0.9027  -0.4785   0.9859   2.4998
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -5.996686   0.168517 -35.585  < 2e-16 ***
## female         -0.481039   0.045446 -10.585  < 2e-16 ***
## black           0.389612   0.075421   5.166 2.39e-07 ***
```

```
## age             0.046847   0.001581  29.627  < 2e-16 ***
## regionNE        0.148269   0.066584   2.227   0.0260 *
## regionS         0.007580   0.061531   0.123   0.9020
## regionW         0.113286   0.062793   1.804   0.0712 .
## factor(hsize)2  0.051834   0.068020   0.762   0.4460
## factor(hsize)3  0.153500   0.080330   1.911   0.0560 .
## factor(hsize)4  0.044942   0.085410   0.526   0.5988
## factor(hsize)5  0.079785   0.082928   0.962   0.3360
## bmi             0.135067   0.005122  26.370  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 14102  on 10350  degrees of freedom
## Residual deviance: 11656  on 10339  degrees of freedom
## AIC: 11680
##
## Number of Fisher Scoring iterations: 4
```

Passage en indicatrice (facteur) des variables régions et hsize.

```
df$region = as.factor(df$region)
df$hsize = as.factor(df$hsize)
```

Résultat identique au précédent

```
#fit = glm(highbp ~ female + black + age + region + hsize + bmi, family=bin
omial, data=df)
#summary(fit)
```

Affichage des Odds Ratios

Utiliser la fonction *odds.ratio()* de la librairie **questionr**. Elle se base sur l'output par défaut, et permet de récupérer les intervalles de confiance.

```
#install.packages("questionr")
library(questionr)
```

Récupération du tableau de résultat avec les Odds et les Odds Ratio

```
odds.ratio(fit)

## Waiting for profiling to be done...

##                        OR     2.5 %  97.5 %         p
## (Intercept)    0.0024870 0.0017829 0.0035 < 2.2e-16 ***
## female         0.6181410 0.5653929 0.6757 < 2.2e-16 ***
## black          1.4764085 1.2735604 1.7118 2.394e-07 ***
## age            1.0479612 1.0447360 1.0512 < 2.2e-16 ***
## regionNE       1.1598246 1.0179235 1.3215   0.02596 *
## regionS        1.0076089 0.8931207 1.1368   0.90195
## regionW        1.1199518 0.9902745 1.2667   0.07121 .
## factor(hsize)2 1.0532010 0.9217579 1.2034   0.44603
## factor(hsize)3 1.1659073 0.9961738 1.3649   0.05602 .
## factor(hsize)4 1.0459670 0.8847453 1.2366   0.59876
```

```
## factor(hsize)5 1.0830545 0.9206296 1.2743   0.33600
## bmi            1.1446134 1.1332515 1.1562 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Changement de référence

Avec `relevel`

```
df$region <- relevel(df$region, ref = "NE")
df$hsize  <- relevel(df$hsize,  ref = "2")

fit = glm(highbp ~ female + black + age + region + hsize + bmi, family=bino
mial, data=df)
summary(fit)
```

Pour info, comment on transforme une variable en indicatrice manuellement.

- On peut créer les indicatrices avec la fonction `ifelse()`
- Utiliser la librairie **`fastDummies`** (solution ci-dessous)

```
# install.packages('fastDummies')
library(fastDummies)
```

On génère les 4 indicatrices (nommées region_NE, region_MW, region_S, region_W)

```
df <- dummy_cols(df, select_columns = 'region')
```

Si on met NE à la référence, elle est omise dans la combinaison linéaire (formule)

```
fit = glm(highbp ~ female + black + age + region_MW + region_S + region_W + hsize
+ bmi, family=binomial, data=df)
summary(fit)

##
## Call:
## glm(formula = highbp ~ female + black + age + region_MW + region_S +
##     region_W + hsize + bmi, family = binomial, data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7034  -0.9027  -0.4785   0.9859   2.4998
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.848417   0.170850 -34.231  < 2e-16 ***
## female      -0.481039   0.045446 -10.585  < 2e-16 ***
## black        0.389612   0.075421   5.166 2.39e-07 ***
## age          0.046847   0.001581  29.627  < 2e-16 ***
## region_MW   -0.148269   0.066584  -2.227   0.0260 *
## region_S    -0.140689   0.066333  -2.121   0.0339 *
## region_W    -0.034983   0.066975  -0.522   0.6014
## hsize2       0.051834   0.068020   0.762   0.4460
```

```
## hsize3       0.153500   0.080330   1.911   0.0560 .
## hsize4       0.044942   0.085410   0.526   0.5988
## hsize5       0.079785   0.082928   0.962   0.3360
## bmi          0.135067   0.005122  26.370  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 14102  on 10350  degrees of freedom
## Residual deviance: 11656  on 10339  degrees of freedom
## AIC: 11680
##
## Number of Fisher Scoring iterations: 4
```

Par exemple pour l'indicatrice *region_NE*:

```
table(df$region,df$region_NE)

##
##          0    1
##   MW 2774    0
##   NE    0 2096
##   S  2853    0
##   W  2628    0
```

Interaction

Assez classiquement, les variables sont croisées dans la fonction `glm` avec une *

Par exemple pour une intéraction entre les variables *female* et *black*

```
fit = glm(highbp ~ female*black + age + region + hsize + bmi, family=binomial, dat
a=df)
summary(fit)

##
## Call:
## glm(formula = highbp ~ female * black + age + region + hsize +
##     bmi, family = binomial, data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7006  -0.9024  -0.4787   0.9857   2.4996
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -5.999657   0.169779 -35.338  < 2e-16 ***
## female       -0.478798   0.048008  -9.973  < 2e-16 ***
## black         0.400735   0.107668   3.722 0.000198 ***
## age           0.046843   0.001581  29.621  < 2e-16 ***
## regionNE      0.148250   0.066585   2.226 0.025982 *
## regionS       0.007575   0.061531   0.123 0.902021
## regionW       0.113237   0.062794   1.803 0.071340 .
## hsize2        0.052159   0.068058   0.766 0.443448
```

```
## hsize3        0.153927   0.080385   1.915 0.055510 .
## hsize4        0.045308   0.085450   0.530 0.595950
## hsize5        0.079991   0.082943   0.964 0.334838
## bmi           0.135135   0.005144  26.270  < 2e-16 ***
## female:black -0.021402   0.147827  -0.145 0.884889
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 14102  on 10350  degrees of freedom
## Residual deviance: 11656  on 10338  degrees of freedom
## AIC: 11682
##
## Number of Fisher Scoring iterations: 4
```

Estimation des effets marginaux (AME)

- Wrapper de la commande *margins* de Stata.
- Le nom de la librairie est identique.
- Les AME (Average Marginal Effect) sont calculés avec la fonction du même nom.

```
# install.packages("margins")
library(margins)
```

Pour estimer les *effets marginaux* de toutes les variables présentent dans le modèle (attention le modèle ne doit pas comporter d'effet d'interaction), on a juste à appliquer seulement l'objet `fit` à la fonction.

```
ame = margins(fit)
summary(ame)

##    factor     AME     SE        z      p   lower   upper
##       age  0.0089 0.0003  35.2816 0.0000  0.0085  0.0094
##     black  0.0745 0.0144   5.1900 0.0000  0.0464  0.1027
##       bmi  0.0258 0.0009  29.9378 0.0000  0.0241  0.0275
##    female -0.0919 0.0085 -10.7862 0.0000 -0.1086 -0.0752
##    hsize2  0.0099 0.0129   0.7666 0.4433 -0.0154  0.0353
##    hsize3  0.0294 0.0153   1.9188 0.0550 -0.0006  0.0594
##    hsize4  0.0086 0.0162   0.5306 0.5957 -0.0232  0.0404
##    hsize5  0.0152 0.0158   0.9656 0.3343 -0.0157  0.0461
##  regionNE  0.0284 0.0127   2.2253 0.0261  0.0034  0.0533
##   regionS  0.0014 0.0117   0.1231 0.9020 -0.0215  0.0244
##   regionW  0.0216 0.0120   1.8035 0.0713 -0.0019  0.0451
```

Autres mise en forme de l'output

Plusieurs librairies proposent des fonctions qui permettent une meilleure visualisation des résultats que l'output de défaut. 2 exemple avec **`jtools`** et **`gtsummary`** (voir le site de J.Larmarange pour plus de détails). Pour les tableaux **`gtsummary`** est supérieur.

```
fit = glm(highbp ~ female + black + age + region + hsize + bmi, family=bino
mial, data=df)
```

jtools

```
# install.packages('jtools')
# install.packages("huxtable")
# install.packages("ggstance")
#install.packages("broom.mixed")
 library(jtools)
```

Outputs type console

- Fonction **`summ`**
- Détails: https://jtools.jacob-long.com/articles/summ.html

```
summ(fit)

## MODEL INFO:
## Observations: 10351
## Dependent Variable: highbp
## Type: Generalized linear model
##   Family: binomial
##   Link function: logit
##
## MODEL FIT:
## <U+03C7>²(11) = 2445.87, p = 0.00
## Pseudo-R² (Cragg-Uhler) = 0.28
## Pseudo-R² (McFadden) = 0.17
## AIC = 11679.66, BIC = 11766.60
##
## Standard errors: MLE
## ------------------------------------------------
##                       Est.   S.E.   z val.      p
## ----------------- ------- ------ -------- ------
## (Intercept)         -6.00   0.17   -35.59   0.00
## female              -0.48   0.05   -10.58   0.00
## black                0.39   0.08     5.17   0.00
## age                  0.05   0.00    29.63   0.00
## regionNE             0.15   0.07     2.23   0.03
## regionS              0.01   0.06     0.12   0.90
## regionW              0.11   0.06     1.80   0.07
## hsize2               0.05   0.07     0.76   0.45
```

```
## hsize3               0.15   0.08     1.91   0.06
## hsize4               0.04   0.09     0.53   0.60
## hsize5               0.08   0.08     0.96   0.34
## bmi                  0.14   0.01    26.37   0.00
## ------------------------------------------------
```

```
summ(fit, confint=TRUE)
```

```
## MODEL INFO:
## Observations: 10351
## Dependent Variable: highbp
## Type: Generalized linear model
##   Family: binomial
##   Link function: logit
##
## MODEL FIT:
## <U+03C7>²(11) = 2445.87, p = 0.00
## Pseudo-R² (Cragg-Uhler) = 0.28
## Pseudo-R² (McFadden) = 0.17
## AIC = 11679.66, BIC = 11766.60
##
## Standard errors: MLE
## ------------------------------------------------------------
##                       Est.    2.5%   97.5%   z val.      p
## ----------------- ------- ------- ------- -------- ------
## (Intercept)         -6.00   -6.33   -5.67   -35.59   0.00
## female              -0.48   -0.57   -0.39   -10.58   0.00
## black                0.39    0.24    0.54     5.17   0.00
## age                  0.05    0.04    0.05    29.63   0.00
## regionNE             0.15    0.02    0.28     2.23   0.03
## regionS              0.01   -0.11    0.13     0.12   0.90
## regionW              0.11   -0.01    0.24     1.80   0.07
## hsize2               0.05   -0.08    0.19     0.76   0.45
## hsize3               0.15   -0.00    0.31     1.91   0.06
## hsize4               0.04   -0.12    0.21     0.53   0.60
## hsize5               0.08   -0.08    0.24     0.96   0.34
## bmi                  0.14    0.13    0.15    26.37   0.00
## ------------------------------------------------------------
```

```
summ(fit, confint=TRUE, exp=TRUE)
```

```
## MODEL INFO:
## Observations: 10351
## Dependent Variable: highbp
## Type: Generalized linear model
##   Family: binomial
##   Link function: logit
##
## MODEL FIT:
## <U+03C7>²(11) = 2445.87, p = 0.00
## Pseudo-R² (Cragg-Uhler) = 0.28
## Pseudo-R² (McFadden) = 0.17
## AIC = 11679.66, BIC = 11766.60
##
## Standard errors: MLE
## ------------------------------------------------------------
##                     exp(Est.)   2.5%   97.5%   z val.      p
## ----------------- ----------- ------ ------- -------- ------
## (Intercept)              0.00   0.00    0.00   -35.59   0.00
## female                   0.62   0.57    0.68   -10.58   0.00
```

```
## black                    1.48   1.27    1.71     5.17   0.00
## age                      1.05   1.04    1.05    29.63   0.00
## regionNE                 1.16   1.02    1.32     2.23   0.03
## regionS                  1.01   0.89    1.14     0.12   0.90
## regionW                  1.12   0.99    1.27     1.80   0.07
## hsize2                   1.05   0.92    1.20     0.76   0.45
## hsize3                   1.17   1.00    1.36     1.91   0.06
## hsize4                   1.05   0.88    1.24     0.53   0.60
## hsize5                   1.08   0.92    1.27     0.96   0.34
## bmi                      1.14   1.13    1.16    26.37   0.00
## ------------------------------------------------------------
```

Outputs type html

- Fonction `*export_summ*`
- Exportation des tableaux sous word ok

```
export_summs(fit, error_pos = "right")

## Registered S3 methods overwritten by 'broom':
##   method            from
##   tidy.glht         jtools
##   tidy.summary.glht jtools
```

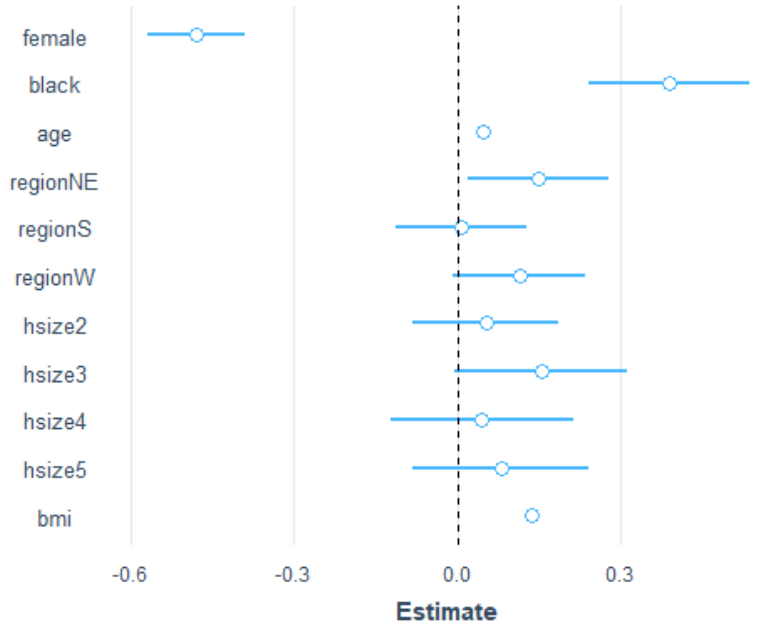
	Model 1	
(Intercept)	-6.00 ***	(0.17)
female	-0.48 ***	(0.05)
black	0.39 ***	(0.08)
age	0.05 ***	(0.00)
regionNE	0.15 *	(0.07)
regionS	0.01	(0.06)
regionW	0.11	(0.06)
hsize2	0.05	(0.07)
hsize3	0.15	(0.08)
hsize4	0.04	(0.09)
hsize5	0.08	(0.08)
bmi	0.14 ***	(0.01)
N	10351	
AIC	11679.66	
BIC	11766.60	
Pseudo R2	0.28	

*** p < 0.001; ** p < 0.01; * p < 0.05.

forest plots

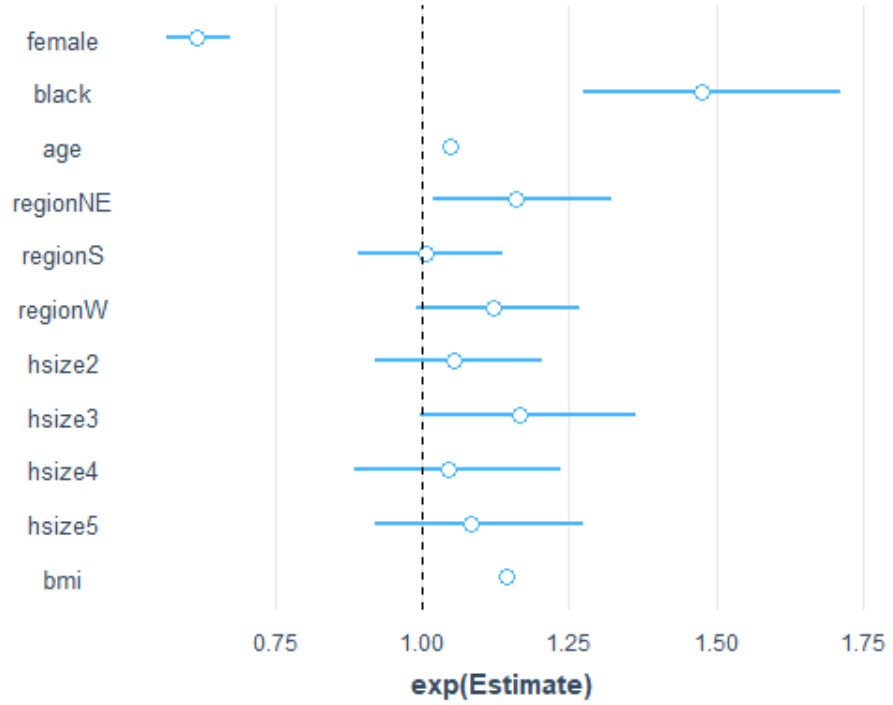
- Fonction `*plot_summ*`
- Détail: https://jtools.jacob-long.com/reference/plot_summs.html

```
plot_summs(fit)
```



Sous forme d'Odds ratio

```
plot_summs(fit,exp=TRUE)
```



Gtsummary

- Rendu du tableau supérieur à celui de jtools.
- Exportable dans word.
- Détail: https://www.danieldsjoberg.com/gtsummary/articles/tbl_regression.html

```
#install.packages("gtsummary")
library(gtsummary)
```

Avec report des Odds ratio

```
tbl_regression(fit, exponentiate = TRUE)
```

Characteristic	OR[1]	95% CI[1]	p-value
female	0.62	0.57, 0.68	<0.001
black	1.48	1.27, 1.71	<0.001
age	1.05	1.04, 1.05	<0.001
region			
MW	—	—	
NE	1.16	1.02, 1.32	0.026
S	1.01	0.89, 1.14	>0.9
W	1.12	0.99, 1.27	0.071
hsize			
1	—	—	
2	1.05	0.92, 1.20	0.4
3	1.17	1.00, 1.36	0.056
4	1.05	0.88, 1.24	0.6
5	1.08	0.92, 1.27	0.3
bmi	1.14	1.13, 1.16	<0.001

[1]OR = Odds Ratio, CI = Confidence Interval

Par défaut le nombre d'observations n'est pas reporté (bien dommage). Pour les ajouter on empile avec la fonction **`add_glance_table(include = nobs)`**:

```
tbl_regression(fit) %>% add_glance_table(include = nobs)
```

Characteristic	log(OR)[1]	95% CI[1]	p-value
female	-0.48	-0.57, -0.39	<0.001
black	0.39	0.24, 0.54	<0.001
age	0.05	0.04, 0.05	<0.001
region			
MW	—	—	
NE	0.15	0.02, 0.28	0.026
S	0.01	-0.11, 0.13	>0.9
W	0.11	-0.01, 0.24	0.071
hsize			
1	—	—	
2	0.05	-0.08, 0.19	0.4
3	0.15	0.00, 0.31	0.056
4	0.04	-0.12, 0.21	0.6
5	0.08	-0.08, 0.24	0.3
bmi	0.14	0.13, 0.15	<0.001
No. Obs.	10,351		

[1]OR = Odds Ratio, CI = Confidence Interval

Qualité du modèle

Test du rapport des vraisemblances

On va comparer la vraisemblance du modèle estimé jusqu'à présent, avec celle ajoutant la variable rural.

- On installe et charge la librairie **`lrmest`**
- On estime le modèle en ajoutant la variable *rural* (modèle *fit2*)
- On utilise la fonction **`lrtest`**: comme on emboîte, penser à mettre le modèle teste (degrés de liberté le plus élevé en premier)

```
#install.packages("lmtest")
library(lmtest)

fit2 = glm(highbp ~ female + black + age + region + hsize + bmi + rural, family=binomial, data=df)

lrtest(fit2, fit)
```

```
## Likelihood ratio test
##
## Model 1: highbp ~ female + black + age + region + hsize + bmi + rural
## Model 2: highbp ~ female + black + age + region + hsize + bmi
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  13 -5826.9
## 2  12 -5827.8 -1 1.8373     0.1753
```

AUC (courbe de ROC)

- La courbe de ROC (sensibilité sur les ordonnées et spécificité sur les abscisses + report de l'aire sous la courbe) est disponible dans plusieurs librairies. On utilisera **pROC** - Attention, même si cela ne change rien, la fonction reporte la spécificité en abscisse inversée (de 1 à 0). D'autres logiciels reportent sur cet axe (1-spécificité) sur une échelle croissante. Le résultat est en tout point identique. - Avant de tracer la courbe et récupéré la valeur de l'aire, il faut générer les probabilités prédites par le modèle dans un objet.

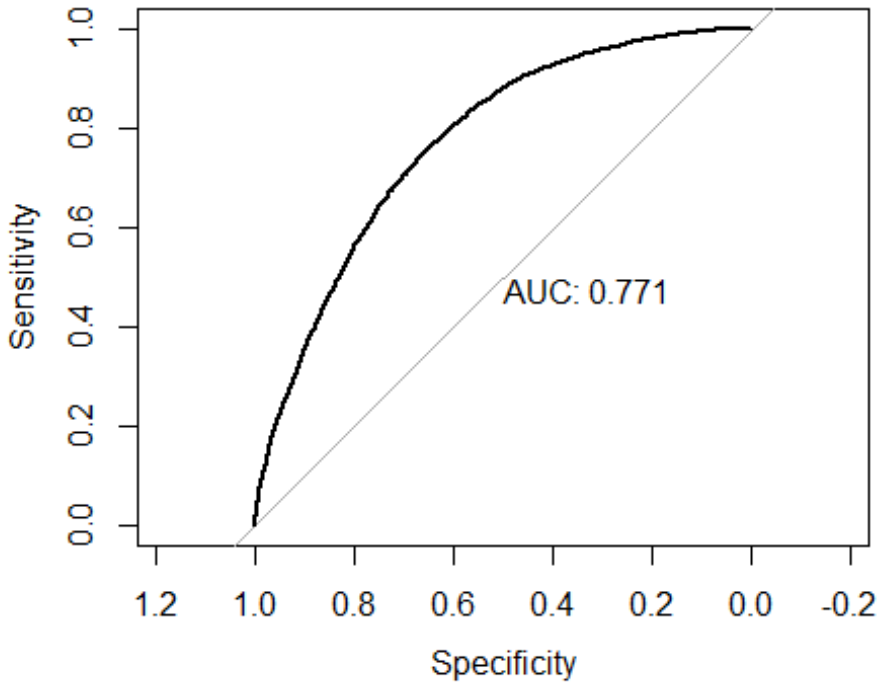
```
#install.packages("pROC")
library(pROC)
```

Prédiction dans l'objet roc avec la function **predict**

```
roc = predict(fit, newdata = df, type = "response")
```

Courbe et valeur de l'AUC (0.77…c'est pas mal)

```
test_roc = roc(df$highbp ~ roc, plot = TRUE, print.auc = TRUE)
```



Introduction de pondérations avec estimation robuste

- Cas d'un échantillonage aléatoire simple (attention ce n'est pas le cas de l'exemple, mais on a fait comme si)
- On introduit des poids de sondage avec estimation robuste de la variance (permet de neutraliser l'hétéroscédasticité générée mécaniquement)
- Librairie **`survey`**
- Warning: toujours utiliser des pondérations normalisées. Si les poids disponibles sont bruts (somme des poids = taille de la population dans laquelle l'échantillon a été tiré) on divise les poids par la moyennes des poids bruts. La somme des poids normalisés est alors égale à la taille de l'échantillon).

```
#install.packages("survey")
library(survey)
```

Avec un plan de sondage simple, le paramétrage des infos relatives au plan de sondage est limité à l'inscription de la variable de pondération, ici *w*. On utilise la fonction **`svydesign()`**, dont les informations contenues dans l'objet seront utilisées pour lors de l'estimation.

- Le modèle est estimé avec la fonction **`svyglm`** avec une précaution dans la syntaxe.
- L'introduction de pondération conduisant à multiplier les valeurs des variables utilisées dans le modèle par un nombre, l'indicatrice de la variable analysée n'est donc plus estimée avec des valeurs égales à 0 ou 1, mais avec des valeurs égales à 0 si y=0 et w si y=1.
- Dans le cas d'une régression logistique on parle d'une estimation quasi-binomiale (car y est toujours =0). On l'a souligné rapidement c'est ce type d'estimation, plus générale que dans la situation binaire, qui permet d'utiliser le modèle logistique pour estimer directement des proportions comme variable dépendante. On retrouve également ce genre d'estimation avec un modèle de poisson lorsque les valeurs de la variable dépendante ne sont pas entières.

On utilise la fonction **`svyglm()`**, on modifie juste le type d'estimation avec ***`family=quasi-binomial`*** et on ajoute l'objet *pw* généré avec la fonction `svydesign`.

```
fit = svyglm(highbp ~ female + black + age + region + factor(hsize) + bmi,
pw, family=quasibinomial, data=df)
summary(fit)

## Call:
## svyglm(formula = highbp ~ female + black + age + region + factor(hsize)
+
##     bmi, design = pw, family = quasibinomial, data = df)
##
## Survey design:
## svydesign(ids = ~1, data = df, weights = ~w)
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -6.174463   0.199390 -30.967  < 2e-16 ***
## female         -0.610840   0.054254 -11.259  < 2e-16 ***
## black           0.373650   0.093827   3.982 6.87e-05 ***
## age             0.048663   0.001847  26.350  < 2e-16 ***
## regionNE        0.175838   0.075850   2.318   0.0205 *
## regionS        -0.007909   0.073587  -0.107   0.9144
## regionW         0.150577   0.076282   1.974   0.0484 *
## factor(hsize)2  0.080490   0.084451   0.953   0.3406
## factor(hsize)3  0.188753   0.095480   1.977   0.0481 *
## factor(hsize)4  0.090865   0.099706   0.911   0.3621
## factor(hsize)5  0.086419   0.096324   0.897   0.3697
## bmi             0.140629   0.006237  22.548  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasibinomial family taken to be 0.9829967)
##
## Number of Fisher Scoring iterations: 4
```

Remarque : le paramètre de dispersion (ou d'échelle), contraint à 1 dans le modèle sans pondération, est égal à 0.98. Il est forcément proche de 1 car la moyenne des pondérations normalisées est égale à 1 (sa valeur sera plus faible si on estime des proportions).

STATA

Tout est dans l'aide ou presque, il n'y a aucune difficulté pour estimer un modèle logistique avec Stata ou pour effectuer des diagnostiques et standards.

- Fonction **`logit`** ou **`logistic`**, identiques à ceci près que la seconde affiche les Odds Ratio par défaut
- On peut estimer le modèle avec **`glm`**. Pas d'utilité particulière avec les commandes précédentes
- Pour les effets marginaux on utilise la très populaire commande **`margins`**
- Pour la qualité de l'ajustement ou le test du rapport de vraisemblance, Stata dispose d'une série de commandes très simples. On peut conseiller néanmoins d'installer et utiliser la librairie **`spost13`** qui dispose d'une série de fonctions très utiles, en particulier **`fitstat`**.

Estimation du modèle avec logit

Par rapport au déroulé de la formation on part du modèle final avec les variables *female*, *black*, *age*, *region*, *hsize* et *bmi*.

Ouverture de la base:

```
use "hypertension.dta", clear
```

Format et entrées des variables :

- Variables dépendante: elle doit avoir les valeurs $\{0,1\}$ (cela permet de ne pas oublier sur quelles valeurs on estime). Elle peut avoir des labels affectés aux deux modalités.
- Covariables : elles doivent être en format numérique
  - Les variables quantitatives sont entrées telles quelles si elles ne sont pas croisées
  - Idem pour les indicatrices comme *female* et *black*. Elles peuvent avoir des labels affectés aux modalités.
  - Si la variable catégorielle n'est pas sous forme d'indicatrice, comme *region* ou *hsize*, on doit les indicer : **`i.nom_variable`** ou **`ib#.nom_variable`** avec # le niveau de réféfrence sélectionné.

On ne déclinera pas toutes les options d'affichage, mais on ajoutera l'option **`baselevel`** qui ajoute la modalité en référence pour les variables catégorielles.

Estimation du modèle

Paramètres estimés: logit et contraste logistique

```
logit highbp female black age i.region i.hsize bmi, baselevel


Logistic regression                             Number of obs     =     10,351
                                                LR chi2(11)       =    2445.87
                                                Prob > chi2       =     0.0000
Log likelihood = -5827.8325                     Pseudo R2         =     0.1734

------------------------------------------------------------------------------
      highbp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |     -0.481      0.045   -10.58   0.000       -0.570      -0.392
       black |      0.390      0.075     5.17   0.000        0.242       0.537
         age |      0.047      0.002    29.63   0.000        0.044       0.050
             |
      region |
          NE |      0.000      (base)
          MW |     -0.148      0.067    -2.23   0.026       -0.279      -0.018
           S |     -0.141      0.066    -2.12   0.034       -0.271      -0.011
           W |     -0.035      0.067    -0.52   0.601       -0.166       0.096
             |
       hsize |
          1  |      0.000      (base)
          2  |      0.052      0.068     0.76   0.446       -0.081       0.185
          3  |      0.153      0.080     1.91   0.056       -0.004       0.311
          4  |      0.045      0.085     0.53   0.599       -0.122       0.212
          5  |      0.080      0.083     0.96   0.336       -0.083       0.242
             |
         bmi |      0.135      0.005    26.37   0.000        0.125       0.145
       _cons |     -5.848      0.171   -34.23   0.000       -6.183      -5.514
------------------------------------------------------------------------------
```

L'output est classique et plutôt *clean*. En tête figure le nombre d'observation utilisée, le test du rapport des vraisemblances avec le modèle sans covariable, et le pseudoR2 standard (voir plus bas).

Si on veut lire les paramètres sous forme d'OR, on ajoute simplement l'option **or**.

```
logit highbp female black age i.region i.hsize bmi, baselevel or


Logistic regression                               Number of obs     =     10,351
                                                  LR chi2(11)       =    2445.87
                                                  Prob > chi2       =     0.0000
Log likelihood = -5827.8325                       Pseudo R2         =     0.1734

------------------------------------------------------------------------------
      highbp | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |      0.618      0.028   -10.58   0.000        0.565       0.676
       black |      1.476      0.111     5.17   0.000        1.274       1.712
         age |      1.048      0.002    29.63   0.000        1.045       1.051
             |
      region |
          NE |      1.000  (base)
          MW |      0.862      0.057    -2.23   0.026        0.757       0.982
           S |      0.869      0.058    -2.12   0.034        0.763       0.989
           W |      0.966      0.065    -0.52   0.601        0.847       1.101
             |
       hsize |
           1 |      1.000  (base)
           2 |      1.053      0.072     0.76   0.446        0.922       1.203
           3 |      1.166      0.094     1.91   0.056        0.996       1.365
           4 |      1.046      0.089     0.53   0.599        0.885       1.237
           5 |      1.083      0.090     0.96   0.336        0.921       1.274
             |
         bmi |      1.145      0.006    26.37   0.000        1.133       1.156
       _cons |      0.003      0.000   -34.23   0.000        0.002       0.004
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

Warning: la colonne affiche Odds Ratio alors que le terme constant estime un Odds. C'est précisé dans une petite note, qui a été ajouté suite à de nombreux commentaires des utilisateurs. Néanmoins cette note n'a pas adaptée à la présence de termes d'interaction, qui ne sont pas non plus des OR mais des rapports d'OR. On est pas ici au niveau de l'output de SAS qui n'affiche que les Odds Ratio dans le tableau **Odds Ratio**.

Changement de référence

- Stata prend la valeur la plus faible pour fixer la référence.
- On peut simplement changer le niveau en changeant la valeur de l'indice **ib#**. Pour la variable région, si on veut faire passer la référence à *MW* (valeur=2 de la variable): **ib2.region**.

```
logit highbp female black age ib2.region i.hsize bmi

Logistic regression                             Number of obs     =     10,351
                                                LR chi2(11)       =    2445.87
                                                Prob > chi2       =     0.0000
Log likelihood = -5827.8325                     Pseudo R2         =     0.1734

------------------------------------------------------------------------------
      highbp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |     -0.481      0.045   -10.58   0.000       -0.570      -0.392
       black |      0.390      0.075     5.17   0.000        0.242       0.537
         age |      0.047      0.002    29.63   0.000        0.044       0.050
             |
      region |
          NE |      0.148      0.067     2.23   0.026        0.018       0.279
           S |      0.008      0.062     0.12   0.902       -0.113       0.128
           W |      0.113      0.063     1.80   0.071       -0.010       0.236
             |
       hsize |
           2 |      0.052      0.068     0.76   0.446       -0.081       0.185
           3 |      0.153      0.080     1.91   0.056       -0.004       0.311
           4 |      0.045      0.085     0.53   0.599       -0.122       0.212
           5 |      0.080      0.083     0.96   0.336       -0.083       0.242
             |
         bmi |      0.135      0.005    26.37   0.000        0.125       0.145
       _cons |     -5.997      0.169   -35.59   0.000       -6.327      -5.666
------------------------------------------------------------------------------
```

Intéractions

Les interactions entre deux variables ou plus sont posées avec **#** ou **##**.

Entre deux variables binaires

On va croiser les variable *female* et *black*.

- Solution 1 : **`female#black`**. Permet de croiser directement les variables sans ajouter un terme d'interaction. Stata génère les 4 indicatrices correspondant aux états croisée. En référence on trouvera la situation pour female=0 et black=0.

```
logit highbp female#black age ib2.region i.hsize bmi


Logistic regression                             Number of obs     =     10,351
                                                LR chi2(12)       =    2445.89
                                                Prob > chi2       =     0.0000
Log likelihood =  -5827.822                     Pseudo R2         =     0.1734

------------------------------------------------------------------------------
      highbp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
female#black |
        0 1  |      0.401      0.108     3.72   0.000        0.190       0.612
        1 0  |     -0.479      0.048    -9.97   0.000       -0.573      -0.385
        1 1  |     -0.099      0.104    -0.96   0.337       -0.303       0.104
             |
         age |      0.047      0.002    29.62   0.000        0.044       0.050
             |
      region |
          NE |      0.148      0.067     2.23   0.026        0.018       0.279
           S |      0.008      0.062     0.12   0.902       -0.113       0.128
           W |      0.113      0.063     1.80   0.071       -0.010       0.236
             |
       hsize |
           2 |      0.052      0.068     0.77   0.443       -0.081       0.186
           3 |      0.154      0.080     1.91   0.056       -0.004       0.311
           4 |      0.045      0.085     0.53   0.596       -0.122       0.213
           5 |      0.080      0.083     0.96   0.335       -0.083       0.243
             |
         bmi |      0.135      0.005    26.27   0.000        0.125       0.145
       _cons |     -6.000      0.170   -35.34   0.000       -6.332      -5.667
------------------------------------------------------------------------------
```

Solution 2 : **`female##black`** on entre directement un terme d'interaction (différence entre deux contrastes logistiques ou rapport d'Odds Ratio).

```
logit highbp female##black age ib2.region i.hsize bmi

------------------------------------------------------------------------------
      highbp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.female |     -0.479      0.048    -9.97   0.000       -0.573      -0.385
     1.black |      0.401      0.108     3.72   0.000        0.190       0.612
             |
female#black |
        1 1  |     -0.021      0.148    -0.14   0.885       -0.311       0.268
             |
         age |      0.047      0.002    29.62   0.000        0.044       0.050
             |
      region |
          NE |      0.148      0.067     2.23   0.026        0.018       0.279
           S |      0.008      0.062     0.12   0.902       -0.113       0.128
           W |      0.113      0.063     1.80   0.071       -0.010       0.236
             |
       hsize |
           2 |      0.052      0.068     0.77   0.443       -0.081       0.186
           3 |      0.154      0.080     1.91   0.056       -0.004       0.311
           4 |      0.045      0.085     0.53   0.596       -0.122       0.213
           5 |      0.080      0.083     0.96   0.335       -0.083       0.243
             |
         bmi |      0.135      0.005    26.27   0.000        0.125       0.145
       _cons |     -6.000      0.170   -35.34   0.000       -6.332      -5.667
------------------------------------------------------------------------------
```

Effets marginaux (AME)

- Commande **`margins`** (`help margins`)
- Pour les AME, on ajoute en option **`dydx(liste_variable)`**. Si on veut estimer les AME pour toutes les variables introduites dans le modèle : **`dydx(*)`**.

```
margins,  dydx(*)


Average marginal effects                        Number of obs     =     10,351
Model VCE    : OIM

Expression   : Pr(highbp), predict()
dy/dx w.r.t. : female black age 2.region 3.region 4.region 2.hsize 3.hsize 4.hsize
5.hsize bmi

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |     -0.092      0.009   -10.79   0.000       -0.109      -0.075
       black |      0.074      0.014     5.19   0.000        0.046       0.103
         age |      0.009      0.000    35.29   0.000        0.008       0.009
             |
      region |
          MW |     -0.028      0.013    -2.23   0.026       -0.053      -0.003
           S |     -0.027      0.013    -2.12   0.034       -0.052      -0.002
           W |     -0.007      0.013    -0.52   0.601       -0.032       0.019
             |
       hsize |
           2 |      0.010      0.013     0.76   0.446       -0.015       0.035
           3 |      0.029      0.015     1.91   0.056       -0.001       0.059
           4 |      0.009      0.016     0.53   0.599       -0.023       0.040
           5 |      0.015      0.016     0.96   0.335       -0.016       0.046
             |
         bmi |      0.026      0.001    30.09   0.000        0.024       0.027
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

Analyse de la qualité du modèle

Installation de la librairie spost13 (pour la commande **`fitstat`**). La suite de commandes n'est pas stockée dans l'entrepôt Stata: rechercher le dépôt avec `search spost13` => aller dans la page **spost13_ado from** **https://jslsoc.sitehost.iu.edu/stata** et installer les commandes.

Commande fitstat

- Après avoir estimé un modèle, la commande **fitstat** permet de récupéré un ensemble de d'indicateurs de « performance »: déviance, vraisemblance pénalisée, liste de pseudo R2. Par exemple :

```
   fitstat

                      |       logit
----------------------+-------------
Log-likelihood        |
                Model |   -6230.366
        Intercept-only |   -7050.765
----------------------+-------------
Chi-square            |
      Deviance(df=10340) |   12460.731
              LR(df=10) |    1640.800
                p-value |       0.000
----------------------+-------------
R2                    |
               McFadden |       0.116
      McFadden(adjusted) |       0.115
      McKelvey & Zavoina |       0.191
            Cox-Snell/ML |       0.147
  Cragg-Uhler/Nagelkerke |       0.197
                  Efron |       0.148
               Tjur's D |       0.149
                  Count |       0.668
         Count(adjusted) |       0.216
----------------------+-------------
IC                    |
                   AIC |   12482.731
        AIC divided by N |       1.206
             BIC(df=11) |   12562.424
----------------------+-------------
Variance of           |
                     e |       3.290
                 y-star |       4.067
```

Remarque : On peut également afficher les vraisemblances pénalisées avec la commande **`estat ic`**.

- Il permet d'effectuer un test de rapport des vraisemblances entre deux modèles imbriqués (voir ci-dessous).

Test du rapport des vraisemblances

- Le test avec le modèle sans covariable est reporté dans l'output par défaut, il n'y a rien à ajouter.

- Si on veut comparer deux modèles imbriqués, on peut utiliser la commande **lrtest** après avoir estimé deux modèles et sauvegardé leur résultats; ou utiliser la commande fitstat.
  Exemple: premier modèle sans la variable *bmi*, le second avec la variable.

- Avec **lrtest**:
  - On estime le modèle 1
  - On enregistre les résultats du modèle avec **`eststo nom_modèle1`**
  - On répète l'opération pour le second modèle
  - On exécute **`lrtest nom_modèle2 nom_modèle1`**
  - Remarque: on fera en sorte que le second modèle soit celui qui comporte le plus de degré de liberté (ajout d'une ou plus variables)

```
qui logit highbp female black age i.region i.hsize, baselevel

eststo fit1

qui logit highbp female black age i.region i.hsize bmi, baselevel

eststo fit2

lrtest fit2 fit1

Likelihood-ratio test                                 LR chi2(1)  =    805.07
(Assumption: fit1 nested in fit2)                     Prob > chi2 =    0.0000
```

- Avec **fitstat**:
  - On estime le premier modèle
  - On exécute `fitstat` avec l'option **`save`**
  - On estime le second modèle
  - On exécute `fitstat` avec l'option **`diff`**
  - Remarque : permet d'avoir une comparaison pour les autres indicateurs reportés par la commande.

```
qui logit highbp female black age i.region i.hsize, baselevel
qui fitstat, save
qui logit highbp female black age i.region i.hsize bmi, baselevel

fitstat, diff

                         |     Current        Saved   Difference
-------------------------+-----------------------------------------
Log-likelihood           |
                   Model |   -5827.832    -6230.366      402.533
           Intercept-only |   -7050.765    -7050.765        0.000
-------------------------+-----------------------------------------
Chi-square               |
    D(df=10339/10340/-1) |   11655.665    12460.731     -805.066
          LR(df=11/10/1) |    2445.866     1640.800      805.066
                 p-value |       0.000        0.000        0.000
-------------------------+-----------------------------------------
R2                       |
                McFadden |       0.173        0.116        0.057
      McFadden(adjusted) |       0.172        0.115        0.057
      McKelvey & Zavoina |       0.290        0.191        0.099
```

```
            Cox-Snell/ML |       0.210        0.147        0.064
  Cragg-Uhler/Nagelkerke |       0.283        0.197        0.086
                   Efron |       0.215        0.148        0.067
                Tjur's D |       0.216        0.149        0.067
                   Count |       0.702        0.668        0.034
        Count(adjusted) |       0.296        0.216        0.080
-------------------------+------------------------------------------
IC                       |
                     AIC |   11679.665    12482.731     -803.066
        AIC divided by N |       1.128        1.206       -0.078
         BIC(df=12/11/1) |   11766.603    12562.424     -795.821
-------------------------+------------------------------------------
Variance of              |
                       e |       3.290        3.290        0.000
                  y-star |       4.635        4.067        0.568

Note: Likelihood-ratio test assumes saved model nested in current model.

Difference of  795.821 in BIC provides very strong support for current model.
```

Courbe de ROC et AUC

- La courbe et l'aire sous la courbe (ou plutôt entre la courbe et la diagonale) est simplement obtenue avec la commande **`lroc`**. Elle affiche par défaut la courbe, on peut supprimer le graphique avec l'option `nograph`.
- Si on souhaite afficher une matrice de confusion à un seuil donnée (par défaut 0.5), on peut utiliser la commande **`estat classification`**.

```
qui logit highbp female black age i.region i.hsize bmi, baselevel

lroc

Logistic model for highbp

number of observations =    10351
area under ROC curve   =   0.7709
```

Introduction de poids de sondage (estimation robuste)

De nouveau très simple. Pour avoir une estimation robuste de la variance, on déclare la variable de pondération (vérifier qu'elle est bien normée) avec **`pw`**. On peut également utiliser le mode **survey** en déclarant de nouveau la pondération en amont avec **`svyset`**.

```
logit highbp female black age i.region i.hsize bmi [pw=w], baselevel

Logistic regression                             Number of obs     =     10,351
                                                Wald chi2(11)     =    1474.14
                                                Prob > chi2       =     0.0000
```

```
Log pseudolikelihood = -5625.6375               Pseudo R2         =     0.1743

------------------------------------------------------------------------------
             |               Robust
      highbp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |     -0.611      0.054   -11.26   0.000       -0.717      -0.505
       black |      0.374      0.094     3.98   0.000        0.190       0.558
         age |      0.049      0.002    26.35   0.000        0.045       0.052
             |
      region |
          NE |      0.000  (base)
          MW |     -0.176      0.076    -2.32   0.020       -0.325      -0.027
           S |     -0.184      0.078    -2.35   0.019       -0.337      -0.031
           W |     -0.025      0.080    -0.32   0.751       -0.181       0.131
             |
       hsize |
           1 |      0.000  (base)
           2 |      0.080      0.084     0.95   0.341       -0.085       0.246
           3 |      0.189      0.095     1.98   0.048        0.002       0.376
           4 |      0.091      0.100     0.91   0.362       -0.105       0.286
           5 |      0.086      0.096     0.90   0.370       -0.102       0.275
             |
         bmi |      0.141      0.006    22.55   0.000        0.128       0.153
       _cons |     -5.999      0.201   -29.82   0.000       -6.393      -5.604
------------------------------------------------------------------------------
```

Remarque : c'est équivalent à l'utilisation d'un poids non robuste avec **`iw`** (le poids entre simplement en coefficient multiplicateur dans l'estimation de la variance), à laquelle on ajoute une correction de l'hétéroscédasticité en option avec **`vce(robust)`**.

```
logit highbp female black age i.region i.hsize bmi [iw=w], baselevel vce(ro
bust)
```

Enfin en mode *survey*

```
svyset [pw = w]
svy: logit highbp female black age i.region i.hsize bmi
```