Introduction

Damien Cartron, Martin Chevalier, Michel Forsé
et Florence Maillochon

Année universitaire 2017-2018

Objectifs et contenu du cours
Ce que NE fait PAS ce cours

Apprendre à utiliser des **logiciels** pour faire du traitement statistique de données :

- *Introduction au logiciel de statistique SAS* de Y Ando (18h)
- *Introduction aux méthodes quantitatives pour sociologues avec le logiciel R* de R. Yanez (24h)
- Tutoriel R en ligne sur `r.slmc.fr`

Aborder des méthodes économétriques avancées : *Économétrie pour sociologues* de O. Godechot (12h, S1 2018-2019).

Organisation générale
Modalités de validation

30h de cours – 6 ECTS.

Mini-mémoire à rendre en juin 2018 (en papier dans le casier de Michel Forsé ou par mail à michel.forse@ens.fr) :

- entre une dizaine et une vingtaine de pages ;
- avec une introduction posant la problématique sociologique et les hypothèses à vérifier ;
- une analyse de données empirique ;
- une conclusion.

Aucun développement n'est attendu sur le côté « programme informatique » mis en œuvre, le choix du logiciel étant libre (**R**, SAS, Stata, SPSS, SPAD, etc.).

Objectifs et contenu du cours
Ce que fait ce cours

Introduire les principales **méthodes d'analyse statistique multivariée** :

- Méthodes de régression : régressions linéaires et régressions logistiques.
- Méthodes d'analyse de données : ACP, AFC, ACM, CAH.
- Éléments sur le traitement de données longitudinales et l'analyse statistique de réseaux.

Développer un **regard réflexif** sur ces méthodes et les données statistiques utilisées dans le cadre d'un travail empirique.

Faciliter l'appréhension et développer le recul critique sur les **articles de recherches** mobilisant ces méthodes.

Organisation générale
Dates et emplacement des cours

Tous les cours ont lieu au 48 bd Jourdan :

Méthodes de régression (M. Chevalier) : 8 février (14h-17h, R2-02), 15 février (14h-17h, R3-35), 8 mars (13h-16h, R2-02) et 15 mars (13h-16h, R2-02)

Analyse des réseaux (M. Forsé) : 22 mars (13h-16h, R2-02)

Analyses factorielles, classifications et typologies (D. Cartron) : 29 mars (10h-13h, R1-14, et 13h-16h (?), R2-02) et 30 mars (10h-13h, R1-14, et 13h-16h (?), R2-02)

Analyse longitudinale, notions avancées (F. Maillochon) : 5 avril (13h-16h, R2-02)

Organisation générale
Organisation des quatre premières séances

Séance du 8 février : Introduction aux méthodes de régression

Séance du 15 février : Les méthodes de régression linéaire

Séance du 8 mars : Les méthodes de régression sur variable qualitative

Séance du 15 mars : Problématiques communes aux différentes méthodes et compléments

`http://teaching.slmc.fr/mqs2`

Introduction aux méthodes de régression

Martin Chevalier (Insee)

Année universitaire 2017-2018

Plan de la séance

Objectifs de la séance

Présenter à travers un exemple la **démarche générale des méthodes de régression** :

1. Construire un modèle pertinent ;
2. Interpréter les relations entre variables ;
3. Tenter de dégager des relations causales.

Introduire certains points importants qui seront davantage développés par la suite :

- les différentes spécifications de modèle ;
- la mise en évidence d'« effets purs » ;
- la question de la causalité.

Interpréter les relations entre variables

À la différence des outils de la statistique uni- et bi-variée, les méthodes de régression permettent d'analyser les **relations entre de très nombreuses variables**.

Elles décrivent les relations entre :

- les variations d'une variable dite « **expliquée** » (« dépendante »)
- et celles de variables « **explicatives** » (« indépendantes », « covariables », « régresseurs »).

L'objectif de la régression est de **mesurer la force du lien entre la variable expliquée et chacune des variables explicatives**, et de déterminer s'il est ou non **statistiquement significatif**.

Quelques références bibliographiques

Béhagel L. (2006), *Lire l'économétrie*, coll. Repères, La Découverte, 128 p.

Dormont B. (2007) *Introduction à l'économétrie*, Montchrestien, 450 p.

Wooldridge J. (2008), *Introductory econometrics. A modern approach*, 839 p.

Cameron A., Trivedi P. (2005), *Microeconometrics : Methods and applications*, Cambridge University Press, New York, 1 056 p.

Construire un modèle pertinent

Mener une régression, c'est chercher à appliquer un **modèle économétrique** aux données que l'on souhaite exploiter.

- Les méthodes de régression conduisent nécessairement à une **simplification**...
- ...qui est acceptée pour autant qu'elle parvienne à **préserver les phénomènes les plus structurants**.

La construction d'un modèle qui soit **pertinent au regard des données et des hypothèses de travail** est donc une étape essentielle dans la mise en œuvre d'une méthode de régression.

Avant d'aboutir au modèle définitif, il faut en estimer un **très grand nombre** et les évaluer à l'aide d'**outils de diagnostics** plus ou moins complexes.

Tenter de dégager des relations causales

« Corrélation n'est pas causalité » : le lien statistiquement significatif entre deux variables n'indique **pas nécessairement une relation de causalité** ; il conduit même parfois à des **conclusions fallacieuses**.

En règle générale, **les modèles de régression seuls ne suffisent pas établir de lien de causalité** : il faut étayer l'argumentaire sur d'autres éléments, notamment issus de la **littérature**.

Dans certaines configurations néanmoins, le modèle économétrique peut permettre de dégager des relations de cause à effet.

Présentation des données

On génère un **échantillon fictif de 100 personnes** pour lesquelles sont connus le poids, la taille ainsi que la pratique d'une activité physique régulière.

L'objectif de l'étude est de **mesurer l'impact de l'activité physique sur le poids des individus**.

1. On cherche tout d'abord à vérifier que poids et taille sont statistiquement liés : **régression linéaire simple**.
2. On introduit ensuite la pratique d'une activité physique régulière : **régression linéaire multiple**.

Le poids est la **variable expliquée**, la taille et l'activité physique sont les variables explicatives.

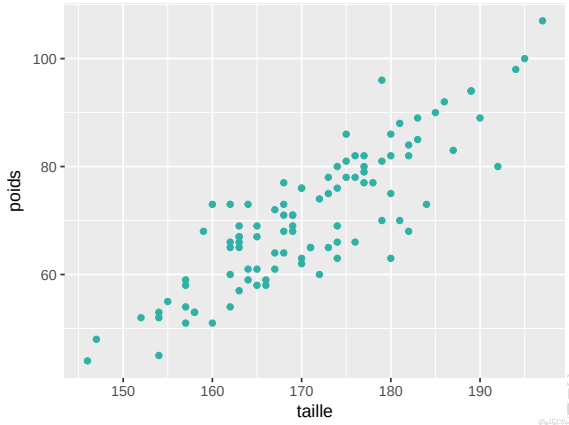
Exemple : poids, taille et activité physique
Simulation des données

```r
# Initialisation du générateur de nombres pseudo-aléatoires de R
set.seed(1)

# Taille de l'échantillon simulé
n <- 100

# Simulation des variables
femme <- rbinom(n, 1, 0.5)
sport <- femme * rbinom(n, 1, 0.65) + (1 - femme) * rbinom(n, 1, 0.45)
taille <- (1 - femme)*rnorm(n, 177, 8) + femme*rnorm(n, 165, 8)
imc <- sport * rnorm(n, 22, 1) + (1 - sport) * rnorm(n, 26, 1)
poids <- imc * (taille / 100) ** 2
regime <- (imc > 23 + rnorm(n, 0, 2.5)) * 1
l.poids <- regime * (poids + rnorm(n, 2, 2)) +
  (1 - regime) * (poids + rnorm(n, 0, 2))

# Arrondi pour les variables numériques
taille <- round(taille, 0)
poids <- round(poids, 0)
l.poids <- round(l.poids, 0)

# Constitution de la table de données d
d <- data.frame(poids,taille,sport)
```


Exemple : poids, taille et activité physique
Les dix premières observations

poids	taille	sport
88	181	0
79	177	0
61	164	1
65	163	0
67	165	0
69	174	1
44	146	1
63	170	1
71	168	0
80	192	1

Poids et taille : régression linéaire simple
Représentation du nuage de points



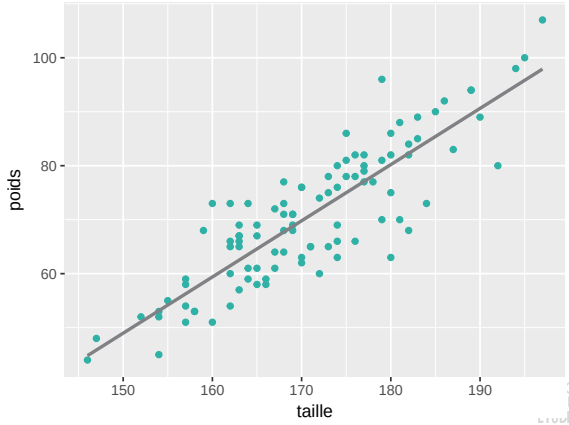
Poids et taille : régression linéaire simple
Corrélation entre les deux variables

On a affaire à deux variables de nature quantitatives donc on peut commencer par calculer le coefficient de corrélation linéaire : **0,8646**.

Le coefficient de corrélation linéaire et le graphique semblent indiquer une **liaison linéaire** entre ces deux variables.

Intuitivement, cela signifie que **le nuage de points du graphique pourrait être bien résumé par une droite** d'équation $y = ax + b$ (avec a et b à estimer).

Poids et taille : régression linéaire simple
Relation linéaire entre taille et poids



Poids et taille : régression linéaire simple
Modélisation

On modélise donc la relation entre le poids et la taille par une équation de la forme :

$$poids_i = \beta_0 + \beta_1 \times taille_i + \varepsilon_i$$

- β_0 et β_1 sont des **paramètres** : inconnus au départ, ils sont estimés à partir des données ;
- β_0 est appelé la **constante** (ou *intercept* en anglais) : il existe des modèles avec ou sans constante ;
- β_1 est le paramètre associé à la variable *taille* : sa valeur s'interprète comme le **nombre moyen** de kg par centimètre supplémentaire dans l'échantillon ;
- ε est le **résidu** : il mesure l'écart du modèle aux données réelles. Si le modèle était parfait, ε vaudrait 0 pour tous les individus.

Poids et taille : régression linéaire simple
Modélisation

Estimer un modèle revient à **déterminer la valeur des paramètres** (ici β_0 et β_1) qui **maximise l'ajustement du modèle aux données**.

Cela revient à chercher $\hat{\beta}_0$ et $\hat{\beta}_1$ tels que, **à partir de la seule taille d'un individu**, on soit en mesure de **déterminer son poids en se trompant en moyenne le moins possible**.

En général, cet objectif revient à **minimiser la somme des carrés des résidus** ε_i : $\sum_{i=1}^{n} \varepsilon_i^2$. Ce n'est cependant pas le seul objectif possible.

Poids et taille : régression linéaire simple
Estimation et interprétation des coefficients

En estimant ce modèle sur les données de l'exemple, on aboutit aux valeurs suivantes :

$$\hat{\beta}_0 = -107,17 \quad \text{et} \quad \hat{\beta}_1 = 1,04$$

Concrètement, cela signifie qu'**en moyenne dans l'échantillon, une taille d'un centimètre supplémentaire est associée à une masse d'environ un kilogramme supplémentaire**.

Le fait que le coefficient $\hat{\beta}_0$ soit négatif peut étonner : personne n'a de poids négatif !

Mais il s'agit de la **constante**, c'est-à-dire du **poids qu'aurait un individu de taille nulle** (ce qui est absurde).

Poids et taille : régression linéaire simple
Valeurs prédites par le modèle

Plus généralement, il est possible de **calculer les poids « prédits » par le modèle** à partir de la taille des individus.

Il s'obtient en **appliquant la formule du modèle pour chaque individu** :

$$poids_i^{M1} = -107,17 + 1,04 \times taille_i$$

Pour chaque individu, la **valeur du résidu** correspond à la **différence entre le poids prédit par le modèle et le poids réel** :

$$\varepsilon_i^{M1} = poids_i - poids_i^{M1}$$

Poids et taille : régression linéaire simple
Valeurs prédites par le modèle

poids	taille	sport	poids_m1
88	181	0	81,24
79	177	0	77,08
61	164	1	63,55
65	163	0	62,51
67	165	0	64,59
69	174	1	73,96
44	146	1	44,81
63	170	1	69,79
71	168	0	67,71
80	192	1	92,7

Activité physique : régression linéaire multiple
Intuition et motivation

On cherche désormais à introduire la pratique d'une activité physique régulière dans le cadre de ce modèle.

À taille égale, on s'attend à ce qu'une personne qui pratique une activité physique régulière ait un poids inférieur.

Pour confirmer cette intuition, on calcule des **statistiques descriptives** selon la pratique d'une activité physique ou non.

Activité physique	Taille	Poids
Oui	166,76 cm	60,84 kg
Non	174,24 cm	78,76 kg

Activité physique : régression linéaire multiple
Intuition et motivation

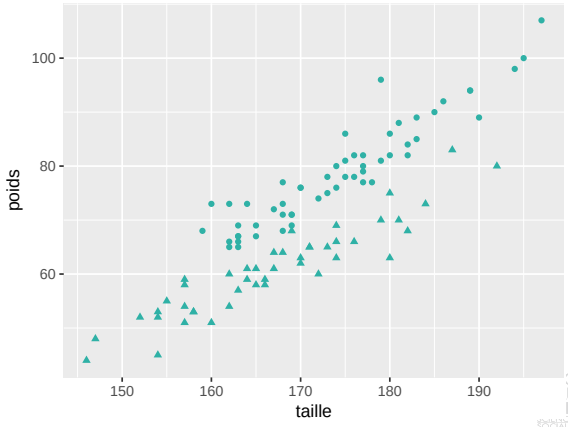
Ces statistiques ne permettent pas de conclure :

- D'une part, les personnes qui pratiquent une activité physique régulière sont **en moyenne plus légères** ;
- Mais d'autre part, elles sont également **en moyenne plus petites**.

On met ici en évidence un **effet de structure** : comme les personnes qui font du sport sont en moyenne à la fois plus légères et plus petites, il est difficile de déterminer si, à taille donnée, l'activité physique est associée à un poids plus faible.

On utilise alors la **régression linéaire multiple** pour **faire la part des choses** et **mesurer l'effet propre** de la pratique d'une activité physique régulière sur le poids à taille donnée.

Activité physique : régression linéaire multiple
Représentation du nuage de points



Activité physique : régression linéaire multiple
Modélisation

On modélise donc la relation entre le poids, la taille et la pratique d'une activité physique régulière par une équation de la forme :

$$poids_i = \beta_0 + \beta_1 \times taille_i + \beta_2 \times sport_i + \varepsilon_i$$

$sport_i$ est une variable indicatrice valant 1 si l'individu pratique une activité physique régulière, et 0 sinon.

Dans le cadre de ce modèle, le coefficient β_2 s'interprète comme l'**effet propre** d'une activité physique régulière sur le poids **à taille donnée**.

Activité physique : régression linéaire multiple
Estimation et interprétation des coefficients

L'estimation conduit aux valeurs suivantes :

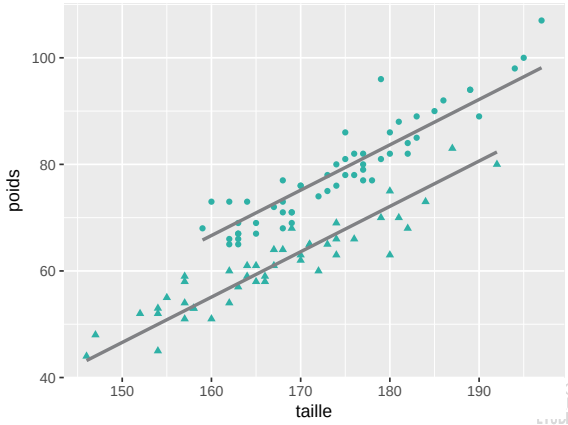
$$\hat{\beta}_0 = -69,48 \quad \hat{\beta}_1 = 0,85 \quad \text{et} \quad \hat{\beta}_2 = -11,55$$

La valeur de $\hat{\beta}_2$ est de -11,55 : en moyenne **et à taille égale**, les personnes qui pratiquent une activité physique régulière pèsent environ 11,55 kg de moins que les autres.

Par rapport au modèle 1, **la valeur des coefficients $\hat{\beta}_0$ et $\hat{\beta}_1$ a changé** :

- dans le modèle 1, le $\hat{\beta}_1$ incorpore à la fois le lien « direct » entre la taille et le poids et le lien « indirect » dû au fait que les personnes petites pratiquent davantage une activité physique régulière ;
- dans le modèle 2, ces effets sont séparés entre $\hat{\beta}_1$ et $\hat{\beta}_2$, et $\hat{\beta}_0$ s'ajuste.

Activité physique : régression linéaire multiple
Représentation de la prédiction du modèle 2



Valeurs prédites par le modèle

$$poids_i^{M2} = -69,48 + 0,85 \times taille_i - 11,55 \times sport_i$$

poids	taille	sport	poids_m1	poids_m2
88	181	0	81,24	84,52
79	177	0	77,08	81,12
61	164	1	63,55	58,5
65	163	0	62,51	69,2
67	165	0	64,59	70,9
69	174	1	73,96	67,01
44	146	1	44,81	43,18
63	170	1	69,79	63,6
71	168	0	67,71	73,46
80	192	1	92,7	82,32

Qualité d'ajustement du modèle

L'examen des valeurs prédites par le modèle (par le biais du graphique notamment) semble indiquer qu'il **rend bien compte de la variabilité des données**.

Selon les modélisations, des **outils plus précis** peuvent être utilisés pour confirmer ce diagnostic (R^2, AIC, etc.).

Une fois le modèle validé, on peut **se concentrer sur l'interprétation des relations** entre variables et en particulier sur leur caractère **statistiquement significatif**.

Imprécision de l'estimation

Les coefficients d'un modèle de régression sont estimés avec une certain **imprécision**.

Intuitivement, cette imprécision est d'autant plus importante :

- que l'**ajustement du modèle est mauvais** ;
- que **le nombre d'observations intervenant dans le modèle est faible**.

Exemple Si l'estimation de $\hat{\beta}_2$ devait être trop imprécise, alors il ne serait **pas possible de conclure à un lien significatif** entre poids et activité physique, en dépit d'une valeur assez élevée en valeur absolue (-11,55 kg).

Erreur standard des coefficients

Pour chaque coefficient, il est possible de calculer une **erreur standard** (*standard error*) qui est **d'autant plus grande que l'estimation est imprécise**.

Variable	Coeff.	Estimation	Erreur Standard
constante	β_0	-69,48	6,24
taille	β_1	0,85	0,04
sport	β_2	-11,55	0,76

Remarque Ces erreurs standards sont obtenues à partir de la **matrice de variance-covariance** du modèle, qui est estimée en même temps que les coefficients.

Construction d'un intervalle de confiance

À partir de l'estimation ponctuelle et de l'erreur standard (se), il est facile de construire un **intervalle de confiance** (IC) :

$$IC_{95\%} = [\hat{\beta}_2 - 1,96 \times se_2; \hat{\beta}_2 + 1,96 \times se_2]$$

Variable	Estimation	Erreur Standard	IC à 95 %
constante	-69,48	6,24	[-81,71 ;-57,25]
taille	0,85	0,04	[0,77 ;0,93]
sport	-11,55	0,76	[-13,04 ;-10,06]

Remarque La valeur 1,96 correspond au quantile à 97,5 % d'une loi normale centrée réduite. On peut en effet montrer que β_2 converge en loi vers une loi normale de moyenne $\hat{\beta}_2$ et d'écart-type se_2.

Construction d'un intervalle de confiance

Si 0 appartient à l'intervalle de confiance à 95 % d'un coefficient, alors celui-ci n'est pas significativement différent de 0 au seuil de 5 %.

Autrement dit, dans ce cas on a plus de 5 % de chances de se tromper en affirmant que la vraie valeur du coefficient est différente de 0.

Construire un intervalle de confiance à 95 % est donc une méthode simple (surtout si on arrondit 1,96 à 2) pour juger de la significativité d'un coefficient.

Une méthode **rigoureusement équivalente** consiste à interpréter la **p-valeur** associée au test de significativité du coefficient.

Test de significativité des coefficients

D'un point de vue statistique, déterminer si l'activité physique régulière est significativement associée à un poids en moyenne plus faible revient à tester l'hypothèse :

$$H_0 : \beta_2 = 0 \quad \text{contre} \quad H_1 : \beta_2 \neq 0$$

Si l'on rejette l'hypothèse H_0 avec un faible risque de se tromper, alors on peut affirmer qu'activité physique régulière et poids sont statistiquement liés.

Pour ce faire, on s'appuie sur une **statistique de test** qui fait intervenir l'estimation ponctuelle et l'erreur standard du coefficient β_2.

Test de siginficativité des coefficients

On définit ainsi la statistique t_2 :

$$t_2 = \frac{|\hat{\beta}_2|}{se_2}$$

1. Quand t_2 est « grand », $\hat{\beta}_2$ est grand devant son erreur standard se_2 : l'imprécision liée à l'estimation est faible et le risque que la vraie valeur de β_2 soit nulle est infime.
2. Quand t_2 est « petit », $\hat{\beta}_2$ est petit devant son erreur standard se_2 : l'imprécision liée à l'estimation est importante et le risque que la vraie valeur de β_2 soit nulle est non-négligeable.

Moralité **Plus t_2 est grand, plus on a tendance à rejeter l'hypothèse H_0 de nullité de β_2.**

Test de siginficativité des coefficients

Pour un seuil statistique de 5 %, la « valeur critique » à partir de laquelle on peut rejeter l'hypothèse nulle est de 1,96.

Concrètement, on rejette $H_0 : \beta_2 = 0$ si la statistique de test t_2 est plus grande que 1,96.

Ici c'est le cas ($t_2 = \frac{|-11,55|}{0,76} = 15,20 > 1,96$) : on peut donc affirmer avec un risque de se tromper inférieur à 5 % qu'activité physique régulière et poids sont statistiquement liés.

Remarque À nouveau, la valeur 1,96 correspond au quantile à 97,5 % d'une loi normale centrée réduite.

Interprétation de la p-valeur

Concrètement, les logiciels de statistique présentent la valeur de la statistique de test (t_2 dans l'exemple) ainsi que la **p-valeur** (*p-value*) associée.

La p-valeur correspond à la probabilité (le risque) que l'on a de se tromper en rejetant l'hypothèse

$$H_0 : \beta_2 = 0 \quad \text{contre} \quad H_1 : \beta_2 \neq 0$$

Variable	Estimation	Erreur Standard	Stat. t	P-valeur
constante	-69,48	6,24	11,13	<0,001
taille	0,85	0,04	23,82	<0,001
sport	-11,55	0,76	15,17	<0,001

Interprétation de la p-valeur

Plus la p-valeur est faible, moins on a de risque de se tromper en affirmant que la variable expliquée est statistiquement liée à la variable explicative.

En particulier, si la p-valeur d'un coefficient est inférieure à 0,05, alors on peut rejeter l'hypothèse H_0 au seuil de 5 %.

Les seuils statistiques conventionnels en sociologie sont :

- * 10 % (p-valeur $< 0,10$) : lien faiblement significatif ;
- ** 5 % (p-valeur $< 0,05$) : lien significatif ;
- *** 1 % p-valeur $< 0,01$) : lien très significatif.

Dans le modèle 2, les p-valeurs des trois coefficients sont inférieures à 0,01 : ils sont tous trois significativement différents de 0 au seuil de 1 %.

Impact de l'activité physique : conclusion

Enfin, les variables en jeu laisse **assez peu de doute quand à la nature des liens de causalité** :

- **taille et le poids** : cela fait écho au concept de **masse corporelle** et à l'indice du même nom ;
- **activité physique et poids** : si dans des cas extrêmes on peut envisager que le poids rende plus difficile une activité physique, de manière générale **il semble bien plus probable que ce soit l'activité physique qui ait un impact négatif sur le poids**.

Relativement simple ici, l'identification des relations de causalité est parfois plus complexe.

Intuition et motivation

Un exemple classique dans lequel la causalité est ambiguë est celui des régimes amaigrissants.

Supposons qu'une partie de l'échantillon ait suivi un régime au cours de l'année précédente. On souhaite mesurer l'efficacité de ces régimes en contrôlant toujours par la taille et la pratique d'une activité physique régulière.

On estime donc un modèle de régression linéaire multiple de la forme :

$$poids_i = \beta_0 + \beta_1 \times taille_i + \beta_2 \times sport_i + \beta_3 \times regime_i + \varepsilon_i$$

Intuition et motivation

Par construction, l'ajustement de ce modèle est **au moins aussi bon que celui du modèle 2** (mais plus difficile à évaluer sans les outils qui seront vus lors de la séance 2).

Variable	Estimation	Erreur Standard	Stat. t	P-valeur
constante	-70,16	6,09	-11,52	<0,001
taille	0,84	0,03	24,20	<0,001
sport	-10,57	0,84	-12,54	<0,001
regime	2,05	0,83	2,47	0,015

La valeur positive de l'estimateur de β_3 peut surprendre : **avoir suivi un régime amaigrissant est statistiquement associé à un poids en moyenne de l'ordre de 2 kg plus élevé**.

Intuition et motivation

En l'état, on pourrait être tenté de voir dans ses résultats une illustration de l'**inefficacité des régimes amaigrissants** et en particulier de l'« effet yo-yo » qui les accompagne souvent.

Cependant, quand on les interroge les personnes qui ont fait un régime au cours de l'année précédente **déclarent avoir *perdu* en moyenne 2 kg**.

Ces éléments sont confirmés par les mesures de poids dont on dispose par ailleurs :

Régime	Poids passé	Poids présent	Évolution
Oui	76,80	75,34	- 1,46
Non	63,07	63,44	+ 0,37

Intuition et motivation

Comment expliquer alors l'estimation du coefficient β_3 dans le modèle ? Est-il possible que le modèle soit faux ?

En réalité, c'est **bien plutôt l'interprétation qu'on a implicitement donné au coefficient qui est fausse**.

Le modèle statistique en lui-même permet de mesurer des **associations entre des variables** et non des relations de causalité.

La question à se poser : est-ce que le fait que le coefficient soit positif et significatif **dans ce contexte** signifie nécessairement que faire un régime induit une prise de poids ?

L'efficacité des régimes : exemple d'endogénéité
Endogénéité et causalité inverse

Ce coefficient positif doit en fait s'interpréter comme le fait que les ***personnes qui ont fait un régime*** pèsent en moyenne 2 kg de plus que les autres.

Or ces personnes ne sont **pas réparties au hasard dans l'échantillon** : **ce sont de fait les personnes qui présentent l'indice de masse corporelle le plus élevé qui ont le plus souvent décidé de faire un régime**.

D'où le résultat : **quand bien même le régime a eu une véritable efficacité au niveau individuel, dans la mesure où les personnes qui l'ont adopté ont un poids en moyenne plus élevé, régime amaigrissant et poids sont liés positivement**.

L'efficacité des régimes : exemple d'endogénéité
Endogénéité et causalité inverse

Cet exemple simple illustre le mécanisme de la **causalité inverse** : on s'attend certes à ce que le fait de faire un régime induise une diminution de poids, mais inversement il peut s'avérer que le fait d'être en surpoids incite davantage à faire un régime.

La causalité inverse est une des formes les plus spectaculaire d'**endogénéité**, qui est susceptible d'inverser totalement le sens d'un coefficient par rapport à la valeur attendue.

D'autres formes d'endogénéité peuvent survenir quand des **variables explicatives importantes ne sont pas intégrées au modèle** ou quand la **variable expliquée est mesurée imparfaitement**.

Remarques conclusives
L'importance du choix du modèle

Les méthodes de régression reposent sur une **modélisation des données** : la réflexion sur le choix du modèle est essentielle pour aboutir à des résultats qui tiennent la route.

Il est particulièrement essentiel de **tester un grand nombre de modèles** pour être en mesure de choisir la spécification la plus appropriée.

Plus généralement, **la nature de la variable dépendante** commande un certain type de modélisation :

- Les modèles **linéaires ou log-linéaires** correspondent assez bien aux variables **quantitatives continues** ;
- Les modèles **logistiques ou probit** correspondent aux variables **qualitatives dichotomiques ou polytomiques**.

Remarques conclusives
« Effet pur » et inférence

L'apport principal des méthodes de régression est de permettre de **décomposer les effets des différentes variables**.

Cependant, l'application non-réfléchie des coefficients de régression peut conduire à des **fictions statistiques** (Halbwachs,1935) :

> *« Cela revient [...] à se demander comment vivrait un chameau, si, restant chameau, il était transporté dans les régions polaires, et comment vivrait un renne si, restant un renne, il était transporté dans le Sahara. »*

À la recherche de l' « effet pur » et du raisonnement « toutes choses égales par ailleurs », on peut préférer **l'inférence statistique sous un modèle donné** (nécessairement imparfait).

Remarques conclusives
La question de la causalité

La mise en évidence de liens causaux est en arrière plan de toute la construction mathématique et épistémologique des modèles économétriques.

Cependant, **rares sont les situations où le modèle peut à lui seul** conduire à des conclusions fermes en termes de liens causaux :

- expériences contrôlées ;
- « expériences naturelles » et discontinuités ;
- données de panel suffisamment riches.

En règle générale, la mise en évidence d'effets causaux ne peut donc pas reposer sur le seul modèle, mais sur son **articulation avec la littérature du champ de recherche ou avec d'autres travaux empiriques** (quantitatifs ou qualitatifs).

Remarques conclusives
Les trois finalités des méthodes de régression

De manière générale, la mise en œuvre d'un modèle de régression passe donc par **trois étapes** :

1. Construire un modèle pertinent :
2. Décrire les relations entre variables et juger de leur significative ;
3. Interpréter ces relations en termes de relations causales.

Ces trois « étapes » peuvent également être vues comme **trois finalités distinctes des méthodes de régression** :

1. **Prédire** : modèles de prévision macro-économiques, apprentissage statistique, etc.
2. **Décrire** : analyse de données en sciences sociales ;
3. **Expliquer** : analyse statistique de réseaux, modélisation micro-économique des comportements, etc.

Remarques conclusives
Les trois finalités des méthodes de régression

En fonction de la finalité, l'usage des méthodes de régression et les outils de validation peuvent différer.

1. Dans une optique de prévision, **le choix des variables à intégrer dans le modèle peut être automatisé** afin de maximiser la qualité de la prévision et **les coefficients n'ont pas nécessairement à avoir une interprétation**.
2. Dans une perspective plus descriptive, le choix des variables explicatives est **lié à la problématique d'étude** et l'essentiel de l'interprétation porte sur **le signe et la significativité des coefficients**.
3. Dans une perspective plus fortement explicative, le choix des variables est **déterminé par le modèle théorique sous-jacent** et **c'est en général la valeur même des coefficients qui est interprétée** (élasticité, etc.).

Régression linéaire

Martin Chevalier (Insee)

Année universitaire 2017-2018

Régression linéaire
Objectifs de la séance

Introduire la **régression linéaire multiple** par les moindres carrés ordinaires (MCO).

Savoir **interpréter les coefficients** d'une régression dans des contextes variés.

Savoir utiliser les **statistiques d'ajustement de modèle**.

Régression linéaire
Données d'exemple : Salaire dans l'EEC

Données du quatrième trimestre de l'**Enquête emploi en continu** (EEC) 2012 (page de présentation sur le site de l'Insee).

Informations sur l'emploi, les salaires, la situation socio-économique des personnes interrogées.

Sélection d'un sous-échantillon au 1/20ème : 1 783 observations.

Plusieurs modèles sont construits pour essayer de modéliser les **écarts de salaire entre femmes et hommes**.

Régression linéaire
Données d'exemple : Salaire dans l'EEC

```
# Importation des données depuis teaching.slmc.fr
e <- read.csv(
  "http://teaching.slmc.fr/mqs2/ee12t4.csv"
  , stringsAsFactors = FALSE
)

# Dimension de la table
dim(e)
## [1] 1783   19

# Nom des variables
names(e)
##  [1] "ident"     "noi"       "extri1613" "sexe"
##  [5] "age"       "cse"       "acteu"     "stc"
##  [9] "tam1d"     "aidref"    "salred"    "tpp"
## [13] "ddipl"     "nbagenf"   "duhab"     "pub3fp"
## [17] "naia"      "fordat"    "ancentr"
```


Régression linéaire
Données d'exemple : Salaire dans l'EEC

```
# Statistiques descriptives sur la variable de salaire
summary(e$salred)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##      31    1250    1600    1851    2168   11423    1004

# Restriction aux observations avec un salaire
e <- e[!is.na(e$salred), ]

# Nombre d'observations restantes
nrow(e)
## [1] 779

# Ecart de salaire femme-homme (non-pondéré)
tapply(e$salred, e$sexe, mean)
##        1        2
## 2177.308 1513.940

# Ecart de salaire femme-homme (pondéré)
sapply(split(e, e$sexe), function(x) weighted.mean(x$salred, x$extri1613))
##        1        2
## 2230.164 1539.230
```


La régression par les moindres carrés ordinaires
Quand utiliser la régression par les MCO ?

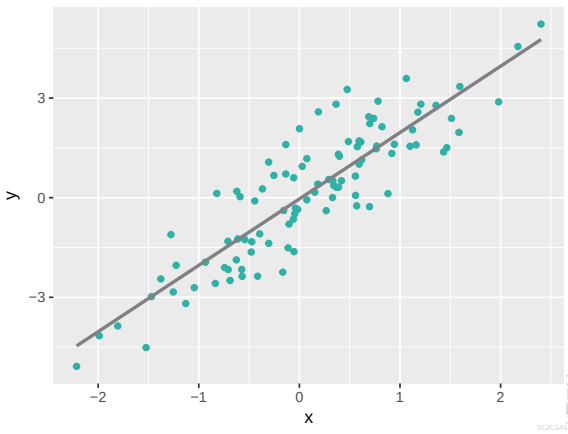
La régression par les MCO est utilisée quand la variable expliquée est **quantitative et continue**.

Exemples Prix, salaire, etc.

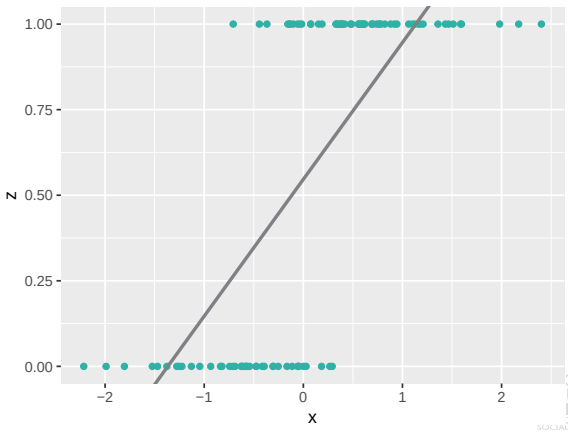
Elle est en revanche inappropriée :

- quand la variable expliquée est qualitative (dichotomique ou polytomique) : **modèles logistiques** ;
- quand la variable expliquée est quantitative discrète : **modèle de Poisson ou négatif-binomiaux**.

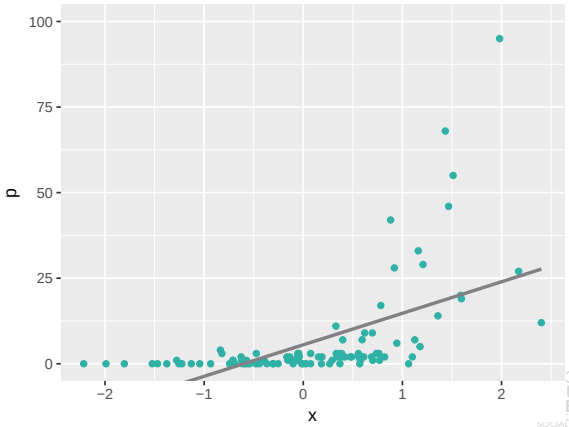
La régression par les moindres carrés ordinaires
Quand utiliser la régression par les MCO ?



La régression par les moindres carrés ordinaires
Quand utiliser la régression par les MCO ?



Quand utiliser la régression par les MCO ?



La forme du modèle

La régression linéaire multiple par les moindres carrés ordinaires (MCO, OLS en anglais) est un modèle de régression qui s'écrit sous la forme :

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \dots \beta_p x_{p,i} + \varepsilon_i \qquad i = 1, \dots, n$$

- Y est la **variable expliquée** ;
- $X = (1\ X_1\ \dots\ X_p)$ est la **matrice de variables explicatives** ;
- $\beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}$ est le vecteur de paramètres ;
- ε est le résidu.

Principe de l'estimation

L'estimation du modèle est menée en cherchant à **maximiser l'ajustement de la prédiction aux vraies valeurs** de Y.

Mathématiquement, on traduit cet objectif par la minimisation de la **somme des carrés des résidus** (SCR) : $\sum_{i=1}^{n} \varepsilon_i^2$.

Dans le cas des MCO, l'estimation des coefficients β peut être obtenue directement avec une formule simple faisant intervenir X et Y (et des opérations matricielles).

Pour résumer Dans un modèle de régression linéaire multiple, la valeur de β qui maximise l'ajustement du modèle aux données est obtenue par un calcul direct.

Principe de l'estimation

$$salaire_i = \beta_0 + \beta_1 age_i + \beta_2 femme_i + \varepsilon_i \qquad i = 1, \dots, n$$

- age est une variable quantitative
- femme est une variable indicatrice

Intégrer ces deux variables dans un modèle de **régression multivariée** permet :

- d'interpréter la relation entre âge et salaire **à sexe égal par ailleurs** ;
- d'interpréter la relation entre sexe et salaire **l'âge étant égal par ailleurs**.

Remarque Le plus souvent c'est le **logarithme du salaire** que l'on modélise par une régression linéaire.

Principe de l'estimation

Principe de l'estimation

```r
# Estimation du modèle et stockage de ses résultats
# dans l'objet m1
m1 <- lm(salred ~ age + femme, data = e)

# Affichage des résultats élémentaires du modèle
m1
##
## Call:
## lm(formula = salred ~ age + femme, data = e)
##
## Coefficients:
## (Intercept)          age        femme
##     1267.33        21.46      -644.94
```


Principe de l'estimation

```r
# Affichage des résultats détaillés du modèle
summary(m1)
##
## Call:
## lm(formula = salred ~ age + femme, data = e)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2255.7  -636.1  -167.3   388.7  9533.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1267.327    161.036   7.870 1.19e-14 ***
## age           21.464      3.557   6.033 2.48e-09 ***
## femme       -644.939     80.537  -8.008 4.26e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1123 on 776 degrees of freedom
## Multiple R-squared:  0.1185, Adjusted R-squared:  0.1163
## F-statistic: 52.17 on 2 and 776 DF,  p-value: < 2.2e-16
```


Significativité des coefficients

L'estimation du modèle n'étant parfaite (les résidus ne sont jamais tous exactement nuls), les coefficients sont estimés avec une certaine imprécision.

Pour juger de la significativité statistique d'un coefficient β_j dans une régression par les MCO, **on pose le test** :

$$H_0 : \beta_j = 0 \quad \text{contre} \quad H_1 : \beta_j \neq 0$$

au seuil statistique α désiré (0,05 ou 0,01 en général).

La régression par les moindres carrés ordinaires
Significativité des coefficients

En pratique, cela revient :

- soit à comparer la valeur d'une **statistique de test** aux quantiles d'une loi normale centrée réduite ;
- soit à interpréter la **p-valeur** ;
- soit à construire l'**intervalle de confiance** au niveau correspondant au seuil α désiré et à vérifier si 0 y appartient.

Pour aller plus loin La statistique de test est $t = \frac{\hat{\beta}_j}{\hat{se}_j}$ (où $\hat{se}_j$ est l'erreur-standard de β_j) et suit en fait sous l'hypothèse H_0 une loi de Student à $n-1$ degrés de liberté.

Ses quantiles sont très proches de ceux d'une loi normale centrée réduite dès lors que le nombre d'observations dépasse 100.

La régression par les moindres carrés ordinaires
Significativité des coefficients

```
# Statistique de test et p-valeur
coef(summary(m1))
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 1267.32696 161.035678  7.869852 1.192889e-14
## age           21.46361   3.557485  6.033366 2.484983e-09
## femme       -644.93936  80.536756 -8.008013 4.258557e-15

# Intervalle de confiance à 95 %
confint.default(m1)
##                 2.5 %     97.5 %
## (Intercept)  951.70284 1582.95109
## age           14.49107   28.43615
## femme       -802.78850 -487.09022
```


Interprétation des coefficients
Adapter l'interprétation à la situation

L'interprétation d'un coefficient diffère **selon la nature et la manière** dont est intégrée la variable explicative correspondante dans le modèle :

- variable quantitative ;
- variable quantitative et son carré ;
- variable qualitative ;
- interaction entre plusieurs variables.

Interprétation des coefficients
Variable explicative quantitative

L'interprétation de la valeur du coefficient associé à une variable explicative quantitative est directe.

Elle correspond à l'**augmentation moyenne dans l'échantillon** de la valeur de la variable expliquée associée à une **augmentation de 1** de la valeur de la variable explicative, les autres variables étant égales par ailleurs.

Exemple Dans le premier modèle l'estimateur du coefficient associé à l'âge $\hat{\beta}_1$ vaut 21,46 : à une augmentation de 1 an est en moyenne associée une augmentation de 21,46 euros, à sexe égal par ailleurs.

Cet « **effet marginal** » est le même où que l'on se place dans la distribution de la variable explicative et pour tous les individus.

Interprétation des coefficients
Variable explicative quantitative

Interprétation des coefficients
Variable explicative quantitative et son carré

Il est tout à fait possible dans un modèle de régression linéaire d'introduire une variable explicative accompagnée d'**une ou plusieurs de ses puissances**.

Exemple

$$salaire_i = \beta_0 + \beta_1 age_i + \beta_2 femme_i + \beta_3 age_i^2 + \varepsilon_i$$

Remarque Le modèle reste alors bien **linéaire en ses paramètres** : les relations entre variables et paramètres sont uniquement des sommes et des multiplications.

Ce faisant, il devient possible de capter des **relations non-linéaires** : avec un carré en particulier, des relations apparaissant graphiquement comme des **paraboles**.

Interprétation des coefficients
Variable explicative quantitative et son carré

Interprétation des coefficients
Variable explicative quantitative et son carré

Dans cette situation on autorise l'effet marginal de la variable explicative à **ne pas être le même selon la position** dans la distribution de la variable explicative.

Exemple Le gain moyen de salaire associé à une année supplémentaire peut être supérieur en début à ce qu'il est en fin de carrière.

Passer un certain niveau, la relation entre variable explicative et variable expliquée **peut s'inverser**.

Pour aller plus loin Mathématiquement, on peut alors montrer que

$$\left(\frac{\partial salaire}{\partial age}\right)_i = \hat{\beta}_1 + 2\hat{\beta}_3 age_i$$

Interprétation des coefficients
Variable explicative qualitative

Pour être intégrées dans un modèle, les variables explicatives doivent toutes être de type numérique.

Mais bien souvent en sociologie, **les variables explicatives d'intéret sont de nature qualitative** (diplôme, catégorie sociale, etc.) : il n'est pas possible de les exploiter directement sous une forme numérique.

Exemple La PCS est **codée par des nombres**, par exemple 2 pour les artisans et 6 pour les ouvriers. Intégrer cette variable telle quelle reviendrait à affirmer qu'un artisan correspond à 3 fois moins de quelque chose (?) qu'un ouvrier.

Interprétation des coefficients
Variable explicative qualitative

Pour intégrer des variables de nature qualitative au modèle, on procède par dichotomisation :

1. Pour chaque modalité de la variable (ex : "1" et "2" pour le sexe), on crée une **variable indicatrice** valant 1 si l'individu présente la modalité et 0 sinon.
2. Ce sont ces variables indicatrices que l'on intègre au modèle, **sauf une** : la **modalité de référence**.

Pour aller plus loin D'un point de vue mathématique, intégrer toutes les variables indicatrices d'une même variable qualitative introduirait dans le modèle une colinéarité (≈ redondance) parfaite avec la constante.

Interprétation des coefficients
Variable explicative qualitative

```
# Intégration de toutes les indicatrices associées
# à une variable qualitative
lm(salred ~ age + femme + homme, data = e)
##
## Call:
## lm(formula = salred ~ age + femme + homme, data = e)
##
## Coefficients:
## (Intercept)          age        femme        homme
##     1267.33        21.46      -644.94           NA
# Note : la variable homme est neutralisée
# automatiquement pour pouvoir estimer le modèle
# (sinon : colinéarité parfaite).
```


Interprétation des coefficients
Variable explicative qualitative

L'interprétation des variables qualitatives **diffère** de celle des variables quantitatives et **fait intervenir la modalité de référence**.

On interprète ainsi le coefficient associée à la modalité j d'une variable qualitative comme la **différence moyenne** dans la valeur de la variable expliquée **entre les individus présentant la modalité** j et **ceux présentant la modalité de référence**, toutes les autres variables du modèle étant égales par ailleurs.

Exemple Dans le premier modèle le coefficient associé à l'indicatrice « femme » vaut -644,94 : au fait d'être une femme (plutôt qu'un homme) est associé en moyenne un salaire inférieur de l'ordre de -644,94 euros, à âge égal par ailleurs.

Interprétation des coefficients
Variable explicative qualitative

Le choix de la modalité de référence importe peu pour l'interprétation quand la variable qualitative est par nature dichotomique (sexe, etc.).

Dans ce cas en effet, la valeur du coefficient concerné est juste multipliée par -1.

```
coef(lm(salred ~ age + femme, data = e))
## (Intercept)         age       femme
##  1267.32696    21.46361  -644.93936
coef(lm(salred ~ age + homme, data = e))
## (Intercept)         age       homme
##   622.38761    21.46361   644.93936
```

Remarque Dans tous les cas la manière d'intégrer une variable qualitative n'affecte que ses coefficients, pas ceux des autres variables du modèle (sauf la constante).

Interprétation des coefficients
Variable explicative qualitative

Le choix de la modalité de référence **affecte en revanche l'interprétation des variables qualitatives polytomiques**, en particulier quand elles sont ordonnées :

- En choisissant une modalité « extrême », les coefficients auront **davantage tendance à être significatifs** aux seuils usuels ;
- Mais en choisissant une modalité « centrale », **l'interprétation des tests de significativité pourra être beaucoup plus riche**.

Exemple

```
# On compare trois modèles
m_infbac <- lm(salred ~ age + sexe + bac + supbac, data = e)
m_bac <- lm(salred ~ age + sexe + infbac + supbac, data = e)
m_supbac <- lm(salred ~ age + sexe + infbac + bac, data = e)
```


Interprétation des coefficients
Variable explicative qualitative

```
# Coefficients relatifs au diplôme dans m_infbac
coef(summary(m_infbac))[4:5, ]
##         Estimate Std. Error   t value     Pr(>|t|)
## bac     571.7658   97.75837  5.848766 7.303926e-09
## supbac 1025.9798   86.21312 11.900507 4.152211e-30

# Coefficients relatifs au diplôme dans m_bac
coef(summary(m_bac))[4:5, ]
##         Estimate Std. Error   t value     Pr(>|t|)
## infbac -571.7658   97.75837 -5.848766 7.303926e-09
## supbac  454.2140  102.26736  4.441437 1.023730e-05

# Coefficients relatifs au diplôme dans m_supbac
coef(summary(m_supbac))[4:5, ]
##         Estimate Std. Error    t value     Pr(>|t|)
## infbac -1025.980   86.21312 -11.900507 4.152211e-30
## bac     -454.214  102.26736  -4.441437 1.023730e-05
```


Interprétation des coefficients
Interaction entre plusieurs variables

Bien souvent, les hypothèses de recherche impliquent **plusieurs variables explicatives du modèle** :

- le lien entre âge et salaire diffère-t-il selon le sexe ?
- le lien entre avoir un diplôme supérieur au bac et salaire diffère-t-il selon le sexe ?

Les modèles construits précédemment **ne permettent pas** de répondre à ce genre de question.

Pour ce faire, il faut introduire des **interactions entre variables explicatives** :

$$salaire_i = \beta_0 + \beta_1 age_i + \beta_2 femme_i + \beta_3 age_i \times femme_i + \varepsilon_i$$

```
# Création de la variable d'interaction
e$age_femme <- e$age * e$femme

# Estimation du nouveau modèle
m3 <- lm(salred ~ age + femme + age_femme, data = e)

# Résulats sommaires
m3
##
## Call:
## lm(formula = salred ~ age + femme + age_femme, data = e)
##
## Coefficients:
## (Intercept)          age        femme    age_femme
##      760.50        33.42       363.23       -24.02
```


Dans cet exemple, **on autorise l'effet marginal de l'âge sur le salaire à varier selon le sexe** :

- pour les hommes, l'effet marginal de l'âge est capté par β_1 ;
- pour les femmes, l'effet marginal de l'âge est capté par $\beta_1 + \beta_3$;

On interprète alors :

- pour les hommes, une année supplémentaire est en moyenne associée à unn salaire de 33,42 euros supérieur ;
- pour les femmes, une année supplémentaire est en moyenne associée à unn salaire de 9,39 euros supérieur

Le test de significativité de β_3 permet de déterminer le caractère statistiquement significatif de l'**écart en termes de progression salariale avec l'âge entre hommes et femmes**.

```
# Coefficients du modèle m3
coef(summary(m3))
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 760.49951 218.662260  3.477964 5.334133e-04
## age          33.41808   4.985237  6.703409 3.920373e-11
## femme       363.23416 307.172556  1.182508 2.373667e-01
## age_femme   -24.02409   7.067146 -3.399405 7.098578e-04
```


Pour aller plus loin Pour tester si âge et salaire sont statistiquement liés chez les femmes, il faudrait mener le test :

$$H_0 : \beta_1 - \beta_3 = 0 \quad \text{contre} \quad H_1 : \beta_1 - \beta_3 \neq 0$$

Ce test est un **test d'hypothèse complexe** faisant intervenir **plusieurs coefficients**.

En pratique, on peut le mener en comparant le pouvoir prédictif (*cf. infra*) du modèle où on contraint $\beta_1 = \beta_3$ à celui où on autorise $\beta_1 \neq \beta_3$.

Il est également possible de faire interagir une variable qualitative **avec une variable quantitative et son carré** :

$$\begin{aligned} salaire_i = \beta_0 + \beta_1 age_i + \beta_2 femme_i + \beta_3 age_i^2 \\ + \beta_4 age_i \times femme_i + \beta_5 age_i^2 \times femme_i + \varepsilon_i \end{aligned}$$

```
m4 <- lm(
  salred ~ age + femme + I(age^2) + femme:(age + I(age^2))
  , data = e
)
```

Remarque Déterminer la significativité de l'association entre l'âge et le salaire implique de tester la nullité de plusieurs paramètres simultanément (test d'hypothèse complexe).

On peut enfin très facilement intégrer des interactions **entre deux variables qualitatives** en définissant une nouvelle **variable croisée**.

Exemple Croisement du sexe et du niveau de diplôme

sexe \ diplôme	**inf_bac ou bac**	**supbac**
homme	homme_infbac	homme_supbac
femme	femme_infbac	femme_supbac

On intègre alors cette nouvelle variable en choisissant une **nouvelle modalité de référence** :

$$\begin{aligned} salaire_i = \beta_0 + \beta_1 age_i + \beta_2 homme\_supbac_i + \beta_3 femme\_infbac_i \\ + \beta_4 femme\_supbac_i + \varepsilon_i \end{aligned}$$

Interprétation des coefficients
Interaction entre plusieurs variables

$$salaire_i = \beta_0 + \beta_1 age_i + \beta_2 homme\_supbac_i + \beta_3 femme\_infbac_i + \beta_4 femme\_supbac_i + \varepsilon_i$$

L'interprétation des coefficients s'effectue classiquement **par rapport à la modalité de référence** :

- β_2 : écart moyen de salaire entre hommes non-diplômés et diplômés du supérieur (à âge égal par ailleurs) ;
- β_3 : écart moyen de salaire entre hommes et femmes non-diplômés du supérieur (*idem*) ;
- β_4 : écart moyen de salaire entre hommes non-diplômés du supérieur et femmes diplômées du supérieur (*idem*).

Remarque Pour tester la significativité de l'écart moyen de salaire entre femmes non-diplômées et diplômées du supérieur, il suffit de **changer la modalité de référence**.

Interprétation des coefficients
Interaction entre plusieurs variables

```
# Création des variables d'interaction
e$homme_infbac <- e$homme * (e$infbac | e$bac)
e$homme_supbac <- e$homme * e$supbac
e$femme_infbac <- e$femme * (e$infbac | e$bac)
e$femme_supbac <- e$femme * e$supbac

# Intégration dans le modèle
m5 <- lm(
  salred ~ age + homme_supbac + femme_infbac + femme_supbac,
  data = e
)
coef(summary(m5))
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)   845.21499 156.147710  5.4129195 8.276664e-08
## age            24.31037   3.375283  7.2024693 1.403514e-12
## homme_supbac 1011.54241 115.597647  8.7505450 1.316197e-17
## femme_infbac -575.71231  92.170455 -6.2461698 6.935511e-10
## femme_supbac   80.24201 110.727652  0.7246791 4.688679e-01
```


Interprétation des coefficients
Interaction entre plusieurs variables

```
# Changement de la modalité de référence
m5b <- lm(
  salred ~ age + homme_infbac + homme_supbac + femme_supbac,
  data = e
)
coef(summary(m5b))
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)   269.50268 161.222354  1.671621 9.500326e-02
## age            24.31037   3.375283  7.202469 1.403514e-12
## homme_infbac  575.71231  92.170455  6.246170 6.935511e-10
## homme_supbac 1587.25472 117.848055 13.468654 2.667410e-37
## femme_supbac  655.95432 113.575731  5.775480 1.111143e-08
```


Juger de la qualité d'une régression par les MCO
Différents types d'indicateurs

Pour juger de la qualité de l'ajustement des modèles de régression linéaire, deux types de statistiques sont disponibles :

- **les indicateurs du type** R^2 : ils rendent compte de la part de la variance de Y expliquée par le modèle et mesurent son caractère prédictif.
- **le test de significativité globale** : il teste la nullité simultanée de tous les coefficients (sauf la constante) et rend compte du caractère explicatif du modèle.

Dans tous les cas ces indicateurs sont construits à partir de la **somme des carrés des résidus** : plus elle est faible devant la variance de Y, meilleur est le modèle.

Juger de la qualité d'une régression par les MCO
Le R^2 et le R^2 ajusté

Le R^2 correspond au ratio de la variance de y expliquée par le modèle (somme des carrés expliqués ou SCE) sur la variance totale de Y (somme des carrés totaux ou SCT) :

$$R^2 = \frac{SCE}{SCT} = \frac{SCT - SCR}{SCT} = 1 - \frac{SCR}{SCT}$$

avec SCR la somme des carrés des résidus (ce que cherche à minimiser l'estimation du modèle).

Le R^2 est compris entre 0 et 1. Plus il est proche de 1, plus le modèle explique bien la variabilité de Y.

Par construction, **le R^2 augmente dès qu'on ajoute une variable** (même sans aucun lien avec Y). Pour corriger ce phénomène on privilégie le R^2 **ajusté**.

Juger de la qualité d'une régression par les MCO
Le test de Fisher de significativité globale

Le test de significativité globale repose sur les mêmes bases théoriques que les tests de significativité ou les tests plus complexes sur un ou plusieurs coefficients.

Il teste simultanément la nullité de tous les coefficients du modèle sauf la constante.

Si on ne peut pas rejeter ce test à un seuil raisonnable, cela signifie qu'**on ne peut pas écarter l'hypothèse que le modèle ne nous apprenne rien de plus qu'une simple moyenne** (pas de variable explicative).

En pratique, **on interprète la p-valeur du test calculée par le logiciel**. La statistique de test est appelée **F-stat** car elle suit sous H_0 une loi de Fisher.

Juger de la qualité d'une régression par les MCO
Indicateurs de qualité

```
#   Estimations de modèles de plus en plus complets
m1 <- lm(salred ~ age + femme, data = e)
m2 <- lm(salred ~ age + femme + I(age^2), data = e)
m3 <- lm(salred ~ age + femme + I(age^2) + infbac + supbac, data = e)
m4 <- lm(salred ~ age + femme + I(age^2) + bac + supbac, data = e)
m5 <- lm(salred ~
  age + femme + I(age^2) + infbac + supbac + tpp
  , data = e
)
m6 <- lm(salred ~
  age + femme + I(age^2) + infbac + supbac + tpp + public
  , data = e
)

# Centralisation des  modèles dans une liste
modeles <- list(m1, m2, m3, m4, m5, m6)
names(modeles) <- c("m1", "m2", "m3", "m4", "m5", "m6")
```


Juger de la qualité d'une régression par les MCO
Indicateurs de qualité

```
# Résultats détaillés du modèle 1
summary(m1)
##
## Call:
## lm(formula = salred ~ age + femme, data = e)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2255.7  -636.1  -167.3   388.7  9533.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1267.327    161.036   7.870 1.19e-14 ***
## age           21.464      3.557   6.033 2.48e-09 ***
## femme       -644.939     80.537  -8.008 4.26e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1123 on 776 degrees of freedom
## Multiple R-squared:  0.1185, Adjusted R-squared:  0.1163
## F-statistic: 52.17 on 2 and 776 DF,  p-value: < 2.2e-16
```

Indicateurs de qualité

```
# Comparaison des R2
as.matrix((sapply(modeles, function(m) summary(m)$r.squared)))
##         [,1]
## m1 0.1185271
## m2 0.1298831
## m3 0.2611738
## m4 0.2611738
## m5 0.3228561
## m6 0.3232154

# Comparaison des R2 ajustés
as.matrix(sapply(modeles, function(m) summary(m)$adj.r.squared))
##         [,1]
## m1 0.1162553
## m2 0.1265149
## m3 0.2563949
## m4 0.2563949
## m5 0.3175933
## m6 0.3170708
```


Indicateurs de qualité

```
# Modèle avec salaire en logarithme
m7 <- lm(I(log(salred))~
  age + femme + I(age^2) + infbac + supbac + tpp + public
  , data = e
)

# Modèle avec salaire en logarithme
summary(m7)$r.squared
## [1] 0.4579202

# Modèle avec salaire en logarithme
coef(summary(m7))
##                 Estimate   Std. Error    t value      Pr(>|t|)
## (Intercept)  6.0223454397 0.2029605350  29.672495 1.164115e-129
## age          0.0668411029 0.0102172906   6.541960  1.105845e-10
## femme       -0.2528731593 0.0363933622  -6.948332  7.870216e-12
## I(age^2)    -0.0006254238 0.0001236448  -5.058229  5.292871e-07
## infbac      -0.2733488449 0.0454610397  -6.012816  2.813199e-09
## supbac       0.2040462649 0.0475606124   4.290236  2.011802e-05
## tpp         -0.7386474749 0.0469001102 -15.749376  1.160515e-48
## public       0.0657518253 0.0432146094   1.521518  1.285398e-01
```

Méthodes quantitatives pour la sociologie 2

Régression logistique dichotomique

Martin Chevalier (Insee)

Année universitaire 2017-2018

Introduction : Modéliser des données qualitatives
Limites de la régression linéaire classique

Les modèles de régression linéaire « classiques » (moindres carrés ordinaires notamment) ont été pensés pour modéliser une **variable expliquée quantitative continue**.

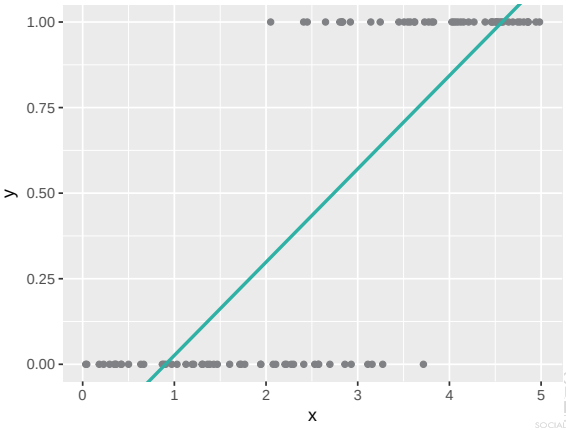
Exemple Salaire net mensuel en fonction du sexe.

Pourtant, le plus souvent en sociologie les variables d'intérêt ne sont pas de nature quantitative mais **qualitative**.

Exemple Chômage en fonction du niveau de diplôme.

Les modèles de régression linéaires classiques ne sont **pas les mieux adaptés** pour modéliser des données qualitatives.

Introduction : Modéliser des données qualitatives
Limites de la régression linéaire classique



Introduction : Modéliser des données qualitatives
Solution : Généraliser le modèle linéaire

Pour modéliser ce type de **données dichotomiques**, on utilise un **modèle linéaire général** du type :

$$f(y_i) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_p x_{p,i} + \varepsilon_i$$

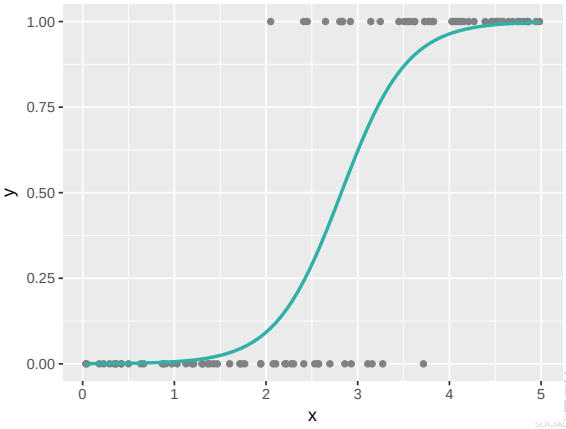
Linéaire ?

- Ce modèle est « **linéaire en ses coefficients** » : les opérations entre les X et β sont uniquement des **sommes ou des multiplications**...
- ... mais il peut modéliser des relations *non-linéaires* grâce notamment à la **fonction de lien** f.

Exemples de modèle non-linéaire en ses coefficients

$$y_i = \beta_0 + x_i^{\beta_1} + \varepsilon_i$$

Introduction : Modéliser des données qualitatives
Régression logistique dichotomique



Introduction : Modéliser des données qualitatives
Objectifs de la séance

Introduire le **modèle linéaire général** et l'appliquer au cas des **données dichotomiques**.

Insister sur les **spécificités de la régression logistique dichotomique** :

- interprétation des coefficients ;
- évaluation de la pertinence du modèle ;
- utilisation des *odds-ratio* et calcul d'effets marginaux moyens ;
- tests d'hypothèses complexes.

Introduction : Modéliser des données qualitatives
Données utilisées pour les exemples

Les données utilisées pour les exemples s'appuient sur une étude de la **probabilité d'être en emploi stable** à partir de l'enquête Emploi en continu 2012T4.

Plusieurs questions de l'enquête Emploi en continu permettent de déterminer la stabilité du contrat de travail :

- **type de contrat** (`CONTRA`) : CDI, CDD, contrat saisonnier, intérim, apprentissage ou alternance ;
- **appartenance à la fonction publique** (`CHPUB`) ;
- **statut au sein de la fonction publique** (`TITC`) : titulaire, stagiaire ou contractuel.

On considère comme **en contrat stable les individus** :

- soit sous contrat de droit privé (y compris contractuels du public) en CDI ;
- soit fonctionnaires titulaires.

Introduction : Modéliser des données qualitatives
Données utilisées pour les exemples

```
# Lecture du sous-échantillon
e <- readRDS(gzcon(url(
  "http://teaching.slmc.fr/mqs2/ee12t4_logit.rds"
)))

# Restriction aux actifs occupés
# (et restriction aux moins de 70 ans)
e <- e[e$ACTEU == "1", ]
e$age <- as.numeric(e$AGE)
e <- e[e$age < 70, ]

# Création de la variable stable
e$stable <- 1 * (e$CONTRA %in% "1" |
  (e$CHPUB %in% c("1", "2", "3") & e$TITC %in% "2"))
```

Nombre de contrats considérés comme stables	751
Nombre de contrats considérés comme instables	257

Données utilisées pour les exemples

On cherche tout particulièrement à mesurer la relation entre stabilité du contrat et variables **socio-démographiques** :

- **âge** : les plus jeunes sont-ils surreprésentés parmi les titulaires d'un contrat de travail instable ?
- **sexe** : constate-t-on un écart entre hommes et femmes en matière de stabilité du contrat de travail ?
- **diplôme** : un diplôme élevé protège-t-il de la précarité associée à un contrat de travail instable ?

On souhaite autant que possible prendre également en compte le **secteur d'activité agrégé** de l'entreprise (primaire, industrie, construction ou tertiaire).

Principe d'estimation

Contrairement au modèle linéaire classique, il n'existe **aucune formule qui donne directement** la valeur de $\hat{\beta}$.

On utilise donc des **algorithmes d'optimisation** pour maximiser une certaine **fonction objectif**, la **(log-)vraisemblance** du modèle.

La **forme de la log-vraisemblance** dépend de la **spécification du modèle** :

- la **distribution supposée** de Y : gaussienne, binomiale, gamma, poissonienne, etc.
- la **fonction de lien** utilisée : identité, logarithme, inverse, logit, etc.

Estimation par maximum de vraisemblance

L'algorithme utilisé est **itératif** : il **maximise la log-vraisemblance** ℓ_n jusqu'à atteindre un **maximum**.

Itération 1 $\ell_n =$ -28,62 $\beta_0 =$ -3,63 $\beta_1 =$ 1,33

Itération 2 $\ell_n =$ -24,16 $\beta_0 =$ -5,55 $\beta_1 =$ 1,98

Itération 3 $\ell_n =$ -23,04 $\beta_0 =$ -7,05 $\beta_1 =$ 2,50

Itération 4 $\ell_n =$ -22,92 $\beta_0 =$ -7,75 $\beta_1 =$ 2,75

Itération 5 $\ell_n =$ **-22,92** $\beta_0 =$ **-7,86** $\beta_1 =$ **2,79**

Régression logistique dichotomique

Le modèle logistique dichotomique est une **spécification** du modèle linéaire général, que l'on peut réécrire :

$$\mathit{logit}\left[\mathbb{P}(y_i = 1|X_i)\right] = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_p x_{p,i} + \varepsilon_i$$

Ses caractéristiques sont les suivantes :

- la quantité modélisée est la **probabilité** que la variable Y prenne la modalité 1 plutôt que la modalité 0 ;
- il appartient à la **famille binomiale** au sein des modèles linéaires généraux ;
- sa **fonction de lien** est la fonction *logit* :

$$\mathit{logit}(p) = \ln\left(\frac{p}{1-p}\right)$$

Modèle linéaire général

Formulation du modèle linéaire général

$$f(y_i) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_p x_{p,i} + \varepsilon_i$$

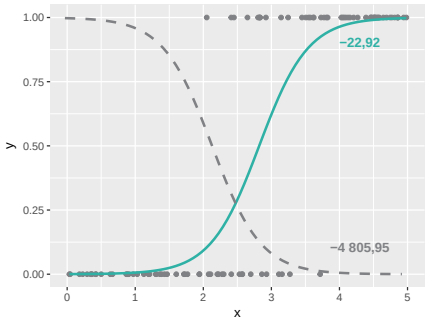
- Y la variable expliquée ;
- $X = (1\ X_1 \cdots X_p)$ la matrice de variables explicatives ;
- ε le résidu ;
- β le vecteur de coefficients à estimer ;
- $f()$ la **fonction de lien**.

Objectifs de l'estimation

1. Trouver les paramètres $\beta_0, \ldots, \beta_p$ qui maximisent l'ajustement du modèle aux données.
2. Pouvoir quantifier la qualité de cet ajustement et ses conséquences sur l'estimation en termes d'inférence.

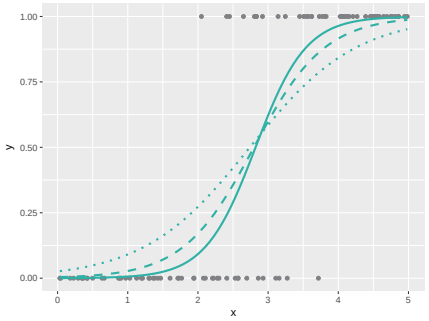
Estimation par maximum de vraisemblance

Plus la vraisemblance est élevée, meilleur est l'ajustement du modèle aux données.



Estimation par maximum de vraisemblance

L'algorithme utilisé est **itératif** : il **maximise la log-vraisemblance** ℓ_n jusqu'à atteindre un **maximum**.



Itérations 1, 2 et 5

Parenthèse : La fonction *logit*

La fonction *logit* est historiquement utilisée pour exprimer sur $\mathbb{R}$ une proportion p définie sur $]0;1[$.

1. $\dfrac{p}{1-p}$ est la **cote** associée à la proportion p (comme dans les paris hippiques). Elle est à valeurs dans $\mathbb{R}^+$.

Exemple Une probabilité de succès de 20 % correspond à une cote de 0,25 soit 1 succès pour 4 échecs. Dans les paris hippiques, on retourne le rapport et on dira « 4 contre 1 ».

2. $\mathit{logit}(p) = \ln\left(\dfrac{p}{1-p}\right)$ est donc bien à valeurs dans $\mathbb{R}$.

Estimer un modèle logistique dichotomique
Inférence

Comme en régression linéaire classique, les paramètres du modèle sont estimés avec une certaine **imprécision**.

En plus de la valeur de $\hat{\beta}$, l'algorithme produit la matrice de variance-covariance dont on extrait les **erreurs standards** des coefficients.

Pour déterminer si un coefficient β_k est statistiquement différent de 0, on pose le **test statistique** :

$$H_0 : \beta_k = 0 \quad \text{contre} \quad H_1 : \beta_k \neq 0$$

On peut alors montrer que sous H_0 :

$$z = \frac{\hat{\beta}_k}{se(\hat{\beta}_k)} \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N}(0,1)$$

avec $se(\hat{\beta}_k)$ l'**erreur-standard** de $\hat{\beta}_k$.

Estimer un modèle logistique dichotomique
Inférence

Il est dès lors possible de **tester la significativité** du coefficient β_k pour un risque de première espèce α donné (5 % ou 1 % en général) :

- en comparant la statistique de test au quantile à $1 - \alpha/2$ % d'une loi normale centrée réduite.

Rappel	90%	95%	97,5%	99%	99,5%
$q_\gamma^{\mathcal{N}(0,1)}$	1,28	1,64	1,96	2,33	2,58

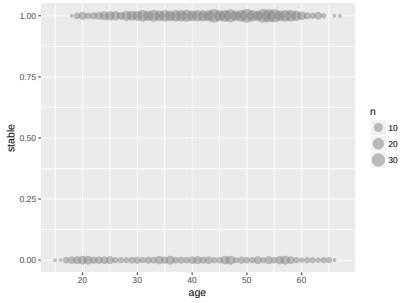
- en interprétant la **p-valeur** : on peut rejeter H_0 au seuil α si la p-valeur est inférieure à α ;
- en construisant l'intervalle de confiance au seuil $1 - \alpha$:

$$IC_{1-\alpha\ \%}(\hat{\beta}_k) = \left[\hat{\beta}_k - q_{1-\alpha/2}^{\mathcal{N}(0,1)} se(\hat{\beta}_k); \hat{\beta}_k - q_{1-\alpha/2}^{\mathcal{N}(0,1)} se(\hat{\beta}_k)\right]$$

Application : Probabilité d'être en emploi stable
Relation entre âge et stabilité du contrat

On s'intéresse tout d'abord à la relation entre **âge et stabilité du contrat** :

$$logit\left[\mathbb{P}(stable_i = 1 | age_i)\right] = \beta_0 + \beta_1 \times age_i + \varepsilon_i$$



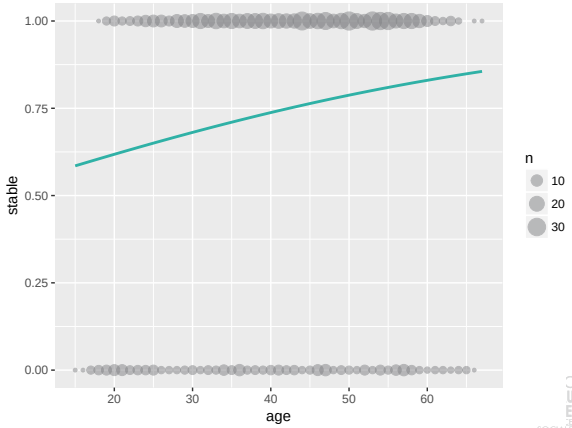
Application : Probabilité d'être en emploi stable
Relation entre âge et stabilité du contrat

```
# Régression logistique dichotomique simple
glm(
  formula = stable ~ age
  , data = e
  , family = binomial(link = "logit")
)
##
## Call:  glm(formula = stable ~ age, family = binomial(link = "logit"),
##     data = e)
##
## Coefficients:
## (Intercept)          age
##    -0.07058      0.02762
##
## Degrees of Freedom: 1007 Total (i.e. Null);  1006 Residual
## Null Deviance:       1145
## Residual Deviance: 1124  AIC: 1128

# Stockage des résultats dans l'objet  m1
m1 <- glm(formula = stable ~ age, data = e, family = binomial)
```


Application : Probabilité d'être en emploi stable
Relation entre âge et stabilité du contrat

```
# Affichage des principaux résultats
summary(m1)
##
## Call:
## glm(formula = stable ~ age, family = binomial, data = e)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9555  -1.3508   0.6911   0.7987   1.0025
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.07058    0.26021  -0.271    0.786
## age          0.02762    0.00617   4.476  7.6e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1144.5  on 1007  degrees of freedom
## Residual deviance: 1124.2  on 1006  degrees of freedom
## AIC: 1128.2
##
## Number of Fisher Scoring iterations: 4
```


Application : Probabilité d'être en emploi stable
Relation entre âge et stabilité du contrat



Application : Probabilité d'être en emploi stable
Relation entre âge et stabilité du contrat

Contrairement à ce qu'il se passe en régression linéaire, la valeur des coefficients ne peut **pas être interprétée directement**.

Ici le coefficient associé à l'âge est positif, aussi la relation entre âge et stabilité de l'emploi est **positive**.

Pour déterminer si β_1 est significativement différent de 0, on pose le **test statistique** :

$$H_0 : \beta_1 = 0 \quad \text{contre} \quad H_1 : \beta_1 \neq 0$$

La statistique de test vaut $4,48 > 1,96$ donc on peut **rejeter l'hypothèse H_0 au seuil de 5 %**.

La p-valeur est même inférieure à 0,01 donc on peut **rejeter H_0 au seuil de 1 %**.

Application : Probabilité d'être en emploi stable
Estimation du modèle complet

Formulation du modèle

$$\begin{aligned} logit\left[\mathbb{P}(stable_i = 1 | age_i)\right] = \beta_0 &+ \beta_1 age_i + \beta_2 femme_i + \beta_3 infbac_i \\ &+ \beta_4 supbac_i + \beta_5 agri_i + \beta_6 cons_i \\ &+ \beta_7 tert_i + \varepsilon_i \end{aligned}$$

```
# Dichotomisation des variables explicatives
e <- within(e, {
  homme <- (SEXE == "1") * 1
  femme <- (SEXE == "2") * 1
  supbac <- (DDIPL %in% c("1", "3")) * 1
  bac <- (DDIPL == "4") * 1
  infbac <- (DDIPL %in% c("5", "6", "7")) * 1
  agri <- (NAFG4N == "ES") * 1
  indus <- (NAFG4N == "ET") * 1
  cons <- (NAFG4N == "EU") * 1
  tert <- (NAFG4N == "EV") * 1
})
```

Application : Probabilité d'être en emploi stable
Estimation du modèle complet

```
# Estimation du modèle complet (m2)
m2 <- glm(stable ~ age + femme + infbac
          + supbac + agri + cons + tert
  , data = e , family = binomial
)

# Affichage des coefficients estimés par le modèle m2
coef(summary(m2))
##                Estimate  Std. Error    z value     Pr(>|z|)
## (Intercept)  0.33083930 0.373150080  0.8866119 3.752879e-01
## age          0.03152330 0.006558967  4.8061371 1.538744e-06
## femme        0.35261130 0.161064115  2.1892604 2.857792e-02
## infbac      -0.08431424 0.208650357 -0.4040935 6.861440e-01
## supbac       0.21288742 0.219811305  0.9685008 3.327943e-01
## agri        -3.14835955 0.473026619 -6.6557767 2.818078e-11
## cons        -0.76040610 0.350850198 -2.1673241 3.021015e-02
## tert        -0.74364553 0.259468760 -2.8660311 4.156535e-03
```


Application : Probabilité d'être en emploi stable
Estimation du modèle complet

Les coefficients d'un modèle de régression logistique multiple rendent compte d'effets « **tous les autres paramètres du modèle égaux par ailleurs** ».

Exemple

1. Le coefficient associé à la variable cons est négatif.
2. On interprète alors : « À âge, sexe et diplômes **égaux par ailleurs**, le fait de travailler dans la construction est associé à une **probabilité plus faible** d'être en emploi stable **par rapport** aux salariés de l'industrie (modalité de référence) ».

Application : Probabilité d'être en emploi stable
Estimation du modèle complet

À nouveau **la valeur des coefficients n'est pas interprétable en tant que telle**. Il est néanmoins possible d'interpréter :

- le signe des coefficients : relation positive s'ils sont positifs, négative sinon ;
- **au sein d'un même modèle**, l'amplitude relative des coefficients.

Exemple

1. En valeur absolue, le coefficient associé à la variable femme est inférieur à celui associé à la variable tert.
2. On interprète alors : « L'**effet propre** du sexe (à âge, diplôme et secteurs égaux par ailleurs) sur la probabilité d'être en emploi stable est **moindre** que celui associé au fait de travailler dans la construction **plutôt que** dans l'industrie (modalité de référence) ».

Indicateurs de qualité du modèle
Statistiques construites à partir de ℓ_n

Pour comparer deux modèles portant sur la même variable expliquée, on peut comparer les valeurs de leur vraisemblance : **on privilégie le modèle présentant la plus grande vraisemblance**.

Cependant, quand un modèle comporte davantage de variables explicatives, son pouvoir prédictif **augmente mécaniquement** (comme pour le R^2).

On peut alors utiliser des indicateurs qui **pénalisent la vraisemblance par le nombre de variables** (p) :

- *Akaike information criterion* : $AIC = -2\ell_n + 2(p+1)$
- *Bayesian information criterion* (ou critère de Schwartz) : $BIC = -2\ell_n + \ln(n)(p+1)$

Indicateurs de qualité du modèle
Exemple : Stabilité du contrat de travail

```
# Comparaison des log-vraisemblance
logLik(m1)
## 'log Lik.' -562.0967 (df=2)
logLik(m2)
## 'log Lik.' -528.0719 (df=8)

# Comparaison des AIC
AIC(m1)
## [1] 1128.193
AIC(m2)
## [1] 1072.144

# Comparaison des BIC
BIC(m1)
## [1] 1138.025
BIC(m2)
## [1] 1111.47
```


Indicateurs de qualité du modèle
Test de significativité globale

Pour évaluer le **pouvoir explicatif** d'un modèle, on peut comparer sa vraisemblance à celle du modèle ne comportant que la constante.

Il est possible de formaliser cette comparaison dans le cadre du test du **ratio de vraisemblance**.

On peut en effet montrer que sous l'hypothèse H_0 d'égalité des deux vraisemblances,

$$LR = -2\ln\left(\frac{L^0}{L_n}\right) = (-2\ell^0) - (-2\ell_n) \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2_p$$

avec ℓ^0 la log-vraisemblance du modèle ne comportant que la constante.

Indicateurs de qualité du modèle
Exemple : Stabilité du contrat de travail

```
# La "deviance" du modèle correspond à - 2*l_n
m2$deviance
## [1] 1056.144
-2*logLik(m2)
## 'log Lik.' 1056.144 (df=8)

# Calcul de la statistique de test à partir de la deviance
LR <- m2$null.deviance - m2$deviance
LR
## [1] 88.37818

# Mise en oeuvre du test avec le package lmtest
# library(lmtest)
# lrtest(m2)
```


Indicateurs de qualité du modèle
Pourcentage de concordance

Le modèle de régression permet d'obtenir, pour chaque individu de l'échantillon, une probabilité prédite $\hat{p}_i$ sur la base des variables explicatives.

On peut alors classer chaque paire d'observations selon trois catégories :

- **concordante** : $y_1 = 0,\ y_2 = 1$ et $\hat{p}_1 < \hat{p}_2$ ou $y_1 = 1,\ y_2 = 0$ et $\hat{p}_1 > \hat{p}_2$
- **discordante** : $y_1 = 0,\ y_2 = 1$ et $\hat{p}_1 > \hat{p}_2$ ou $y_1 = 1,\ y_2 = 0$ et $\hat{p}_1 < \hat{p}_2$
- **ex-aequo** : $\hat{p}_1 = \hat{p}_2$.

On peut alors calculer un **pourcentage de paires concordantes** rapporté au nombre de paires total.

Indicateurs de qualité du modèle
Exemple : Stabilité du contrat de travail

```
# Fonction de calcul du pourcentage de concordance
conc <- function(m){
  un <- m$fitted.values[m$y == 1]
  zero <- m$fitted.values[m$y == 0]
  t <-rowSums(sapply(un, function(i){
    c(sum(i > zero), sum(i < zero), sum(i == zero))
  }))
  return(c(
    "Pct concordant" = t[1] * 100 / sum(t)
    , "Pct discordant" = t[2] * 100 / sum(t)
    , "Pct ex-aequo"= t[3] * 100 / sum(t)
  ))
}

# Application de la fonction au modèle m2
conc(m2)
## Pct concordant Pct discordant   Pct ex-aequo
##     66.9649287     32.8029553      0.2321159
```


Indicateurs de qualité du modèle
Performance de la classification et courbe ROC

Bien souvent, l'objectif d'un modèle est d'aboutir à une **classification binaire**.

Exemples Le radar détecte-t-il un avion ennemi ? Le message reçu est-il un *spam* ?

Mais en sortie du modèle, on obtient pour chaque individu la probabilité $\hat{p}_i$, et non une valeur 0 ou 1.

Question **Où placer la probabilité seuil** p^* entre les cas à classer comme positifs ($\hat{p}_i > p^*$) et les cas à classer comme négatifs ($\hat{p}_i < p^*$) ?

Indicateurs de qualité du modèle
Performance de la classification et courbe ROC

Le meilleur modèle serait celui qui ne conduirait à **aucun faux négatif et aucun faux positif**.

Mais on a en fait affaire à un arbitrage :

- **Si le seuil est trop haut**, certains individus positifs risquent d'être classés comme négatifs (faux négatifs).
- **Si le seuil est trop bas**, certains individus négatifs risquent d'être classés comme positifs (faux positifs).

Moralité Afin de limiter le risque de faux négatifs on est amené à tolérer un certain nombre de faux positifs, et inversement.

La courbe ROC (*Receiver operating characteristics*) représente cet arbitrage.

Indicateurs de qualité du modèle
Construction de la courbe ROC

1. Estimer le modèle et classer les observations par **probabilités prédites** $\hat{p}_i$ **croissantes** ;
2. Pour chaque observation i :
   - calculer la part des **positifs classés positifs** (**sensibilité**) si $\hat{p}_i$ constitue le seuil entre positif et négatif ;
   - calculer la part des **négatifs classés négatifs** (**spécificité**) si $\hat{p}_i$ constitue le seuil entre positif et négatif ;
3. La courbe ROC est la représentation de la **sensibilité en fonction de la spécificité** (axe inversé).

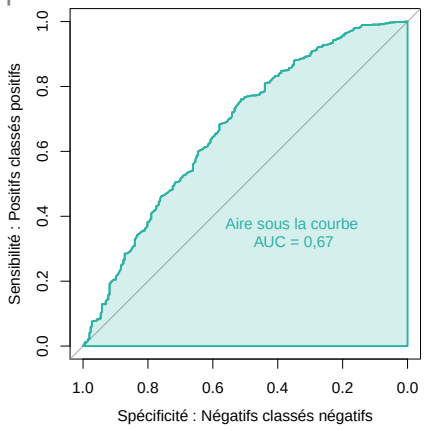
L'**aire sous la courbe** (*Area under the curve* ou AUC) est un **indicateur synthétique de la performance** de classification du modèle.

Indicateurs de qualité du modèle
Exemple : Stabilité du contrat de travail

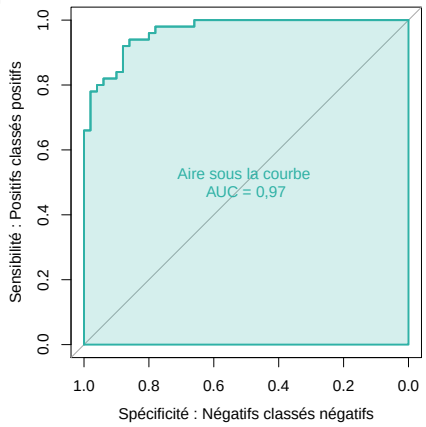
```
# Calcul de la courbe ROC avec le package pROC
library(pROC)
m2_roc <- roc(m2$y ~ m2$fitted.values)
m2_roc
##
## Call:
## roc.formula(formula = m2$y ~ m2$fitted.values)
##
## Data: m2$fitted.values in 257 controls (m2$y 0) < 751 cases (m2$y 1).
## Area under the curve: 0.6708
```

```
# Représentation avec la fonction plot()
par(pty="s") # Pour avoir une zone de tracé carrée
plot(m2_roc
 , xlab = "Spécificité : Négatifs classés négatifs"
 , ylab = "Sensibilité : Positifs classés positifs"
)
```


Indicateurs de qualité du modèle
Exemple : Stabilité du contrat de travail



Indicateurs de qualité du modèle
Comparaison : Données simulées



Odds-ratio et effets marginaux
Définition de l'*odds-ratio*

On rappelle que la « **cote** » (ou *odd*) d'une proportion p est le rapport

$$odd_p = \frac{p}{1-p}$$

Exemple Pour une proportion de 25 %, la cote est de 1/3 (ou 3 contre 1 dans les paris hippiques).

On appelle alors « **rapport des cotes** » (ou *odds-ratio*) des proportions p et q :

$$OR_{p|q} = \frac{\frac{p}{1-p}}{\frac{q}{1-q}}$$

Interprétation Si $p > q$ alors $OR_{p|q} > 1$.

Les *odds-ratio* dans une régression logistique

Mathématiquement, les *odds-ratio* d'un modèle de régression logistique correspondent à l'exponentielle de la valeur des coefficients :

$$OR_{k|ref} = e^{\beta_k} = exp(\beta_k)$$

```
cbind(coef=coef(m2), or = exp(coef(m2)))
##                    coef         or
## (Intercept)  0.33083930 1.39213605
## age          0.03152330 1.03202542
## femme        0.35261130 1.42277800
## infbac      -0.08431424 0.91914238
## supbac       0.21288742 1.23724535
## agri        -3.14835955 0.04292248
## cons        -0.76040610 0.46747655
## tert        -0.74364553 0.47537775
```

Remarque Quand le coefficient est positif, l'*odds-ratio* est supérieur à 1 et inversement.

Interprétation courante des *odds-ratio*

En règle générale, on interprète l'*odds-ratio* associé à une modalité d'une variable qualitative comme le **rapport des chances** pour les individus présentant cette modalité d'être dans la situation modélisée **par rapport aux individus présentant la modalité de référence**.

Exemple L'*odds-ratio* associé au fait d'être une femme est de 1,423 : à âge, diplôme et secteur égaux par ailleurs, les femmes ont **1,423 fois plus de chances** d'être en contrat stables que les hommes.

Néanmoins, **cette interprétation très courante assimile *odds-ratio* et risque relatif**, ce qui pose problème quand l'*odds-ratio* est proche de 1.

Inférence à partir des *odds-ratio*

Pour réaliser une inférence à partir des *odds-ratio*, il suffit de **recalculer les bornes de l'intervalle de confiance au niveau souhaité**.

```
cbind(OR = exp(coef(m2)), exp(confint.default(m2)))
##                     OR      2.5 %    97.5 %
## (Intercept) 1.39213605 0.66997062 2.8927280
## age         1.03202542 1.01884329 1.0453781
## femme       1.42277800 1.03762187 1.9509007
## infbac      0.91914238 0.61063161 1.3835227
## supbac      1.23724535 0.80417785 1.9035292
## agri        0.04292248 0.01698415 0.1084740
## cons        0.46747655 0.23502583 0.9298311
## tert        0.47537775 0.28587662 0.7904949
```


Odds-ratio et risque relatif

Le terme « **risque relatif** » des proportions p et q désigne le rapport : $RR_{p|q} = \frac{p}{q}$.

Les confusions entre risque relatif et *odds-ratio* sont fréquentes.

Si pour les proportions rares les deux quantités sont proches, pour les proportions fréquentes ce n'est pas du tout le cas.

Exemple $p = 0{,}70$, $q = 0{,}40$

- $RR_{p|q} = \frac{0{,}70}{0{,}40} = 1{,}75$
- $OR_{p|q} = \frac{0{,}70/0{,}30}{0{,}40/0{,}60} = 3{,}5$

Pour aller plus loin http://logisticregression.su.se/ .

Définition de l'effet marginal

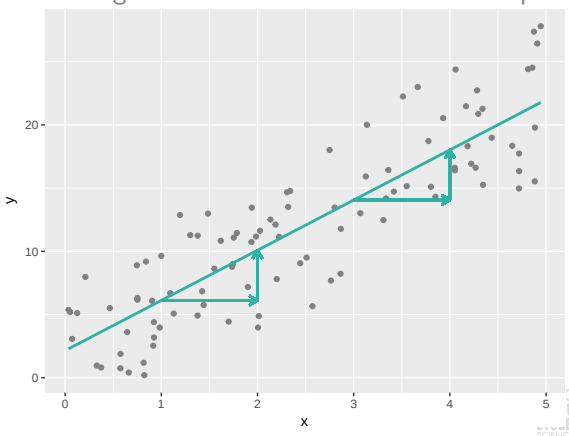
Dans un modèle logistique dichotomique, l'**effet marginal** est un moyen simple pour réexprimer la relation entre une variable explicative et la variable d'intérêt en termes de **points de pourcentages**.

Exemple En moyenne dans l'échantillon et à âge, sexe et diplôme égaux par ailleurs, le fait de travailler dans la construction plutôt que dans l'industrie est associé à une probabilité inférieure d'avoir un emploi stable de l'ordre de 15 points de pourcentage.

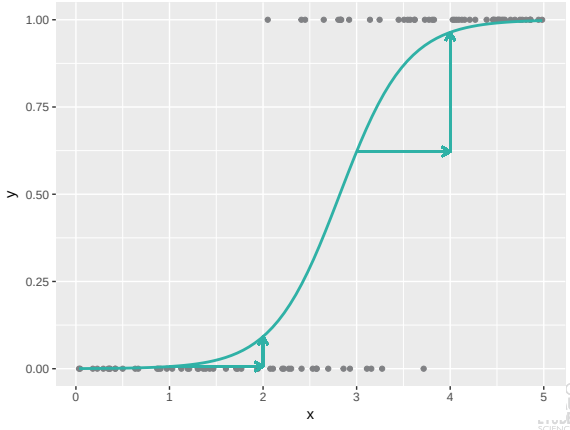
Dans le **modèle linéaire classique**, l'effet marginal de la variable x_j sur Y est tout simplement $\hat{\beta}_j$.

Exemple Ainsi dans $salaire_i = \beta_0 + \beta_1 age_i + \varepsilon_i$ l'effet marginal de la variable *age* est **constant et égal à** $\hat{\beta}_1$.

Effet marginal dans un modèle linéaire classique



Effet marginal dans un modèle logistique



Effet marginal dans un modèle logistique

Dans un modèle de régression logistique dichotomique, l'effet marginal de la variable x_j sur Y peut **varier d'un individu à l'autre**.

Quand la variable x_j est **dichotomique**, le calcul de l'effet marginal de la variable x_j pour l'individu i $\delta_i(x_j)$ est effectué de la façon suivante :

1. on calcule la probabilité de i prédite par le modèle $\hat{p}_{i|x_j=1}$ si x_j **était égale à 1** ;
2. on calcule la probabilité de i prédite par le modèle $\hat{p}_{i|x_j=0}$ si x_j **était égale à 0** ;
3. on calcule l'effet marginal avec :

$$\delta_i(x_j) = \hat{p}_{i|x_j=1} - \hat{p}_{i|x_j=0}$$

Effet marginal dans un modèle logistique

Exemple Dans le modèle

$$\mathbb{P}(stable_i = 1 | age_i, femme_i)) = \beta_0 + \beta_1 age_i + \beta_2 femme_i + \varepsilon_i$$

on calcule l'effet marginal du sexe sur la stabilité de l'emploi pour un individu i $\delta_i(femme)$ de la façon suivante :

1. on calcule $\hat{p}_{i|femme=1} = logit^{-1}(\hat{\beta}_0 + \hat{\beta}_1 age_i + \hat{\beta}_2)$;
2. on calcule $\hat{p}_{i|femme=0} = logit^{-1}(\hat{\beta}_0 + \hat{\beta}_1 age_i)$;
3. l'effet marginal est alors

$$\delta_i(femme) = \hat{p}_{i|femme=1} - \hat{p}_{i|femme=0}$$

Si la relation entre stabilité de l'emploi et le fait d'être une femme est **positive** ($\hat{\beta}_2 > 0$), $\delta_i(femme) > 0$, et inversement.

Exemple : Stabilité du contrat de travail

```
# Fonction de calcul margins()
margins(m2)[1:3]
## $femme
##            Average MFX Std. Error  z value      P>|z|
## femme = 0  0.71396457         NA       NA         NA
## femme = 1  0.77498811         NA       NA         NA
## Diff       0.06102354 0.02789538 2.187586 0.02869976
##
## $infbac
##             Average MFX Std. Error    z value     P>|z|
## infbac = 0  0.75188391         NA         NA        NA
## infbac = 1  0.73736314         NA         NA        NA
## Diff       -0.01452077 0.03599009 -0.4034658 0.6866056
##
## $supbac
##             Average MFX Std. Error   z value     P>|z|
## supbac = 0  0.73252143         NA        NA        NA
## supbac = 1  0.76868772         NA        NA        NA
## Diff        0.03616628 0.03683972 0.9817197 0.3262379
```


Motivation et exemples

Il est fréquent que certaines hypothèses ne puissent pas être testées à l'aide d'un test sur **un seul paramètre** :

- Quelle est la relation entre le niveau de diplôme pris dans son ensemble et stabilité du contrat ?
- Quelle est la relation entre le secteur d'activité pris dans son ensemble et la stabilité du contrat ?

Dans le **cas du diplôme** par exemple, on pose ce test de la façon suivante :

$$H_0 : \beta_3 = \beta_4 = 0 \quad \text{contre} \quad H_1 : \beta_3 \neq 0 \text{ ou } \beta_4 \neq 0$$

Rappel du modèle

$$stable_i = \beta_0 + \beta_1 age_i + \beta_2 femme_i + \beta_3 infbac_i + \beta_4 supbac_i + \beta_5 agri_i + \beta_6 cons_i + \beta_7 tert_i + \varepsilon_i$$

Statistique de test

On peut montrer que sous H_0 :

$$LR = -2 \ln \left(\frac{L_c}{L_{nc}} \right) = (-2\ell_c) - (-2\ell_{nc}) \xrightarrow[n \to +\infty]{\mathcal{L}} \chi^2_q$$

où ℓ_{nc} est la log-vraisemblance du modèle non-contraint, ℓ_c la log-vraisemblance du modèle contraint et q le nombre de restrictions.

Dans le **cas du diplôme** :

1. ℓ_{nc} = -528,07 et ℓ_c = -529,55 et ainsi LR = 2,95 ;
2. La p-valeur du test du ratio de vraisemblance vaut $1 - F^{\chi^2_2}(LR) = 0{,}2291$;
3. On ne peut pas rejeter l'hypothèse nulle au seuil de 5 %.

Effet marginal moyen

L'effet marginal moyen est directement calculé comme la moyenne sur l'échantillon des effets marginaux individuels :

$$\bar{\delta}(x_j) = \frac{1}{n} \sum_{i=1}^{n} \delta_i(x_j) = \frac{1}{n} \sum_{i=1}^{n} \hat{p}_{i|x_j=1} - \frac{1}{n} \sum_{i=1}^{n} \hat{p}_{i|x_j=0}$$
$$= \bar{p}_{|x_j=1} - \bar{p}_{|x_j=0}$$

Interprétation L'effet marginal moyen correspond à l'**augmentation moyenne dans l'échantillon** de la probabilité $\mathbb{P}(Y = 1)$ quand x_j passe de 0 à 1.

Exemple Si dans le modèle de la diapositive précédente $\bar{\delta}(femme) = 0{,}10$, on dira qu'en moyenne dans l'échantillon et à âge égal par ailleurs, le fait d'être une femme est associé à une probabilité d'être en contrat stable supérieure **de 10 points de pourcentage**.

Intérêt de l'effet marginal moyen

L'effet marginal moyen présente plusieurs avantages :

1. Il s'exprime en **termes de probabilités**, ce qui le rend extrêmement intuitif et facile à utiliser.
2. Des **erreurs standards** peuvent être obtenues pour l'effet marginal moyen, ce qui permet de juger de la significativité de l'écart en termes de probabilité.
3. La comparaison d'effets marginaux moyens entre plusieurs modèles emboîtés semble **plus robuste** que la comparaison des *odds-ratio* à l'hétérogénéité inobservée.

Pour aller plus loin Mood C. (2010)
https://doi.org/10.1093/esr/jcp006

Test du ratio de vraisemblance

Principe Comparer la log-vraisemblance de **deux modèles emboîtés** :

- d'une part le modèle **complet** ou **non-contraint** qui comporte tous les paramètres ;
- d'autre part le modèle **contraint** qui correspond au cas où l'hypothèse nulle est vérifiée.

Dans le **cas du diplôme**, le modèle contraint est :

$$stable_i = \beta_0 + \beta_1 age_i + \beta_2 femme_i + \beta_5 agri_i + \beta_6 cons_i + \beta_7 tert_i + \varepsilon_i$$

Intuition Si le modèle non-contraint est **beaucoup plus vraisemblable** que le modèle contraint, alors on a tendance à **rejeter la contrainte**, c'est-à-dire l'hypothèse nulle.

Exemple : Stabilité du contrat de travail

```
# On teste la significativité jointe des coefficients
# associés au secteur d'activité de l'entreprise
m4 <- glm(stable ~ age + femme + infbac + supbac
  , data = e, family = binomial(link = "logit")
)

# Test du ratio de vraisemblance avec le package lmtest
library(lmtest)
lrtest(m2, m4)
## Likelihood ratio test
##
## Model 1: stable ~ age + femme + infbac + supbac + agri + cons + tert
## Model 2: stable ~ age + femme + infbac + supbac
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   8 -528.07
## 2   5 -556.10 -3 56.065  4.069e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# Le secteur d'activité semble très significatif.
```

En guise de conclusion

Le modèle de régression logistique est adapté pour modéliser des **données dichotomiques**.

Comme la plupart des spécifications du modèle linéaire général, son estimation est effectuée par **maximum de vraisemblance** et avec la fonction `glm()` de **R**.

Les **indicateurs de qualité** diffèrent sensiblement de ceux utilisés en régression linéaire classique : *AIC*, *BIC*, pourcentage de concordance, courbe ROC.

L'**interprétation des coefficients** est plus difficile qu'en régression linéaire : utilisation des *odds-ratio* et calculs d'effets marginaux moyens.

Le modèle probit dichotomique

C'est la **fonction de lien** f qui différencie le modèle probit dichotomique du modèle logistique dichotomique.

Dans le modèle probit dichotomique, f est telle que

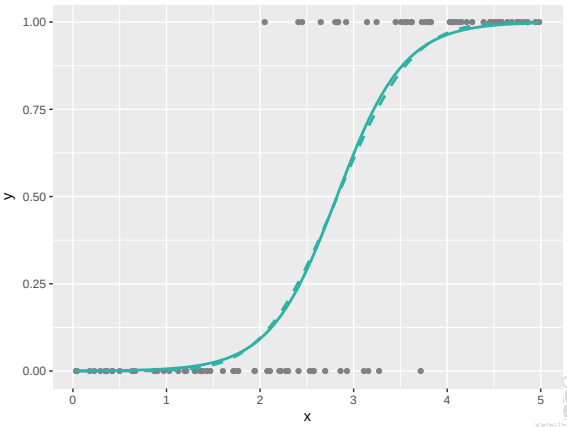
$$p_i = f^{-1}(X\beta) = \Phi(X\beta)$$

où $\Phi(x)$ est la **fonction de répartition de la loi normale centrée réduite**.

Ses coefficients diffèrent mais qualitativement **ses résultats sont proches** de ceux d'un modèle logistique dichotomique.

```
probit <- glm(z ~ x, data = sim_dicho
  , family = binomial(link = "probit")
)
summary(probit)
```


Le modèle probit dichotomique



Retour au modèle linéaire classique

Le modèle linéaire classique peut être vu comme un cas particulier de modèle linéaire général :

1. de la **famille gaussienne** ;
2. avec la **fonction de lien identité**.

```
# Avec la fonction lm()
lin_lm <- lm(y ~ x, data = sim_lin)

# Avec la fonction glm()
lin_glm <- glm(y ~ x, data = sim_lin
  , family = gaussian(link = "identity")
)

# Comparaison des coefficients
identical(lin_lm$coefficients, lin_glm$coefficients)
## [1] TRUE
```


Vraisemblance du modèle logistique dichotomique

L'objectif de cette annexe est de déterminer l'**expression de la (log-)vraisemblance** dans le cas d'une régression logistique dichotomique.

Au-delà de son contenu théorique, elle doit permettre de **mieux comprendre les paramètres** à indiquer au logiciel pour effectuer l'estimation.

En toute généralité, la vraisemblance d'une variable Y sachant les observations X est définie par

$$L_n = \mathbb{P}(y_1, \ldots, y_n | X_1, \ldots, X_n)$$

Il s'agit de la **probabilité d'observer les valeurs** $(y_1, \ldots, y_n)$ **sachant les valeurs** $(X_1, \ldots, X_n)$.

Vraisemblance du modèle logistique dichotomique

Sous les hypothèses que les observations sont indépendantes les unes des autres et qu'elles suivent une même distribution, L_n devient :

$$L_n = \mathbb{P}(y_1|X_1) \times \ldots \times \mathbb{P}(y_n|X_n) = \prod_{i=1}^{n} \mathbb{P}(y_i|X_i)$$

Pour faciliter les manipulations et le fonctionnement des algorithmes, on travaille en général sur la **log-vraisemblance** ℓ_n :

$$\ell_n = \ln(L_n) = \ln\left[\prod_{i=1}^{n} \mathbb{P}(y_i|X_i)\right] = \sum_{i=1}^{n} \ln\left[\mathbb{P}(y_i|X_i)\right]$$

Pour déterminer l'expression de la vraisemblance dans le cas d'un modèle logistique dichotomique, on réexprime $\mathbb{P}(y_i|X_i)$.

Vraisemblance du modèle logistique dichotomique

Dans le cas d'un modèle dichotomique, Y ne peut prendre que deux valeurs (0 ou 1), aussi :

$$\mathbb{P}(y_i|X_i) = \mathbb{P}(y_i = 1|X_i)^{y_i} \times \mathbb{P}(y_i = 0|X_i)^{1-y_i}$$

On dit que les modèles dichotomiques correspondent à la **famille binomiale** de modèles linéaires généraux.

$p_i = \mathbb{P}(y_i = 1|X_i)$ représente la **probabilité de succès**.

Comme

$$\mathbb{P}(y_i = 0|X_i) = 1 - \mathbb{P}(y_i = 1|X_i) = 1 - p_i$$

on peut réécrire :

$$\mathbb{P}(y_i|X_i) = p_i^{y_i} \times (1 - p_i)^{1-y_i}$$

Vraisemblance du modèle logistique dichotomique

La **log-vraisemblance d'un modèle dichotomique** s'écrit ainsi :

$$\begin{aligned}\ell_n &= \sum_{i=1}^{n} \ln\left[\mathbb{P}(y_i|X_i)\right] \\ &= \sum_{i=1}^{n} \ln\left[p_i^{y_i} \times (1 - p_i)^{1-y_i}\right] \\ &= \sum_{i=1}^{n} \left[y_i \ln(p_i) + (1 - y_i) \ln(1 - p_i)\right]\end{aligned}$$

où p_i est la probabilité de $y_i = 1$ sachant les variables explicatives X_i.

C'est la manière de **relier** p_i aux variables explicatives X_i qui distingue les différents modèles de régression pour variable dichotomique.

Vraisemblance du modèle logistique dichotomique

Le modèle logistique dichotomique est le modèle tel que :

$$logit(p_i) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_p x_{p,i}$$

Ainsi

$$p_i = logit^{-1}(X_i\beta) = \frac{e^{X_i\beta}}{1 + e^{X_i\beta}}$$

et donc

$$\begin{aligned}\ell_n &= \sum_{i=1}^{n} [y_i \ln(p_i) + (1 - y_i) \ln(1 - p_i)] \\ &= \sum_{i=1}^{n} \left[y_i \ln\left(\frac{e^{X_i\beta}}{1 + e^{X_i\beta}}\right) + (1 - y_i) \ln\left(1 - \frac{e^{X_i\beta}}{1 + e^{X_i\beta}}\right)\right]\end{aligned}$$

L'estimateur du maximum de vraisemblance $\hat{\beta}$ est obtenu en maximisant la quantité ℓ_n.

Aspects transversaux et approfondissements

Martin Chevalier (Insee)

Année universitaire 2017-2018

Plan de la séance

Hypothèse d'homoscédasticité

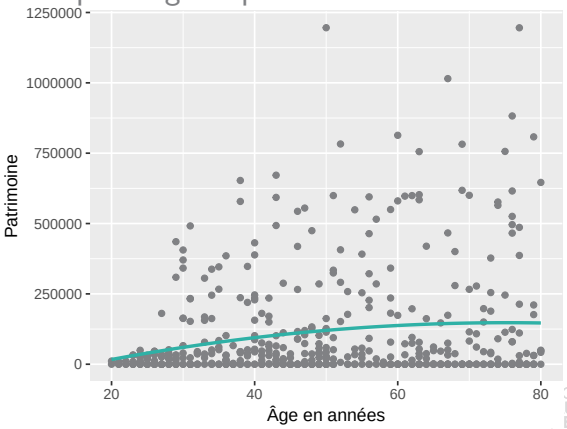
Le modèle linéaire classique comme le modèle linéaire général (régression logistique, probit) reposent sur l'hypothèse d'**homoscédasticité**.

La variance du terme d'erreur est supposée identique dans l'ensemble de l'échantillon :

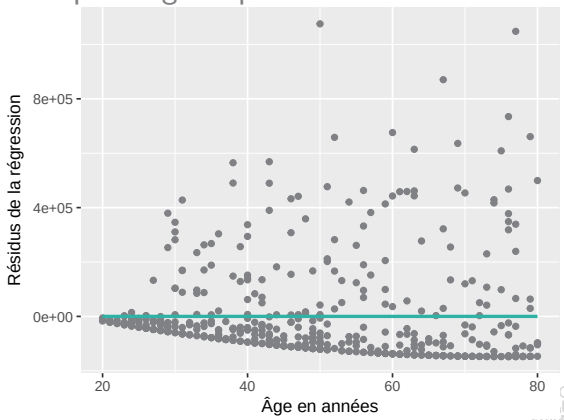
$$\forall i \quad V(\varepsilon_i) = \sigma^2 = \text{constante}$$

Les erreurs standards étant calculées sous cette hypothèse d'homoscédasticité, quand elle n'est pas respectée **les tests sur les paramètres peuvent être faussés**.

Exemple : Âge et patrimoine



Exemple : Âge et patrimoine



Tests d'hétéroscédasticité

Plusieurs tests existent pour évaluer le caractère plus ou moins hétéroscédastique des résidus d'une régression : test de Breusch et Pagan, test de White.

Pour les mettre en œuvre, on régresse le carré des résidus estimés $\hat{\varepsilon}^2$ sur une combinaison des variables explicatives du modèle.

Plus cette régression est explicative, plus on est fondé à penser que les résidus sont hétéroscédastiques (car bien corrélés aux variables explicatives du modèle).

On peut ainsi construire un test à partir du R^2 de cette régression.

Moindres carrés généralisés et quasi-généralisés

Une première méthode pour prendre en compte l'hétéroscédasticité est d'utiliser des modèles qui relâchent l'hypothèse d'homoscédasticité : les moindres carrés généralisés (MCG).

Ces modèles permettent de spécifier manuellement une matrice Ω de variance des résidus :

- la variance des résidus peut ainsi différer d'un individu à l'autre ;
- des covariances non-nulles entre individus peuvent être introduites (individus d'un même ménage, salariés d'une même entreprise, etc.).

En pratique, Ω est elle-même estimée : on parle alors de moindres carrés quasi-généralisés (MCQG).

Erreurs standards robustes

Une seconde méthode pour prendre en compte l'hétéroscédasticité est d'utiliser une estimation dite « robuste » des erreurs standards.

Son principe est de faire intervenir la véritable distribution des résidus dans l'estimation de la matrice de variance-covariance du modèle.

On aboutit ainsi à une estimation dite « sandwich » de la matrice de variance-covariance, qui conduit à des erreurs standards en général (mais pas toujours) plus larges que les erreurs standards classiques.

Exemple : Âge et patrimoine

$$patri_i = \beta_0 + \beta_1 \times age_i + \beta_2 \times age_i^2 + \varepsilon_i$$

```
# Estimation du modèle m2
m2 <- lm(pat ~ age + I(age^2))

# Coefficient estimés
coef(summary(m2))
##                Estimate  Std. Error   t value   Pr(>|t|)
## (Intercept) -94956.4960 78284.70409 -1.212964 0.22571994
## age           6447.6972  3315.33867  1.944808 0.05236177
## I(age^2)       -42.8501    32.30654 -1.326360 0.18532971
```


Exemple : Âge et patrimoine

```
# Fonction robust()
robust <- function(model, cluster = NULL){
  if(!require(multiwayvcov) | !require(lmtest))
    stop("Les packages multiwayvcov et lmtest sont requis.")
  if(is.null(cluster)) cluster <- 1:length(model$residuals)
  return(coeftest(model, cluster.vcov(model, cluster)))
}

# Erreurs standards robustes
robust(m2)
##
## t test of coefficients:
##
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -94956.496  63529.086 -1.4947   0.1356
## age           6447.697   3081.342  2.0925   0.0369 *
## I(age^2)       -42.850     32.714 -1.3098   0.1909
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


La question de la causalité

Il est difficile d'obtenir des résultats proprement causaux à partir de l'analyse statistique seule : cependant la recherche de la causalité est au cœur de la démarche économétrique.

La principale difficulté dans la mise en évidence de relations causales tient au phénomène d'endogénéité : en présence d'endogénéité, tous les coefficients (dont celui de la ou les variables d'intérêt) sont biaisés.

Les deux principales sources d'endogénéité que l'on rencontre en pratique sont le **biais de variable omise** et la **causalité inverse**. L'endogénéité peut également provenir d'erreurs de mesure sur la variable expliquée ou la variable d'intérêt.

Biais de variable omise

Quand une variable importante n'est pas intégrée dans le modèle, tous les coefficients sont susceptibles d'être biaisés.

Ce phénomène se comprend assez bien quand on observe l'**évolution des coefficients lors de l'introduction de nouvelles variables** (modèles imbriqués).

Exemple Introduire le temps partiel modifie la relation entre sexe et salaire : temps partiel et salaire d'une part et temps partiel et sexe d'autre part sont liés.

Le biais de variable omise est le signe des **effets de structure** qui relient les différentes variables explicatives.

Variable omise et biais de sélection

En règle générale, une variable est omise dans un modèle car elle n'est pas directement « observable » par l'enquête.

Exemple Les compétences relationnelles d'un salarié, la motivation d'un élève, etc.

Quand on souhaite évaluer l'efficacité d'un dispositif en comparant les résultats les individus qui y participent et ceux qui n'y participent pas, la **propension à participer au dispositif** est une variable omise souvent cruciale.

Omettre cette variable du modèle conduit à un « **biais de sélection** ».

Causalité inverse

Dans de nombreuses situations, les relations entre variable expliquée et variable d'intérêt ne sont pas univoques : **la causalité peut aller dans les deux sens**.

Exemple Relation entre mauvais état de santé et activité sur le marché du travail :

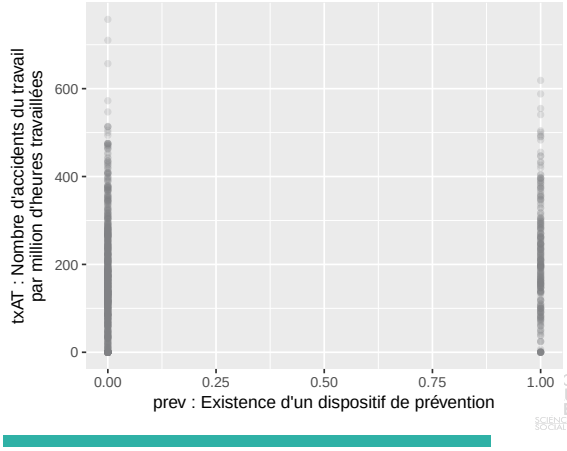
- être en mauvaise santé affecte la participation au marché du travail (recherche d'emploi plus difficile, incapacité, etc.) ;
- mais l'absence de participation au marché du travail peut avoir un impact très négatif sur la santé (conditions de vie, isolement, etc.).

Exemple : Accidents du travail et prévention

On dispose d'un échantillon de 1 000 établissements pour lesquels le taux d'accidents du travail en 2016 est connu (nombre d'accidents du travail par million d'heures travaillées, variable `txAT`).

On cherche à analyser les liens entre la survenue d'accidents du travail et les caractéristiques des établissements, en particulier la mise en place ou non de dispositifs de prévention des risques professionnels (variable `prev`).

Exemple : Accidents du travail et prévention



Exemple : Accidents du travail et prévention

On dispose par ailleurs d'informations sur :

- le secteur d'activité de l'établissement : la variable `cons` indique si l'établissement appartient au secteur de la construction ou non ;
- la taille de l'établissement : la variable `eff` correspond au nombre de salariés de l'établissement ;
- la représentation des salariés : la variable `rp` indique si des représentants du personnel sont présents dans l'établissement ;
- le pourcentage de salarié syndiqués : la variable `pctsy` correspond au pourcentage de salariés de l'établissement qui sont syndiqués (entre 0 et 100).

Exemple : Accidents du travail et prévention

$$txAT_i = \beta_0 + \beta_1 prev_i + \beta_2 cons_i + \beta_3 eff_i + \varepsilon_i$$

```
m3 <- lm(txAT ~ prev + cons + eff)
coef(summary(m3))
##                Estimate Std. Error   t value      Pr(>|t|)
## (Intercept) 161.78881990 5.53076418 29.252525 2.828794e-136
## prev         38.00128753 7.20440493  5.274730  1.630944e-07
## cons        187.26145996 7.89108887 23.730750  4.991724e-99
## eff          -0.08412927 0.01870076 -4.498709  7.641753e-06
```

L'estimation du coefficient associé à la variable `prev` est positif : il y a une **association positive** entre existence de dispositif de prévention des risques et accidents du travail.

On est manifestement dans un cas de **causalité inverse**.

Pourquoi ajouter des variables ne résout pas le problème

En règle générale, ajouter des variables supplémentaires ne peut pas résoudre un problème d'endogénéité :

- la « variable omise » qu'il faudrait ajouter n'est pas présente dans le fichier ;
- rajouter des variables supplémentaires ne peut pas faire disparaître la relation causale ambiguë entre variable dépendante et variable d'intérêt en cas de causalité inverse.

Plusieurs solutions existent cependant, qui toutes reposent sur la construction d'un « **contrefactuel** ».

Utiliser des données de panel

En cas de causalité inverse notamment, l'approche la plus naturelle est de recourir à des données de panel.

Exemple Pour mesurer l'impact d'un régime sur la perte de poids, il paraît judicieux de comparer l'**évolution** du poids selon que l'on a fait un régime ou non.

Deux difficultés majeures pour cette approche :

- les données *rétrospectives* sont affectés par les problèmes de mémoire ou certains effets de désirabilité ;
- les panels *prospectifs* sont complexes et coûteux à mettre en place.

Exploiter des « expériences naturelles »

Dans certaines situations, des évolutions du contexte produisent des « **expériences naturelles** ».

C'est tout particulièrement le cas des **changements de législation** ou encore des **discontinuités** dans l'application de certaines dispositions légales.

L'idée est d'utiliser ces discontinuités pour construire le « contrefactuel ».

Toute cette démarche repose sur l'hypothèse qu'en l'absence de discontinuité, les différents groupes auraient connu des évolutions similaires.

Exemple : Efficacité de la loi SRU

L'article 55 de la loi SRU (Solidarité et renouvellement urbain) prévoit que les communes de plus de 3 500 habitants (1 500 en Île-de-France) disposent d'au moins 20 % de logements sociaux (25 % depuis 2013), sous peine d'amendes.

Pour mesurer l'efficacité de ce dispositif, on peut exploiter **deux discontinuités** :

1. le seuil de taille : les communes proches du seuil sont comparables, mais celles juste au-dessus sont soumises au dispositifs et les autres pas ;
2. la rupture en 2013 : si le dispositif est efficace, on devrait constater une augmentation du nombre de logements sociaux après 2013 dans les communes comportant entre 20 % et de 25 % de logements sociaux.

Mettre au point une « expérience contrôlée »

La mise en place d'une « expérience contrôlée » permet de neutraliser les biais de sélection associés au dispositif dont on souhaite évaluer l'efficacité.

Exemple Les volontaires pour participer à un programme d'accompagnement renforcé à la recherche d'emploi peuvent différer sensiblement de l'ensemble des demandeurs d'emploi.

Inspirées de la recherche médicale, les expériences contrôlées opèrent une **sélection aléatoire des bénéficiaires** du programme à évaluer (on parle de « traités ») : on neutralise ce faisant le biais de sélection.

Cette méthodologie induit cependant un **coût élevé** et pose d'importantes **questions éthiques**.

Utiliser des variables instrumentales

L'utilisation de variables instrumentales est un **prolongement très direct de la logique expérimentale**.

Dans le cadre expérimental, l'assignation au groupe des « traités » est totalement aléatoire. De ce fait, la différence sur la variable d'intérêt entre « traités » et « non-traités » peut être interprétée comme l'effet du seul traitement.

En l'absence d'assignation aléatoire, on recherche **une variable qui ait les mêmes propriétés**, à savoir :

- la plus corrélée possible au fait d'avoir reçu le traitement ;
- *a priori* totalement indépendante de la variable de résultat (sinon par le biais du traitement).

Exemple (tiré de Behagel L. (2006), *Lire l'économétrie*, coll. Repères, La Découverte, 128 p.)

À partir de données américaines, Angrist et Krueger cherchent à mesurer le « rendement de l'éducation », c'est-à-dire le gain salarial associé à une année d'étude supplémentaire.

Le principal problème est qu'**on ne peut pas mesurer le talent ou la motivation des individus**, qui peuvent affecter à la fois la durée d'étude et le niveau de rémunération.

Si on mène une régression naïve, on risquerait de **surestimer** la valeur du coefficient associé au nombre d'années d'étude.

Pour neutraliser ce phénomène, ils utilisent le **trimestre de naissance** comme variable instrumentale.

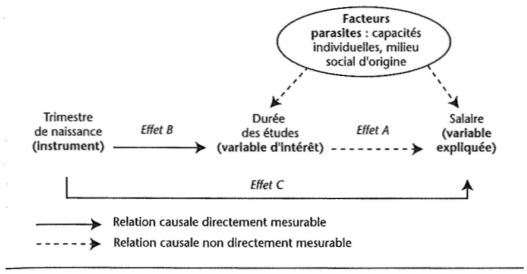
D'une part, **trimestre de naissance et durée des études sont liés** :

- selon le mois de naissance, la scolarité obligatoire commence entre 2 ans et 8 mois et 3 ans et 8 mois ;
- dans tous les cas, la scolarité obligatoire s'achève au seizième anniversaire.

D'autre part, on n'a **aucune raison de penser que le trimestre de naissance affecte directement le niveau de rémunération**.

Le seul canal par lequel ces deux phénomènes sont éventuellement liés est donc la **durée effective de scolarisation**.

Figure IV. **Schématisation du raisonnement de l'estimation par variables instrumentales**



Contrairement aux expériences contrôlées ou naturelle, la liaison entre l'instrument et le « traitement » n'est pas parfaite.

D'où une procédure en deux étapes :

1. Régression de la variable d'intérêt (la durée des études) sur la variable instrumentale (trimestre de naissance) + les variables de contrôle.
2. Régression de la variable expliquée sur la prédiction de la variable d'intérêt par le modèle précédent + les variables de contrôle.

Si la variable instrumentale était totalement assimilable à une assignation aléatoire, elle expliquerait parfaitement la variable d'intérêt qui coïnciderait avec sa prédiction : le premier modèle ne serait alors pas nécessaire (expérimentation aléatoire).

Dans ce cadre « à deux étapes », **les erreurs standards estimées sont néanmoins fausses**.

Le second modèle ne prend en effet pas en compte le fait qu'une de ses variables est en réalité la prédiction d'un autre modèle.

Pour obtenir des erreurs standards et des tests corrects, il faut impérativement mettre en œuvre cette méthode par le biais de **fonctions adaptées**.

On choisit comme variables instrumentales la **présence de représentants du personnel** et le **taux de syndicalisation** :

- l'une et l'autre sont fortement liés à l'existence de dispositifs de prévention des accidents du travail ;
- ils ne sont *a priori* pas liés au taux d'accidents du travail dans l'établissement.

On estime donc deux modèles :

1. $prev_i = \alpha_0 + \alpha_1 rp_i + \alpha_2 pctsy_i + \alpha_3 cons_i + \alpha_4 eff_i + \mu_i$
2. $txAT_i = \beta_0 + \beta_1 \widehat{prev}_i + \beta_2 cons_i + \beta_3 eff_i + \varepsilon_i$

$$prev_i = \alpha_0 + \alpha_1 rp_i + \alpha_2 pctsy_i + \alpha_3 cons_i + \alpha_4 eff_i + \mu_i$$

```
# Régression de première étape
m4 <- lm(prev ~ rp + pctsy + cons + eff)
coef(summary(m4))
##                 Estimate   Std. Error     t value      Pr(>|t|)
## (Intercept) -0.2611916396 0.0227386865 -11.4866634  9.143870e-29
## rp           0.3617925665 0.0203419576  17.7855334  1.165208e-61
## pctsy        0.0224597556 0.0009182782  24.4585517 7.935982e-104
## cons        -0.0083744164 0.0253630744  -0.3301814  7.413324e-01
## eff          0.0004622544 0.0000590602   7.8268342  1.275086e-14

# Récupération de la prédiction du modèle m4
prevhat <- m4$fitted.values
```


$$txAT_i = \beta_0 + \beta_1 \widehat{prev}_i + \beta_2 cons_i + \beta_3 eff_i + \varepsilon_i$$

```
# Régression de seconde étape
m5 <- lm(txAT ~ prevhat + cons + eff)
coef(summary(m5))
##                 Estimate  Std. Error   t value      Pr(>|t|)
## (Intercept) 172.08062758  5.74261993 29.965526 3.714347e-141
## prevhat     -23.69200573 10.66651434 -2.221157  2.656455e-02
## cons        186.85374659  7.98097116 23.412407  6.221360e-97
## eff          -0.05412265  0.01928916 -2.805858  5.116045e-03

# Rappel : modèle initial (sans variable instrumentale)
coef(summary(m3))
##                 Estimate Std. Error   t value      Pr(>|t|)
## (Intercept) 161.78881990 5.53076418 29.252525 2.828794e-136
## prev         38.00128753 7.20440493  5.274730  1.630944e-07
## cons        187.26145996 7.89108887 23.730750  4.991724e-99
## eff          -0.08412927 0.01870076 -4.498709  7.641753e-06
```

Causalité et endogénéité
Exemple : Accidents du travail et prévention

L'utilisation de la prédiction de la variable `prev` en lieu et place de `prev` a pour effet de **« retourner » le coefficient** dans la régression du taux d'accidents du travail.

Dans la régresion originale, ce coefficient était biaisé par l'existence d'une causalité inverse entre accidents du travail et probabilité de mettre en place des dispositifs de prévention.

Avec la régression instrumentale, on « filtre » `prev` de façon à neutraliser cette corrélation.

La régression finale permet donc d'estimer le coefficient associé à `prev` sans le biais lié à l'endogénéité.

Causalité et endogénéité
Exemple : Accidents du travail et prévention

La démarche en deux étapes ne permet cependant pas d'aboutir à des erreurs standards correctes dans le second modèle.

Pour y parvenir, on estime les deux étapes simultanément à l'aide de la fonction `ivreg()` du *package* `AER` :

```
# Installation et chargement du package AER
# install.packages("AER")
library("AER")

# Estimation du modèle m6
m6 <- ivreg(txAT ~ prev + cons + eff | rp + pctsy + cons + eff)
coef(summary(m6))
##                Estimate  Std. Error   t value      Pr(>|t|)
## (Intercept) 172.08062758  5.88337377 29.248631 3.007776e-136
## prev        -23.69200573 10.92795474 -2.168018  3.039353e-02
## cons        186.85374659  8.17658786 22.852289  2.891135e-93
## eff          -0.05412265  0.01976195 -2.738731  6.277748e-03
```


Utiliser des pondérations dans une régression
Économétrie et sondages

La plupart des données sur lesquelles sont estimés les modèles économétriques sont obtenues par le biais d'enquêtes statistiques.

Contrairement aux recensements, ces enquêtes reposent sur un **échantillonnage** des unités à enquêter (*cf.$* cours du S1).

Pour garantir la représentativité des résultats et pouvoir les généraliser à l'ensemble de la population, on utilise une **pondération**.

Faut-il utiliser cette pondération dans des modèles de régression, et si oui comment ?

Utiliser des pondérations dans une régression
Économétrie et sondages

Si l'utilisation des pondérations pour les statistiques descriptives uni- et bi-variées fait consensus, ce n'est pas le cas pour les modèles économétriques de régression.

Un document de travail de l'Insee dresse un état des lieux de cette question : Davezies L., D'Hautfoeuille X. (2009), « Faut-il pondérer ? », *Document de travail*, 23 p.

Le choix de pondérer ou non les régressions va dépendre principalement de la plus ou moins bonne prise en compte de la sélection des individus dans le modèle et dans les pondérations.

Utiliser des pondérations dans une régression
Mécanisme de sélection et pondérations

Cas 1 Le modèle de régression intègre toutes les variables déterminant la sélection des unités.

Ces variables sont notamment :

- les variables du plan de sondage : strates, unités primaires, etc. ;
- les variables utilisées dans la correction de la non-réponse (groupes de réponse homogènes, variables du modèle utilisées pour calculer les probabilités de réponse, etc.).

Dans ce cas, **les résultats pondérés et non-pondérés doivent être très proches**.

En pratique Il est très rare de disposer de l'ensemble de ces informations dans les fichiers de diffusion.

Utiliser des pondérations dans une régression
Mécanisme de sélection et pondérations

Cas 2 Le modèle de régression n'intègre pas toutes les variables déterminant la sélection des unités, mais les pondérations oui.

Dans ce cas, **les résultats pondérés et non-pondérés peuvent différer et il est préférable de pondérer**.

Dans les enquêtes « classiques » sans sur-pondération ou non-réponse trop marquée, les résultats non-pondérés peuvent rester qualitativement proches des résultats pondérés.

En revanche, dans les enquêtes où certaines populations ont été volontairement surreprésentées (par exemple les hauts patrimoine dans les enquêtes Patrimoine), il est impératif de pondérer.

Utiliser des pondérations dans une régression
Mécanisme de sélection et pondérations

Cas 3 Les pondérations n'intègrent pas toutes les variables déterminant la sélection des unités.

Dans ce cas, **les résultats pondérés et non-pondérés peuvent différer et il est préférable de NE PAS pondérer**.

Faute de disposer de toutes les informations déterminant la sélection dans le modèle (cas 1), il est possible de mettre en œuvre des **méthodes économétriques spécifiques de correction du biais de sélection**.

On peut se ramener par exemple à un cas d'application des **variables instrumentales**, où la variable omise est la propension à être sélectionnée dans l'échantillon.

Utiliser des pondérations dans une régression
Utiliser les pondérations en pratique

Il n'est pas toujours facile d'appliquer ces recommandations :

- il n'est **pas évident de déterminer** dans quelle mesure le mécanisme de sélection est bien pris en compte par les pondérations de l'enquête ;
- les caractéristiques du plan de sondage à intégrer au modèle (variable de stratification, etc.) ne figurent en général **pas dans les fichiers de diffusion**.

Sauf cas de sur-échantillonnage spécifique, on recommande d'**estimer les modèles sur données pondérées et non-pondérées** et de **comparer leurs résultats d'un point de vue qualitatif**.

Utiliser des pondérations dans une régression
Exemple : Déterminants du salaire dans l'EEC

```
# Estimation du modèle non-pondéré
m7 <- lm(salred ~ age + femme + infbac + supbac, data = e)
coef(summary(m7))
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 1137.86974 160.145543  7.105223 2.728590e-12
## age           28.10184   3.328956  8.441637 1.534157e-16
## femme       -722.29984  74.390897 -9.709519 4.144809e-21
## infbac      -571.76582  97.758374 -5.848766 7.303926e-09
## supbac       454.21403 102.267365  4.441437 1.023730e-05

# Estimation du modèle pondéré
m7w <- lm(
  salred ~ age + femme + infbac + supbac
  , data = e, weights = extri1613
)
coef(summary(m7w))
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  937.29018 163.521885  5.731894 1.422966e-08
## age           33.33709   3.502043  9.519328 2.159807e-20
## femme       -743.80445  77.307423 -9.621385 8.934513e-21
## infbac      -581.06005 101.130549 -5.745643 1.316388e-08
## supbac       541.78308 103.759430  5.221531 2.282301e-07
```


En guise de conclusion
Vade-mecum sur les méthodes de régression

Utiliser la modélisation appropriée Adapter la modélisation à la nature de la variable dépendante, chercher à maîtriser au mieux les spécificités des méthodes mises en œuvres.

Estimer de nombreuses variantes Intégrer les variables explicatives progressivement pour mesurer leur impact sur la modélisation (R^2, pourcentage de concordance, tests), tester de nombreux modèles pour repérer les relations les plus structurantes.

Tenir compte de l'hétéroscédasticité Utiliser des écart-types robustes pour toutes les estimations.

En guise de conclusion
Vade-mecum sur les méthodes de régression

Travailler en amont la problématique de l'étude La mettre sous forme d'hypothèses ou de questions susceptibles d'être vérifiées empiriquement sur des données quantitatives.

Identifier les sources d'endogénéité Rechercher le ou les dispositifs techniques susceptibles de neutraliser au mieux cette endogénéité.

Connaître les données Maîtriser la documentation et le protocole de recueil des données (importance de la pondération), connaître la signification précise des différentes variables utilisées, repérer et traiter les valeurs aberrantes des variables de la modélisation.

En guise de conclusion
Vade-mecum sur les méthodes de régression

Choisir les modèles les plus pertinents Être parcimonieux dans le choix des modèles à présenter au lecteur tout en soulignant la robustesse des résultats obtenus (modèles complémentaires, variantes en annexe).

Présenter les résultats de façon judicieuse Adapter la présentation au public visé en fournissant toujours à la fois des informations générales sur le modèle et sur les paramètres estimés (valeur, significativité).

Contextualiser et interpréter Toujours replacer les résultats et les modèles dans la problématique de l'étude (pour en souligner les apports) et dans le cadre économétrique qui est le leur (pour en souligner les éventuelles limites).