Statistiques et informatique

appliquées aux sciences sociales

Traiter des données numériques 3

Mardi 21/11/2023 15h30-17h Censier D2

Licence de sciences sociales 3e année
Université Paris 1 Panthéon Sorbonne

Plan de la séance : régressions linéaires

- Généralisation à plusieurs variables
- Exemple de régression linéaire
- Méthode d’analyse de la variance
- Régressions logistiques

Généralisation à deux ou plusieurs variables explicatives

- Avec deux variables explicatives notées x_1 et x_2
- i un individu, on vérifie l'équation $y_i = a_1 \cdot x_{i1} + a_2 \cdot x_{i2} + b + u_i$
- se résout de façon analogue au cas à 1 variable, avec des généralisations de ses propriétés...
- $\bar{u} = 0 ; cov(x_1, u) = 0 ; cov(x_2, u) = 0$
- Pour trouver a_1 et a_2, on résout le système linéaire

$$\begin{cases} a_1 \cdot Var(x_1) + a_2 \cdot Cov(x_1, x_2) = Cov(x_1, y) \\ a_1 \cdot Cov(x_1, x_2) + a_2 \cdot Var(x_2) = Cov(x_2, y) \end{cases}$$

Généralisation à 2 variables explicatives

- Pour trouver b, on applique $\bar{y} = a_1 \cdot \overline{x_1} + a_2 \cdot \overline{x_2} + b$
- On va utiliser la même définition du coefficient de détermination R²…

$$R^2 = \frac{Variance\ expliquée}{Variance\ totale}$$

- …sachant que la décomposition de la variance des y_i s'écrit toujours comme suit :

$$Variance\ totale = Variance\ expliquée + Variance\ des\ résidus$$

avec $Variance\ totale = Var(y)$ et $Variance\ des\ résidus = Var(u)$

mais désormais

$$Variance\ expliquée = a_1^2 \cdot Var(x_1) + 2\,a_1\,a_2 \cdot Cov(x_1, x_2) + a_2^2 \cdot Var(x_2)$$

Généralisation à plusieurs variables explicatives

- On utilise le langage des matrices, tableaux de coefficients d'une application linéaire à deux ou plusieurs variables

$$Y = X \cdot A + B + U \text{ avec } Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}; X = \begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & x_{n3} \end{pmatrix}; A = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}; B = \begin{pmatrix} b \\ b \\ \vdots \\ b \end{pmatrix}; U = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

- On peut noter $V(X) = \begin{bmatrix} Var(x_1) & Cov(x_1, x_2) & Cov(x_1, x_3) \\ Cov(x_2, x_1) & Var(x_2) & Cov(x_2, x_3) \\ Cov(x_3, x_1) & Cov(x_3, x_2) & Var(x_3) \end{bmatrix}$ et $Cov(Y, X) = \begin{pmatrix} Cov(y, x_1) \\ Cov(y, x_2) \\ Cov(y, x_3) \end{pmatrix}$

la matrice des variances-covariances des x_j, symétrique et inversible au sens des matrices si et seulement si les variables x_j sont linéairement indépendantes

Le vecteur des covariances de y avec les x_j

- On résout les équations matricielles $Cov(Y, X) = V(X) \cdot A$ et (pour trouver la constante b) : $\bar{y} = \begin{pmatrix} \bar{x}_1 & \bar{x}_2 & \bar{x}_3 \end{pmatrix} \cdot A + b$

Généralisation à plusieurs variables explicatives

- On va encore utiliser la même définition du coefficient de détermination R^2…

- L'exemple du cours pourra être approfondi avec 2, 3 ou 4 variables explicatives, éventuellement en utilisant tous les pays

Exemple de régression linéaire

Cet article distingue et compare le pouvoir explicatif de quatre types de transmissions culturelles : la transmission domestique osmotique, la transmission domestique stratégique, la transmission scolaire par les pairs, la transmission par l'institution scolaire et ses agents. Pour ce faire, il étudie la manière dont le plaisir de lire se transmet chez les jeunes de 15 ans à partir d'une analyse secondaire des données PISA 2009. Les résultats montrent que la transmission culturelle par l'école est nettement plus efficace que la transmission culturelle qui s'opère au sein de la famille. En examinant l'efficacité relative des quatre modes de transmission culturelle parmi les classes moyennes supérieures, l'enquête met en évidence que, même dans ces milieux, la transmission culturelle par osmose ne suffit pas à assurer la reproduction culturelle. Ce modèle traditionnel de transmission culturelle, défini ici comme le « modèle des héritiers », tend à être remplacé par un nouveau modèle de transmission culturelle, passant tantôt par une transmission domestique stratégique et tantôt par la transmission scolaire. La première stratégie étant privilégiée par les fractions intellectuelles des classes moyennes supérieures, la seconde par leurs fractions managériales.

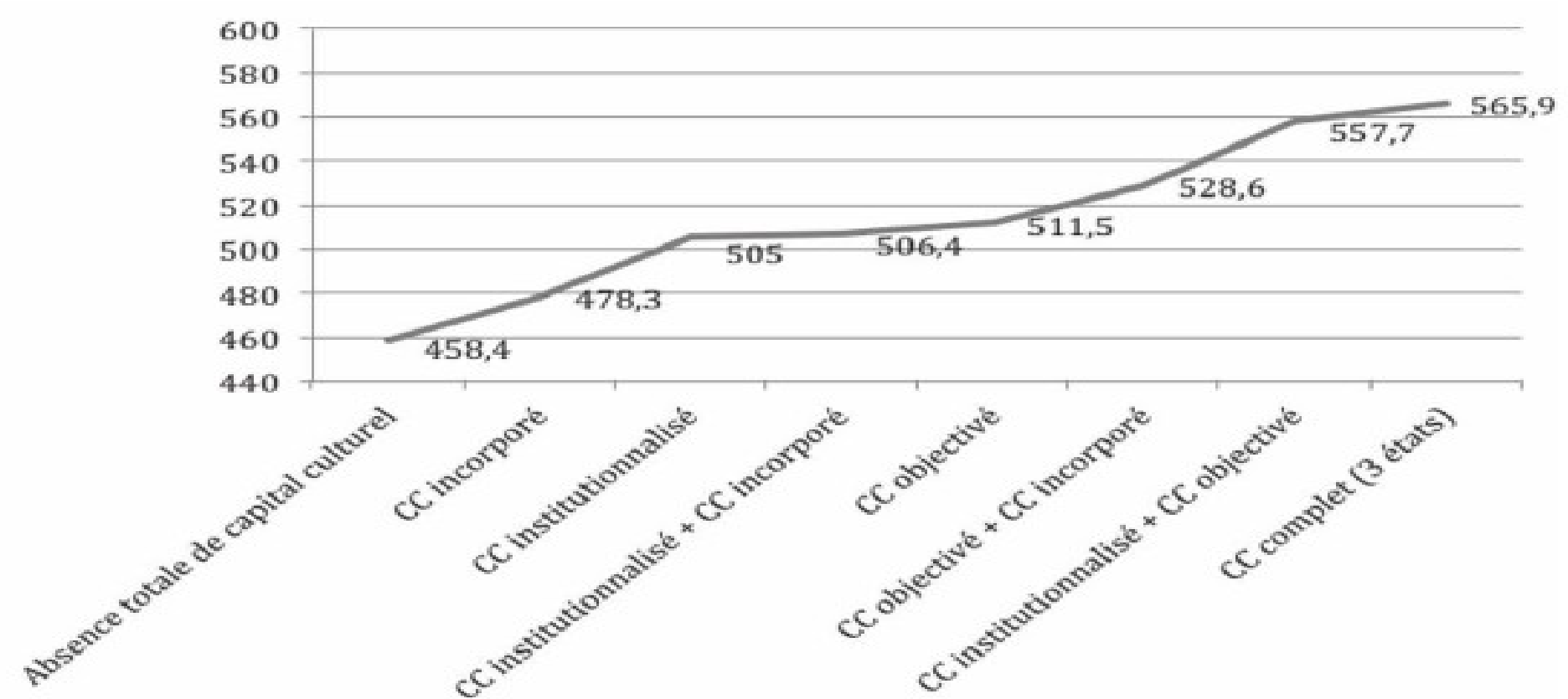
Hugues Draelants, 2016, « Formes et évolutions de la transmission culturelle. Le « modèle des héritiers » à l'épreuve des données PISA 2009 », *Revue française de pédagogie*, 194, http://rfp.revues.org/4967 ; DOI : 10.4000/rfp.4967

Exemple de régression linéaire

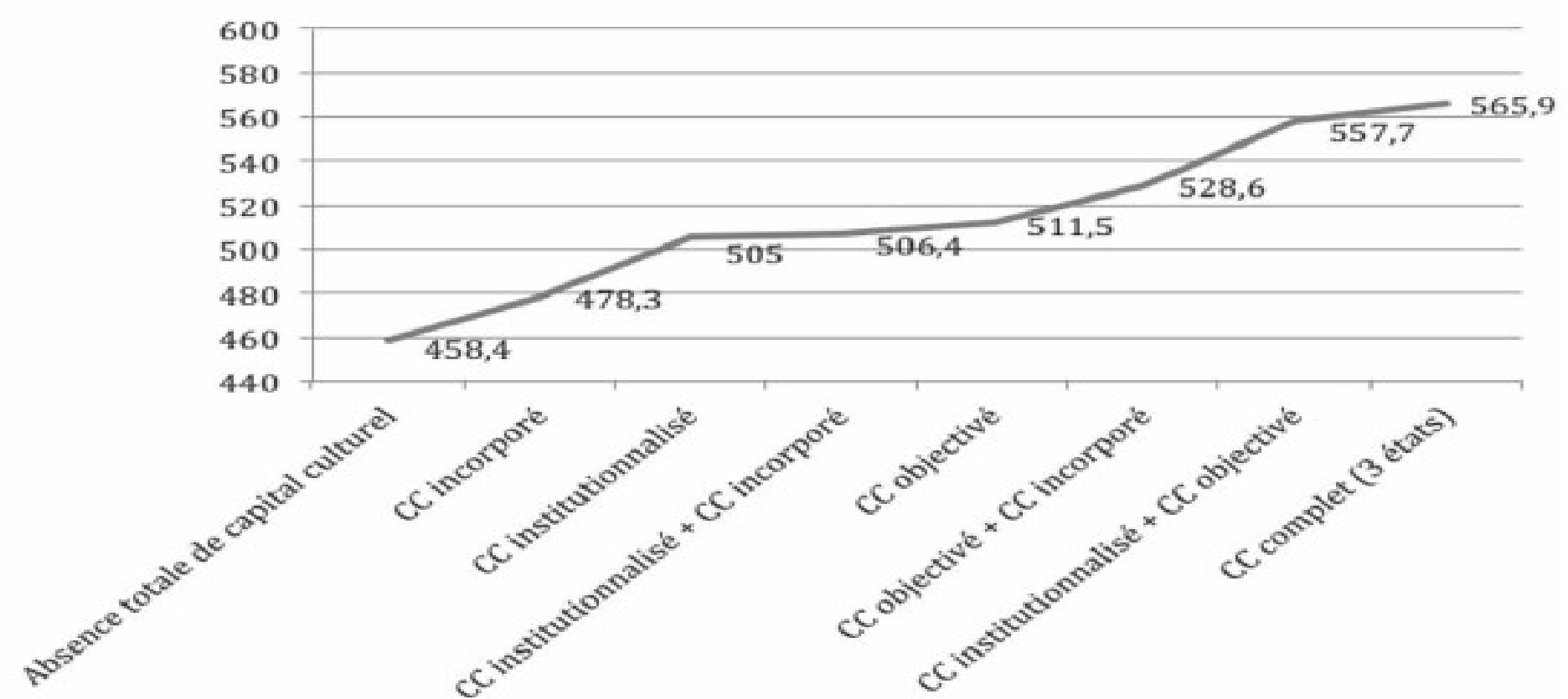
Tableau 1. Typologie des modes de transmission culturelle

	Socialisation verticale (domestique ou parentale)	Socialisation horizontale ou indirecte (par l'école)
Socialisation par imprégnation	Transmission culturelle parentale « osmotique »	Transmission culturelle scolaire par les pairs
Socialisation par inculcation	Transmission culturelle parentale « stratégique »	Transmission culturelle scolaire par l'institution et ses agents

Hugues Draelants, 2016, « Formes et évolutions de la transmission culturelle. Le « modèle des héritiers » à l'épreuve des données PISA 2009 », *Revue française de pédagogie*, 194, http://rfp.revues.org/4967 ; DOI : 10.4000/rfp.4967



Hugues Draelants, 2016, « Formes et évolutions de la transmission culturelle. Le « modèle des héritiers » à l'épreuve des données PISA 2009 », *Revue française de pédagogie*, 194, http://rfp.revues.org/4967 ; DOI : 10.4000/rfp.4967



Hugues Draelants, 2016, « Formes et évolutions de la transmission culturelle. Le « modèle des héritiers » à l'épreuve des données PISA 2009 », *Revue française de pédagogie*, 194, http://rfp.revues.org/4967 ; DOI : 10.4000/rfp.4967

Tableau 2. Capital culturel des parents et pratiques de lecture des enfants (%)

Pratiques de lecture des enfants	30 minutes ou moins par jour	Plus de 30 minutes par jour
Capital culturel institutionnalisé (-) : pas de diplôme du supérieur	72,1	27,9
Capital culturel institutionnalisé (+) : diplôme du supérieur	63,9	36,1
Capital culturel incorporé (-) : petit lecteur	72,1	27,9
Capital culturel incorporé (+) : grand lecteur	61,6	38,4
Capital culturel objectivé (-) : moins de 100 livres	75,7	24,3
Capital culturel incorporé (+) : plus de 100 livres	58,5	41,5

Lecture : 72,1 % des jeunes de 15 ans qui déclarent lors de l'enquête PISA 2009 lire en moyenne moins de 30 minutes par jour ont des parents peu dotés en capital culturel institutionnalisé, c'est-à-dire définis comme n'ayant pas de diplôme d'enseignement supérieur.

Hugues Draelants, 2016, « Formes et évolutions de la transmission culturelle. Le « modèle des héritiers » à l'épreuve des données PISA 2009 », *Revue française de pédagogie*, 194, http://rfp.revues.org/4967 ; DOI : 10.4000/rfp.4967

Disposition à lire par plaisir (CC incorporé) : La variable « JOYREAD » est une variable disponible dans la base de données PISA 2009 (voir OCDE, 2012), construite à partir des 11 propositions suivantes, avec lesquelles les élèves devaient marquer leur degré d'accord : « je lis seulement si je suis obligé ; la lecture est un de mes loisirs préférés ; j'aime bien parler de livres avec d'autres gens ; je trouve difficile de terminer un livre ; je suis content si je reçois un livre en cadeau ; pour moi lire est une perte de temps ; j'aime bien aller dans une librairie ou une bibliothèque ; je lis seulement pour trouver l'information dont j'ai besoin ; je ne parviens pas à rester assis et à lire plus de quelques minutes ; j'aime bien donner mon avis sur les livres que j'ai lus ; j'aime bien échanger des livres avec mes amis ».

Capital culturel institutionnalisé des parents : mesuré ici à partir de la variable PISA « PARED » qui renseigne le plus haut niveau d'éducation parentale exprimé en années d'études

Capital culturel incorporé appréhendé à partir d'une variable PISA « MOTREAD », visant à cerner leur attitude envers la lecture sur la base de leur degré d'accord avec les propositions suivantes : « lire est un de mes hobbies favoris ; je suis content si je reçois un livre en cadeau ; pour moi, lire représente une perte de temps ; j'aime me rendre dans une librai- rie ou en bibliothèque »

Capital culturel objectivé : appréhendé à partir d'un indicateur de possessions culturelles présent dans PISA 2009 (la variable « CULTPOSS ») qui nous renseigne sur la présence de littérature classique, de livres de poésie et d'œuvres d'art dans le lieu de vie

Hugues Draelants, 2016, « Formes et évolutions de la transmission culturelle. Le « modèle des héritiers » à l'épreuve des données PISA 2009 », *Revue française de pédagogie*, 194, http://rfp.revues.org/4967 ; DOI : 10.4000/rfp.4967

Transmission culturelle active

- soutien parental à la lecture durant la petite enfance, à 6 ans, au moment où l'enfant entre en principe à l'école élémentaire. Il s'agit d'un indice factoriel construit à partir de la fréquence rapportée par les parents concernant les activités suivantes : lire des livres à son enfant ; lui raconter des histoires ; jouer avec lui à des jeux en rapport avec l'alphabet ; lui parler de livres que vous avez lus ; jouer ensemble à des jeux en rapport avec les mots ; écrire des lettres ou des mots.

- pratiques actuelles de soutien et d'encouragement à la lecture (ou plus largement à la compréhension de l'écrit), c'est-à-dire lorsque l'enfant est âgé de 15 ans. Cet indice factoriel s'appuie sur la fréquence rapportée par les parents concernant le fait de discuter de livres, de films ou de programmes de télévision avec son enfant ; d'aller avec lui dans une librairie ou une biblio- thèque ; ou encore de parler avec lui de ce qu'il est en train de lire.

Hugues Draelants, 2016, « Formes et évolutions de la transmission culturelle. Le « modèle des héritiers » à l'épreuve des données PISA 2009 », *Revue française de pédagogie*, 194, http://rfp.revues.org/4967 ; DOI : 10.4000/rfp.4967

La transmission culturelle scolaire par les pairs

La méthode utilisée ici pour mesurer l'effet des pairs est [ici] centrée sur le groupe : chaque élève est comparé à un groupe de pairs imposé, celui des élèves fréquentant le même établissement que lui. La propension à lire par plaisir d'un élève est donc comparée au score moyen calculé pour le groupe.

Les indicateurs disponibles nous permettent de distinguer entre trois types d'effets des pairs, ceux qui tiennent à la composition sociale de l'école (nombre moyen d'années d'études suivies par les parents d'élèves par école), ceux qui relèvent de la composition académique de l'école (moyenne obtenue par les élèves d'une école aux épreuves PISA 2009 de lecture) et ceux qui concernent ce que nous avons appelé, par analogie avec les deux effets précédents, la « composition culturelle de l'école » qui n'est autre que la propension moyenne des enfants de l'école fréquentée à avoir une attitude ou des pratiques favorables à la lecture (calculée à partir de la variable JOYREAD présentée plus haut).

Hugues Draelants, 2016, « Formes et évolutions de la transmission culturelle. Le « modèle des héritiers » à l'épreuve des données PISA 2009 », *Revue française de pédagogie*, 194, http://rfp.revues.org/4967 ; DOI : 10.4000/rfp.4967

La transmission culturelle scolaire institutionnelle

- diverses pratiques enseignantes d'encouragement à la lecture (indice factoriel fondé sur le degré d'accord exprimé par les élèves avec les propositions suivantes : « le professeur recommande aux élèves de lire un livre ou un auteur ; le professeur encourage les élèves à exprimer leur opinion sur un texte ; le professeur aide les élèves à faire un lien entre les récits qu'ils lisent et leur propre vie »).

- performance scolaire en lecture de l'élève (sa moyenne aux 5 épreuves PISA 2009 de lecture). Nous faisons en effet l'hypothèse que le niveau en lecture au test PISA reflète les verdicts que l'institution scolaire renvoie habituellement aux élèves et qui leur permettent ou non de se vivre comme des lecteurs compétents et, à ce titre, les incitent à développer des attitudes et pratiques favorables à la lecture [...]

Tableau 5. Corrélations entre l'ensemble des variables utilisées

	Lecture par plaisir	Cap. cult. instit.	Cap. cult. incorporé	Cap. cult. objectivé	Comp. culturelle	Comp. sociale	Comp. acad.	Soutien en primaire	Soutien à 15 ans	Perform. lecture	Stimulation prof.
Disposition à lire par plaisir	1										
Capital culturel institutionnalisé		1									
Capital culturel incorporé			1								
Capital culturel objectivé				1							
Composition culturelle école	0,389				1						
Composition sociale école		0,522			0,449	1					
Composition académique		0,326		0,324	0,694	0,612	1				
Soutien parental en primaire			0,321					1			
Soutien parental à 15 ans			0,365					0,437	1		
Performance en lecture	0,418			0,323	0,523	0,461	0,754			1	
Stimulation professorale											1

Note : seules les corrélations supérieures à 0,3 sont reprises.

Tableau 6. Fraction de la variance expliquée (%) par les quatre principaux types de transmission culturelle (variable dépendante : attitude des jeunes de 15 ans à l'égard de la lecture)

	Variance expliquée (par chaque type de transmission considérée indépendamment)	% du total de variance expliquée
Transmission culturelle parentale osmotique (type 1)	10,1	32,9
Transmission culturelle parentale stratégique (type 2)	6,4	20,8
Transmission culturelle scolaire institutionnelle (type 3)	18,4	59,9
Transmission culturelle scolaire par les pairs (type 4)	15,1	49,2
Transmission culturelle par imprégnation (types 1 et 4)	19,2	62,5
Transmission culturelle par inculcation (types 2 et 3)	23,1	75,2
Transmission culturelle domestique (types 1 et 2)	13,7	44,6
Transmission culturelle scolaire (types 3 et 4)	26,3	85,7
Transmission culturelle verticale (types 1 et 3)	21,9	71,3
Transmission culturelle mixte (types 2 et 4)	19	61,9
Transmission culturelle (types 1, 2 et 3)	25,2	82,1
Transmission culturelle (types 1, 2 et 4)	22,1	72
Transmission culturelle (types 1, 3 et 4)	28,5	92,8
Transmission culturelle (types 2, 3 et 4)	29	94,5
Transmission culturelle (modèle complet)(types 1, 2, 3 et 4)	30,7	

Tableau 8. Effet des différents types de transmission culturelle sur l'attitude envers la lecture selon la profession de la mère (% de variance expliquée)

	Intellectuels	Enseignants	Professionnels	Managers
Transmission domestique osmotique	12,5	7,2	14,6	14,6
Transmission domestique intentionnelle	12,5	9,2	10,5	5
Transmission scolaire par les pairs	10	14,6	17,3	8,8
Transmission scolaire institutionnelle	14,7	20,5	26,3	22
% total de la variance expliquée	39,2	36	39,6	32,4

Analyse de la variance

- Principe proche de la décomposition de la variance dans une régression linéaire
- s'applique au cas d'une seule variable explicative qualitative avec une variable expliquée numérique
- On décompose la variance totale de la variable expliquée en deux termes :

Variance totale = Variance intra-classes + Variance inter-classes

$$Variance\ intra-classes = \sum_{k=1}^{K} \frac{n_k}{n} \cdot \sigma_k^2 \text{ en notant } \sigma_k^2 = \frac{1}{n_k} \cdot \sum_{i \in c_k} (y_i - \bar{y}_k)^2 \text{ et } \bar{y}_k = \frac{1}{n_k} \sum_{i \in c_k} y_i$$

n_k *l'effectif de la classe* c_k *pour k variant de* 1 *à* K

$$Variance\ inter-classes = \sum_{k=1}^{K} \frac{n_k}{n} \cdot (\bar{y}_k - \bar{y})^2$$

Analyse de la variance

- On peut calculer la part de la variance expliquée

$$\eta^2 = \text{Rapport de corrélation interclasses} = \frac{\text{Variance interclasses}}{\text{Variance totale}}$$

- Alternative

$$= \frac{\sum_{k=1}^{K} \frac{n_k}{n} \cdot \sigma_k^2}{\frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{\sum_{k=1}^{K} \frac{n_k}{n} \cdot \sigma_k^2}{\sum_{k=1}^{K} \frac{n_k}{n} \cdot (\bar{y}_k - \bar{y})^2 + \sum_{k=1}^{K} \frac{n_k}{n} \cdot \sigma_k^2}$$

-

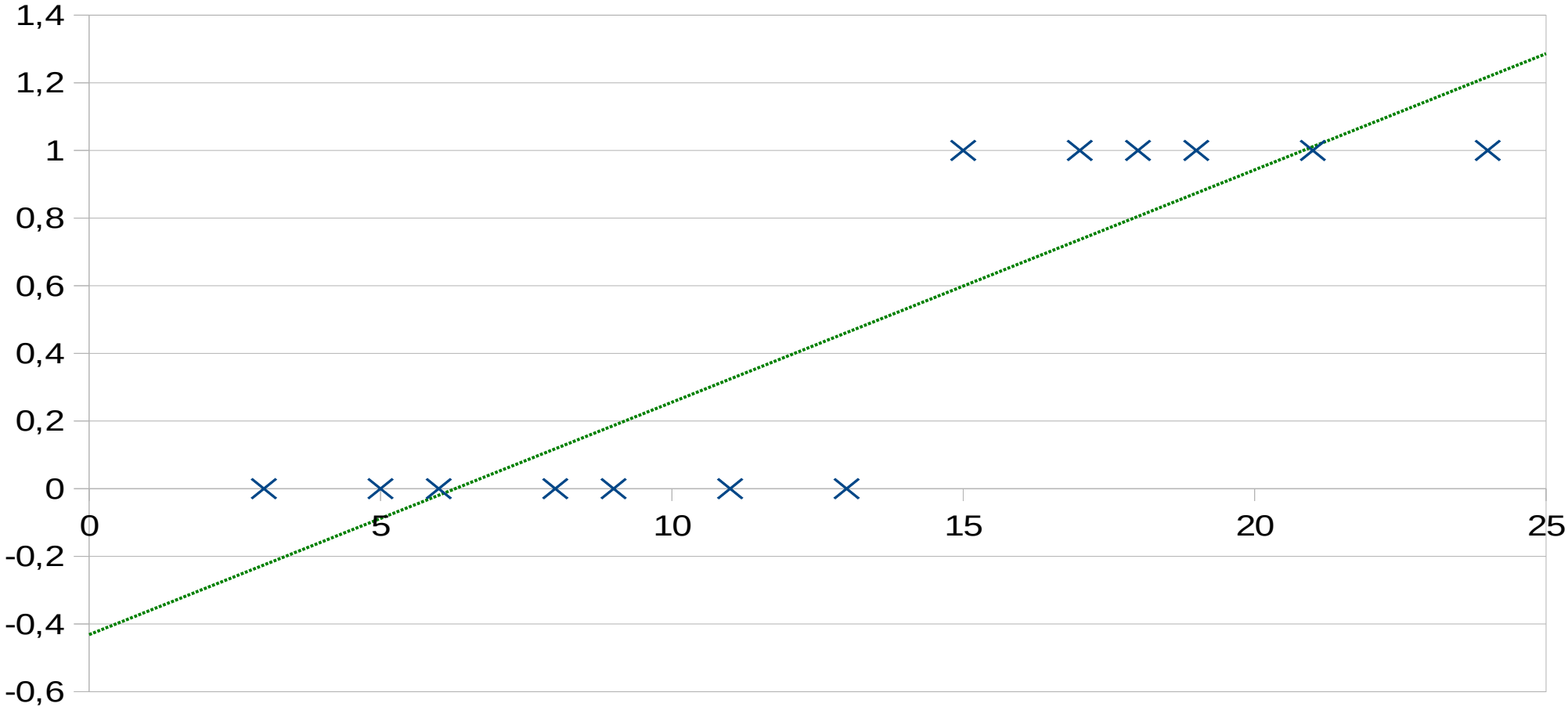
Régressions logistiques

→ diapositive suivante

Régressions logistiques

- Présentation sommaire, liens avec les régressions linéaires
- Spécificités, utilisation des rapports de chances (odds ratio)
- Un exemple d’utilisation

Que donne une régression linéaire avec une variables dépendante dichotomique ?



Principe général

- Expliquer une variable dichotomique à l'aide d'une batterie de variables numériques/dichotomiques
- On se ramène à une modélisation (avec effet de seuil) sur une variable dite latente dépendant linéairement des explicatives
- La variable latente, non observable, et les coefficients de la régression, calculables, sont définis à un facteur multiplicatif près : ce problème d'identification est résolu en postulant que les résidus suivent une loi normale centrée réduite (modèle probit) ou une loi qui en est proche, la loi logistique (modèle logit)
- NB : Une loi normale centrée réduite est une loi gaussienne de moyenne 0 et d'écart-type 1. Pour mémoire, les résidus d'une régression linéaire sont de moyenne nulle mais leur écart-type intervient dans l'analyse de la variance de la variable expliquée ! Il n'a donc aucune raison d'être unitaire !
- Contrairement à la régression linéaire, où l'on dispose d'une résolution analytique (c'est à dire permettant un calcul direct des coefficients et autres indicateurs issus du modèle), on obtient les coefficient par approximations successives, en maximisant la vraisemblance du modèle i.e. pour une observation la probabilité P1 que la variable latente soit supérieure à 0 si Z=1 et 1-P1 si Z=0.
- NB : Cette méthode du maximum de vraisemblance est également utilisable pour les régressions linéaires, et donne des résultats très proches de celle des moindres carrés ordinaires
- Tout comme pour un modèle linéaire, on est en mesure de tester la nullité de chaque coefficient de la régression en donnant la probabilité d'observer le coefficient calculé sous l'hypothèse d'une nullité du « vrai coefficient » et on utilise les seuils usuels (1-5-10%) pour rejeter cette hypothèse et donc considérer un coefficient donné comme « significatif », grâce au chi^2 de Wald

Notation et modélisation (pour information)

- On peut noter Z la variable d'intérêt / dépendante /expliquée, qui est ici dichotomique (valant 0 ou 1), X le vecteur des variables explicatives X_k , k valant de 1 à K

- On suppose qu'il existe une variable Y numérique et continue telle que Z vaut 1 si $Y \geq 0$ et Z vaut 0 sinon.

- Explication de la multiplicité des solutions : pour toute variable latente $Y = \Sigma_k \alpha_k X_k + \beta + U$ solution du problème tel que $Z_i=1$ si $Y_i \geq 0$ et 0 sinon, il en existe une infinité d'autres de type $\lambda \cdot Y$ avec $\lambda>0$. En effet, ces variables vérifient également la condition recherchée. On peut donc choisir une solution telle que les résidus soit de variance égale à 1.

- Pour obtenir une solution, le critère du maximum de vraisemblance remplace celui de la minimisation de la variance des résidus : la vraisemblance est la probabilité que chaque Z soit en accord avec les observations compte-tenu des explicatives et de la distribution postulée pour les résidus, en considérant ces derniers indépendants et de même loi (en général logistique ou normale centrée réduite.

- La probabilité d'un accord avec les observations s'écrit $prob(Y \geq 0|X_k)$ si Z=1 et $prob(Y<0|X_k)$ sinon. Sous l'hypothèse d'indépendance entre les résidus, la vraisemblance est le produit de toutes ces probabilités, pour chaque individu i.

Spécificités du modèle logit

- Il n'existe pas vraiment d'équivalent satisfaisant du R^2 pour mesurer la qualité de l'ajustement.
- On dispose en revanche d'un moyen de calculer la probabilité P1 pour chaque combinaison des explicatives : on peut donc regarder dans quelle mesure cette distribution est en adéquation avec les réalisations de la variable expliquée (par exemple pour le seuil P1>0,5)

→ Ainsi on compare souvent la distribution de la variable dépendante avec celle de la dichotomique valant 1 si P1>0,5, 0 sinon, au moyen d'un tableau croisé (sous SPSS par exemple)

→ On peut aussi calculer la proportion de paires de deux observations concordantes / discordantes / liées d'un individu 1 pour qui Z=1 et d'un individu 2 pour qui Z=0 : la paire est dite concordante si P1>P2, discordante si P1<P2, liée si P1=P2 (statistique donnée par le logiciel SAS). Plus la proportion de paires concordantes est élevée, meilleur est le modèle.

Usages du modèle logit

- En pratique, le modèle logit donne des solutions faciles à interpréter si l'on travaille non avec des probabilités, mais avec des rapports de chance (Odd Ratio)
- En remplaçant les paramètres par leurs exponentielles, en effet, on obtient un coefficient multiplicatif positif qui augmente ou diminue ce rapport de chance selon qu'il est supérieur ou inférieur à 1
- Pour cette raison, la plupart des auteurs utilisent également des explicatives de type dichotomique, en faisant en sorte que la situation de référence soit une situation fréquente ou « moyenne » (et en tout cas possible), ceci afin de mieux interpréter l'impact relatif de chaque modalité explicative étudiée

$$\mathrm{OR} = \frac{P_1}{1 - P_1}$$