Séance : Regression avec variable dépendante dichotomique

Régression logistique et probit

Visseho Adjiwanou, PhD.

Plan de présentation

- Rappel
- Introduction
- Variables dépendantes dichotomiques
  - Exemple
  - Estimation:
    - Pourquoi le MCO n'est pas approprié
    - Logit ou probit?
  - Estimation
  - Interprétation
  - Tests d'hypothèses

Rappel

Quelle méthode de régression ?

- Le type de méthode dépend du type de la **variable dépendante**

Variables dépendantes	Méthodes
Quantitatives continues	Régression linéaire
Qualitative dichotomique	Logistique, probit
Qualitative avec plus de deux catégories nominale	Logit ou probit multinomial
Qualitative avec plus de deux catégories (ordinale)	Logit ou probit ordonné
Durée	Modèle de durée ou de survie

Introduction

Introduction

- La variable dépendante dichotomique est un cas particulier de variable dépendante qualitative où la variable dépendante n'a que deux catégories
  - Succès / perte, malade ou non, entrée dans la sexualité ou non
- Variable dépendante qualitative

Introduction

- L'analyse de régression d'une variable qualitative binaire ou dichotomique est un problème courant en sociologie
- Ces modèles se concentrent sur les déterminants de la probabilité p d'occurrence d'un résultat plutôt que d'un autre résultat qui se produit avec une probabilité de 1-p.
- Exemples:
  - Modéliser si le premier rapport sexuel a eu lieu pendant l'adolescence ou non
  - Modéliser si une personne a utilisé une méthode de contraception moderne ou pas
  - Donnez-moi d'autres exemples

Estimation

Estimation

- Dans l'analyse de régression, nous voulons mesurer comment la probabilité p varie d'un individu à l'autre en fonction des régresseurs (variables indépendantes0.
- Trois principales approches d'estimation sont utilisées:
  1. Le modèle de probabilité linéaire
     - Souvent dans le cas d'un régresseur endogène
  2. Le modèle logit
  3. le modèle probit

1. Le modèle de probabilité linéaire

- $Y_i = \beta_0 + \beta_1 X_i + ... + \beta_k X_k + \epsilon + j$

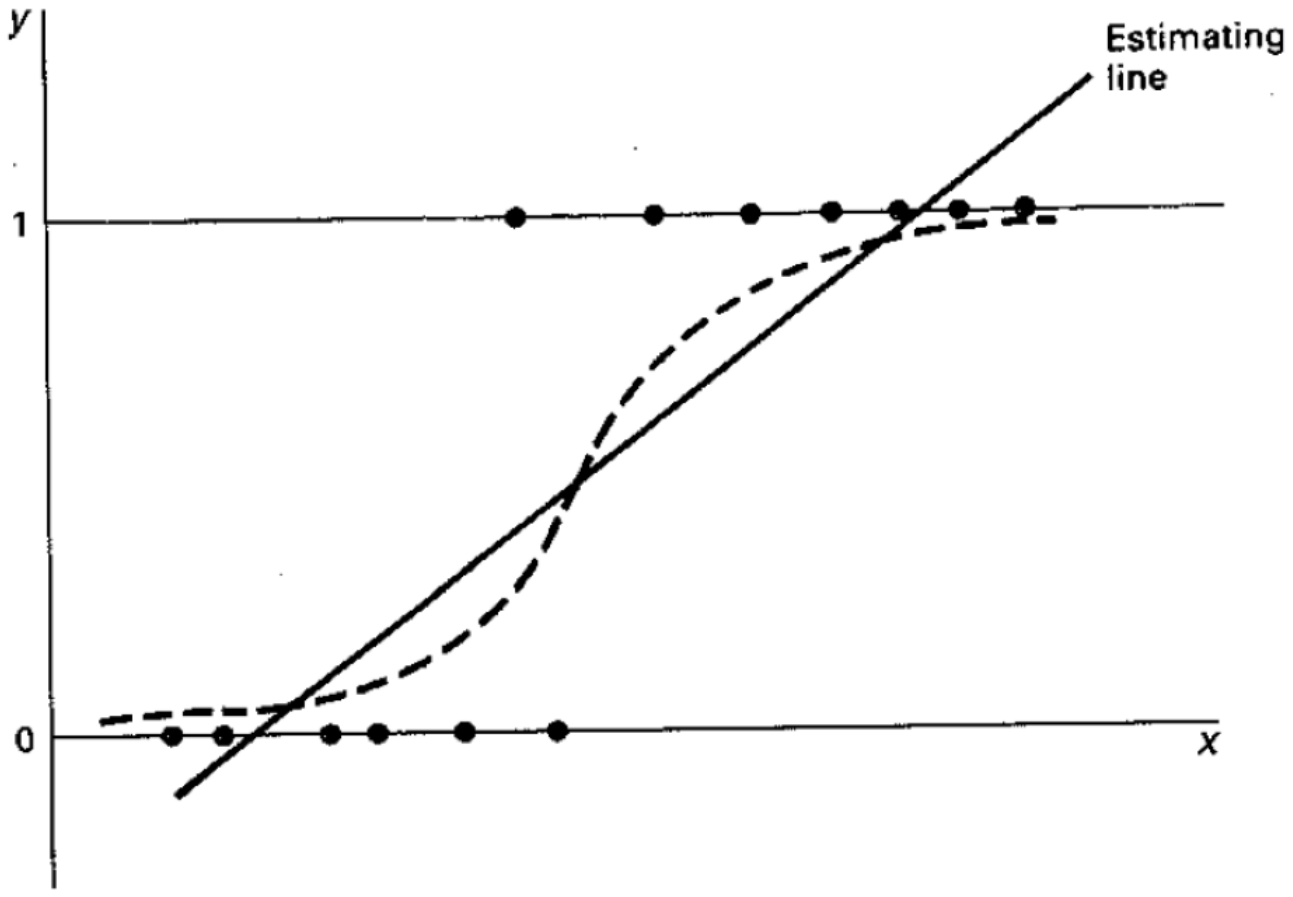
1. Le modèle de probabilité linéaire

- $Y_i = \beta_0 + \beta_1 X_i + ... + \beta_k X_k + \epsilon + j$
- Parce que Y ne peut prendre que deux valeurs, β_j ne peut pas être interprété comme le changement de Y étant donné une augmentation d'une unité de X_j, en maintenant tous les autres facteurs fixes: Y passe de 0 à 1, ou de 1 à 0 (ou ne change pas).

1. Le modèle de probabilité linéaire

- $Y_i = \beta_0 + \beta_1 X_i + ... + \beta_k X_k + \epsilon + j$
- Parce que Y ne peut prendre que deux valeurs, β_j ne peut pas être interprété comme le changement de Y étant donné une augmentation d'une unité de X_j, en maintenant tous les autres facteurs fixes: Y passe de 0 à 1, ou de 1 à 0 (ou ne change pas).
- Le modèle de régression linéaire multiple avec une variable dépendante binaire est appelé le modèle de probabilité linéaire (LPM) car la probabilité de réponse est linéaire dans les paramètres β_j.

1. Le modèle de probabilité linéaire



- Il est évident que la droite d'estimation n'est pas appropriée pour traiter la variable dépendante dichotomique

1. Le modèle de probabilité linéaire

- Formulation
- $Y_i = \beta_0 + \beta_1 X_i + ... + \beta_k X_k + \epsilon_i$
- $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ (simple régression linéaire)
- $Y = X\beta + \epsilon$ sous la forme matricielle
- $E(Y|X) = \beta_0 + \beta_1 X_i + ... + \beta_k X_k$
- Que vaut $E(Y|X)$?

1. Le modèle de probabilité linéaire

- Si Y est discrète:
- $E(Y|X) = \sum_k kP(Y = k|X)$
- Y variable dichotomique prend les valeurs 0 et 1
- $E(Y|X) = 0 * P(Y = 0|X) + 1 * (Y = 1|X)$

$==>$ E(Y|X) = P(Y = 1|X)

- donc, E(Y|X) est interprété comme une probabilité

1. Le modèle de probabilité linéaire

- Si Y est discrète:
- $E(Y|X) = \sum_k kP(Y = k|X)$
- Y variable dichotomique prend les valeurs 0 et 1
- $E(Y|X) = 0 * P(Y = 0|X) + 1 * (Y = 1|X)$

$==>$ E(Y|X) = P(Y = 1|X)

- donc, E(Y|X) est interprété comme une probabilité
- β_j mesure le changement de la probabilité de succès lorsque X_j change, en maintenant les autres facteurs fixes

1. Le modèle de probabilité linéaire

- Exemple: calcul de l'espérance quand Y est discrète
- Soit la variable aléatoire age qui a la distribution suivante:

Age	Nombre de cas
10	5
11	6
12	4

- Quelle est l'espérance de Y (moyenne de Y)

1. Le modèle de probabilité linéaire

- Exemple: calcul de l'espérance quand Y est discrète
- Soit la variable aléatoire age qui a la distribution suivante:

Age	Nombre de cas
10	5
11	6
12	4

- Quelle est l'espérance de Y (moyenne de Y)
- M = (5*10 + 11*6 +4*12)/15

1. Le modèle de probabilité linéaire

- Vous pouvez vous rendre compte que cela est exactement:
- E(Y) = 10*P(Y=10) + 11*P(Y=11) + 12*P(Y=12)
- P(Y=10) = 5/15
- P(Y=11) = 6/15
- P(Y=12) = 4/15

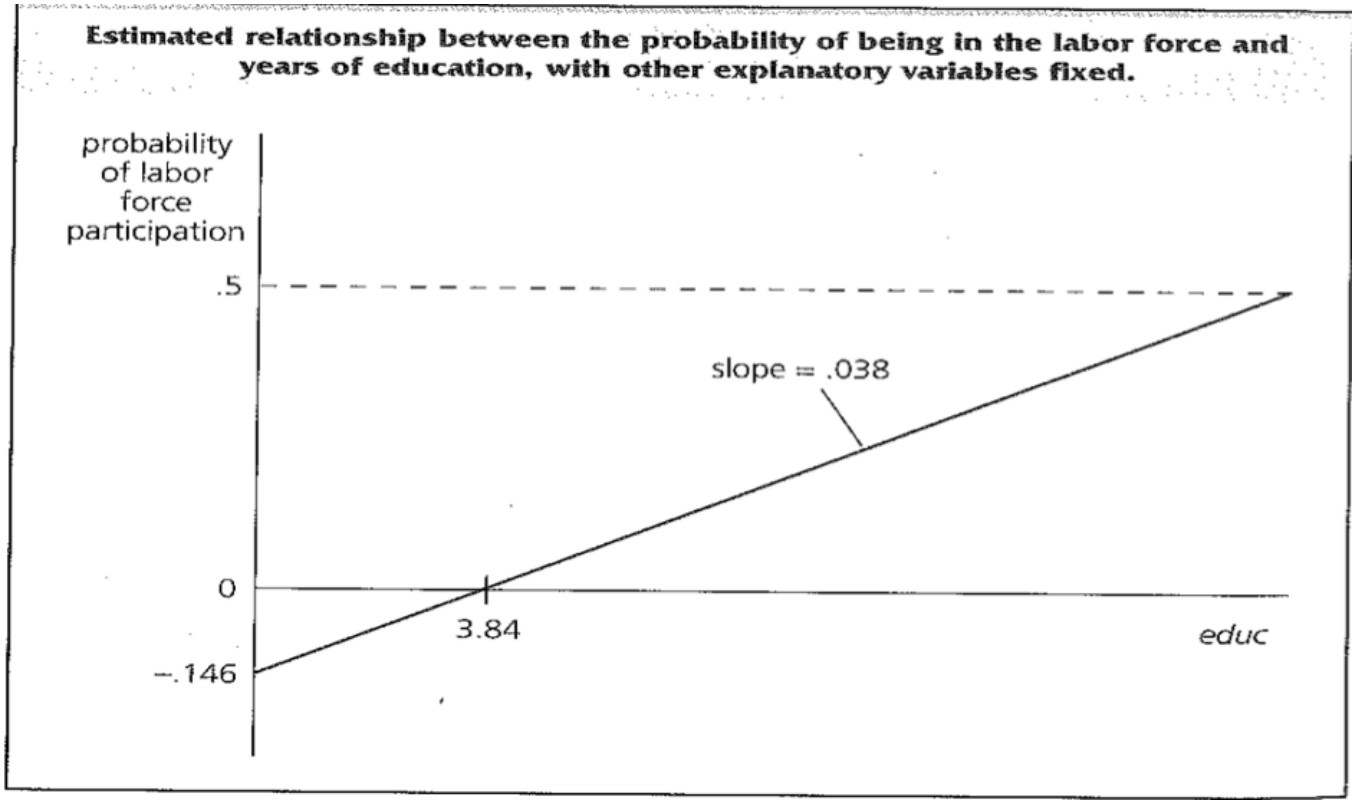
==> E(Y) = 10*5/15 + 11*6/15 + 12*4/15

1. Le modèle de probabilité linéaire

- Deux principaux problèmes avec le LPM:
  1. La probabilité prédite est supérieure à 1 ou inférieure à 0
  2. Les termes d'erreurs sont hétéroscédastiques

1. Le modèle de probabilité linéaire

1. Les valeurs prédites sont illimitées



1. Le modèle de probabilité linéaire

2. Les termes d'erreurs sont hétéroscédastiques

- ϵ_i prend deux valeurs:
- $-X\beta$ si Y $= 0$ avec P(Y $= 0) = 1 - X\beta$
- $(1 - X\beta)$ si Y $= 1$ avec P(Y $= 1) = X\beta$

1. Le modèle de probabilité linéaire

Valeur de ϵ	Probabilité
$-X\beta$	$1 - X\beta$
$(1 - X\beta)$	$X\beta$

- $E(\epsilon) = -X\beta * (1 - X\beta) + (1 - X\beta) * X\beta = 0$

1. Le modèle de probabilité linéaire

Valeur de ϵ	Probabilité
$-X\beta$	$1 - X\beta$
$(1 - X\beta)$	$X\beta$

- $E(\epsilon) = -X\beta * (1 - X\beta) + (1 - X\beta) * X\beta = 0$
- $Var(\epsilon) = (-X\beta)^2 * (1 - X\beta) + (1 - X\beta)^2 * X\beta = X\beta(1 - X\beta)$

1. Le modèle de probabilité linéaire

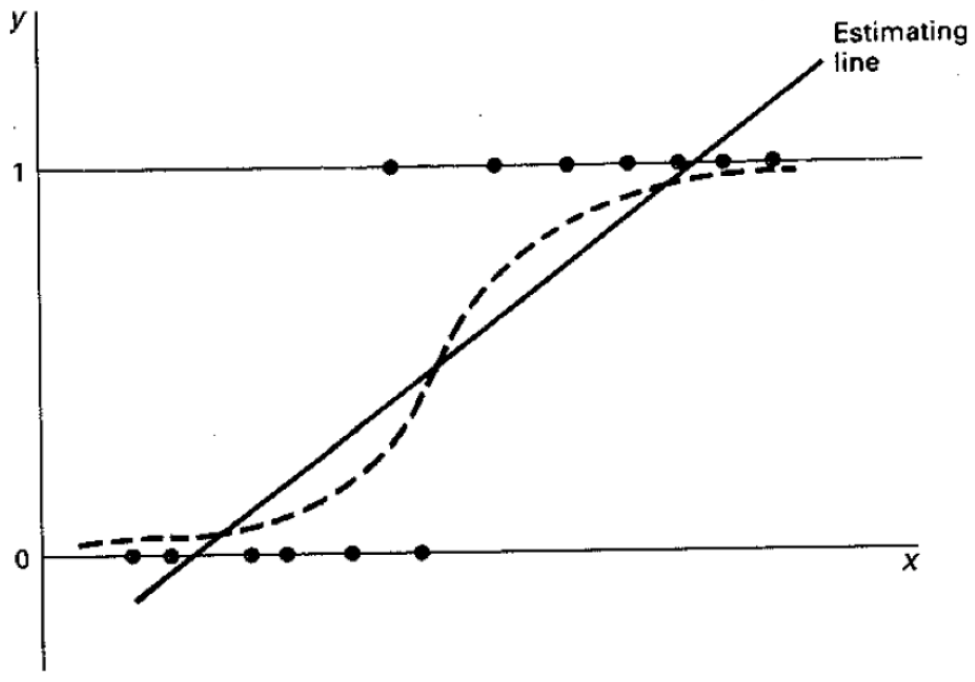
Valeur de ϵ	Probabilité
$-X\beta$	$1 - X\beta$
$(1 - X\beta)$	$X\beta$

- $E(\epsilon) = -X\beta * (1 - X\beta) + (1 - X\beta) * X\beta = 0$
- $Var(\epsilon) = (-X\beta)^2 * (1 - X\beta) + (1 - X\beta)^2 * X\beta = X\beta(1 - X\beta)$
- $Var(\epsilon)$ n'est pas constante

1. Le modèle de probabilité linéaire

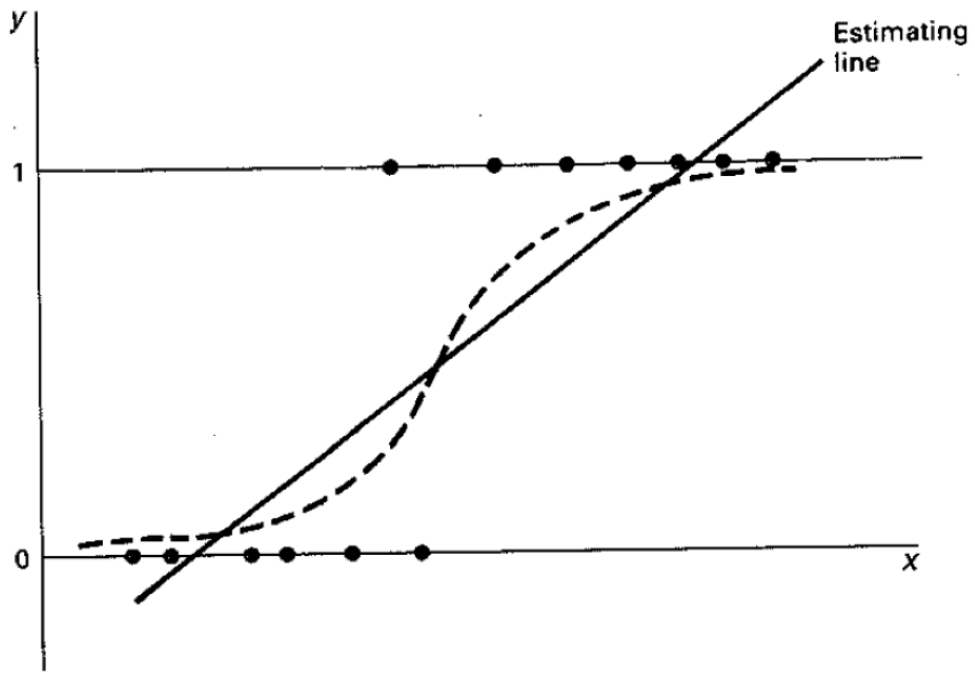
- Ces deux problèmes ne sont pas insurmontables:
  - Changer la valeur des valeurs prédites
    - 0 pour toutes les valeurs négatives
    - 1 pour toutes les valeurs supérieures à 1
- Estimation en contrôlant l'hétéroscédasticité

Modèle Logit / Probit



- Ce qu'il faut, c'est un moyen de "presser" les probabilités estimées à l'intérieur de l'intervalle 0-1

Modèle Logit / Probit



- Ce qu'il faut, c'est un moyen de "presser" les probabilités estimées à l'intérieur de l'intervalle 0-1
- $P(Y = 1) = G(Y_i = \beta_0 + \beta_1 X_i + ... + \beta_k X_k)$

Modèle Logit / Probit

- $P(Y = 1) = G(\beta_0 + \beta_1 X_i + ... + \beta_k X_k)$
- De nombreuses fonctions sont disponibles
- Les deux plus populaires sont:
  - La fonction normale cumulative qui donne le modèle probit
  - La fonction logistique qui donne le modèle logit
- Le modèle log-log complémentaire pour la distribution non symétrique
  - Pour les phénomènes rares, où la probabilité de succès est faible

Modèle Logit / Probit : Formulation

- Forme latente
- $Y_i^* = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} + \epsilon_i$, i allant de 1 à n
- On observe :
  - $Y_i = 1$ si $Y_i^* > 0$
  - $Y_i = 0$ si $Y_i^* < 0$
- $P(Y_i = 1) = P(Y_i^* > 0) = P(X\beta + \epsilon > 0)$

Modèle Logit / Probit : Formulation

- Forme latente
- $Y_i^* = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} + \epsilon_i$, i allant de 1 à n
- On observe :
  - $Y_i = 1$ si $Y_i^* > 0$
  - $Y_i = 0$ si $Y_i^* < 0$
- $P(Y_i = 1) = P(Y_i^* > 0) = P(X\beta + \epsilon > 0)$
- $P(Y_i = 1) = P(\epsilon > -X\beta)$

Modèle Logit / Probit : Formulation

- Forme latente
- $Y_i^* = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} + \epsilon_i$, i allant de 1 à n
- On observe :
  - $Y_i = 1$ si $Y_i^* > 0$
  - $Y_i = 0$ si $Y_i^* < 0$
- $P(Y_i = 1) = P(Y_i^* > 0) = P(X\beta + \epsilon > 0)$
- $P(Y_i = 1) = P(\epsilon > -X\beta)$
- $P(Y_i = 1) = P(\epsilon < X\beta) = \psi(Y_i^*)$

Modèle Logit / Probit : Formulation

- Forme latente
- $Y_i^* = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} + \epsilon_i$, i allant de 1 à n
- On observe :
  - $Y_i = 1$ si $Y_i^* > 0$
  - $Y_i = 0$ si $Y_i^* < 0$
- $P(Y_i = 1) = P(Y_i^* > 0) = P(X\beta + \epsilon > 0)$
- $P(Y_i = 1) = P(\epsilon > -X\beta)$
- $P(Y_i = 1) = P(\epsilon < X\beta) = \psi(Y_i^*)$
- Où ψ est la fonction de distribution cumulative

Modèle Logit / Probit : Estimation

- Les techniques du maximum de vraisemblance sont utilisées pour estimer les paramètres
- Pour chaque observation, la probabilité d'observation Y conditionnelle à X peut s'écrire:
- $P(Y_i = y_i|X) = \psi(x_i\beta)^{y_i}(1 - \psi(x_i\beta))^{1-y_i}$ avec $y_i = 0$ ou 1
- Le logarithme de la vraisemblance de l'observation i peut s'écrire:
- $l_i(\beta) = y_i log[\psi(x_i\beta)] + (1 - y_i) log[(1 - \psi(x_i\beta)]$
- Et la vraisemblance de l'échantillon vaut:

$$L(\beta) = \sum l_i(\beta)$$

Modèle Logit / Probit

Logit

- La fonction de distribution cumulative est la fonction logistique:
- $\psi(t) = \frac{exp(t)}{1+exp(t)}$
- $P(Y = 1|x) = \pi_i = \frac{exp(x\beta)}{1+exp(x\beta)}$
- $logit(\pi_i) = Log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki}$

Modèle Logit / Probit

Probit

- La fonction de densité cumulée est la fonction normale:
- $G(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{1}{2}x^2} dx$
- $P(Y = 1) = G(X\beta)$
- $G^{-1}[P = 1] = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki}$

Modèle Logit / Probit

- Choix entre les deux modèles:
  - Les deux fonctions sont très similaires
  - Le choix est une question de goût en raison de la disponibilité du logiciel
  - Logit populaire dans la santé publique tandis que probit est plus populaire parmi les économistes
  - Logit facilement manipulable : popularisé par la notion de rapport de chances (odd ratio)

Interprétation

Modèle Logit / Probit

Logit

- $Log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki}$
- $d(Log(\frac{\pi_i}{1-\pi_i})/d(X_{1i}))$ donne β_1
- On peut démontrer que:

$$d(\pi_i)/d(X_1) = \beta_1 \pi_i (1 - \pi_i)$$

- β_1 n'explique pas le changement de probabilité dû à un changement d'unité dans la variable X_1

Modèle Logit / Probit

Logit

- $Log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki}$
- $d(Log(\frac{\pi_i}{1-\pi_i})/d(X_{1i}))$ donne β_1
- On peut démontrer que:

$$d(\pi_i)/d(X_1) = \beta_1 \pi_i (1 - \pi_i)$$

- β_1 n'explique pas le changement de probabilité dû à un changement d'unité dans la variable X_1
- $d(\pi_i)/d(X_1)$ dépend de la valeur des autres variables du modèle

Modèle Logit / Probit

Logit

- $Log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki}$
- $d(Log(\frac{\pi_i}{1-\pi_i})/d(X_{1i}))$ donne β_1
- On peut démontrer que:

$$d(\pi_i)/d(X_1) = \beta_1 \pi_i (1 - \pi_i)$$

- β_1 n'explique pas le changement de probabilité dû à un changement d'unité dans la variable X_1
- $d(\pi_i)/d(X_1)$ dépend de la valeur des autres variables du modèle
- Pour interpréter l'effet de X_1, il faut aussi fixer $\pi_i(1 - \pi_i)$

Modèle Logit / Probit

Probit

- $P(Y = 1) = \pi_i = G(X\beta)$
- $d(\pi_i)/d(X_1) = \beta_1 G'(X\beta)$
- Plus difficile que le modèle logit

Modèle Logit / Probit

Logit : Interprétation alternative

- Odd ratio ou rapport de chances

Modèle Logit / Probit

Logit : Interprétation alternative

- Odd ratio ou rapport de chances
- $Log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki}$

Modèle Logit / Probit

Logit : Interprétation alternative

- Odd ratio ou rapport de chances
- $Log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki}$
- $\frac{\pi_i}{1-\pi_i} = exp(\beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki})$

Modèle Logit / Probit

Logit : Interprétation alternative

- Odd ratio ou rapport de chances
- $Log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki}$
- $\frac{\pi_i}{1-\pi_i} = exp(\beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki})$
- $\frac{\pi_i}{1-\pi_i} = exp(\beta_0) exp(\beta_1 X_{1i}) * ... * exp(\beta_k X_{ki})$

Modèle Logit / Probit

Logit : Interprétation alternative

- Odd ratio ou rapport de chances
- $Log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki}$
- $\frac{\pi_i}{1-\pi_i} = exp(\beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki})$
- $\frac{\pi_i}{1-\pi_i} = exp(\beta_0) exp(\beta_1 X_{1i}) * ... * exp(\beta_k X_{ki})$
- C'est ce qu'on appelle une chance ou une côte

Modèle Logit / Probit

Logit : Interprétation alternative

- Odd ratio ou rapport de chances
- $Log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki}$
- $\frac{\pi_i}{1-\pi_i} = exp(\beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki})$
- $\frac{\pi_i}{1-\pi_i} = exp(\beta_0) exp(\beta_1 X_{1i}) * ... * exp(\beta_k X_{ki})$
- C'est ce qu'on appelle une chance ou une côte
- Si vous faites varier X_1 de 0 à 1, vous pouvez calculer le rapport de cette côte.

Autres éléments à considérer

Violation des hypothèses

- Toutes violations d'hypothèses comme dans le cas d'un modèle linéaire affectent les estimations et leurs erreurs standards
  - Variables omises
  - Hétéroscédasticité
  - Multicolinéarité
  - Endogénéité...

Mesure de la "qualité d'ajustement"

- En régression linéaire, ce rôle est joué par R^2 ou $pseudo - R^2$.
- R^2 ou pseudo R^2 ne conviennent pas dans le cas de modèle logit ou probit

Alternative:

- Tableau calculant le nombre de valeurs Y = 1 correctement et incorrectement prédites et le nombre de valeurs Y = 0 correctement et incorrectement prédites
- Une observation est prédite comme Y = 1 si la probabilité estimée dépasse une valeur fixe (souvent la moitié)
- Doit être utilisé avec prudence

Test d'hypothèses

- Test d'hypothèse d'un paramètre
  - t Student est utilisé dans le cas de la modélisation logit
  - la statistique z est utilisée dans le cas de la modélisation probit
- Test d'hypothèse de nombreuses paramètres

Test d'hypothèses

- Test d'hypothèse d'un paramètre
  - t Student est utilisé dans le cas de la modélisation logit
  - la statistique z est utilisée dans le cas de la modélisation probit
- Test d'hypothèse de nombreuses paramètres
- Test du rapport de vraisemblance (LR) (test de Fischer en cas de régression linéaire)

Test d'hypothèses

- Test d'hypothèse d'un paramètre
  - t Student est utilisé dans le cas de la modélisation logit
  - la statistique z est utilisée dans le cas de la modélisation probit
- Test d'hypothèse de nombreuses paramètres
- Test du rapport de vraisemblance (LR) (test de Fischer en cas de régression linéaire)
- $LR = -2[L(RM) - L(UM)]$ suit une loi de chi-deux à m dégrés de liberté

Test d'hypothèses

- Test d'hypothèse d'un paramètre
  - t Student est utilisé dans le cas de la modélisation logit
  - la statistique z est utilisée dans le cas de la modélisation probit
- Test d'hypothèse de nombreuses paramètres
- Test du rapport de vraisemblance (LR) (test de Fischer en cas de régression linéaire)
- $LR = -2[L(RM) - L(UM)]$ suit une loi de chi-deux à m dégrés de liberté
- Où RM et UM sont respectivement le modèle restreint et le modèle non restreint, m est le nombre de contraintes

Test d'hypothèses

- Modèle non contraint

Probit regression

Number of obs	=	9793
LR chi2(9)	=	43.87
Prob > chi2	=	0.0000
Pseudo R2	=	0.0094

Log likelihood = -2315.8856

dead	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
twin	.4499684	.0961671	4.68	0.000	.2614843	.6384525
female	-.0207827	.039677	-0.52	0.600	-.0985482	.0569827
age15_19	.0988493	.0609413	1.62	0.105	-.0205935	.218292
age35_49	-.0973036	.0776271	-1.25	0.210	-.2494499	.0548428
parity1	.1320236	.0611795	2.16	0.031	.012114	.2519333
parity6	.0330353	.0631601	0.52	0.601	-.0907563	.1568269
bambara	.0041792	.0416883	0.10	0.920	-.0775283	.0858867
primary	-.0478053	.0558566	-0.86	0.392	-.1572823	.0616716
secondary	-.149802	.0590777	-2.54	0.011	-.2655921	-.0340119
_cons	-1.544451	.0393228	-39.28	0.000	-1.621523	-1.46738

Test d'hypothèses

- Test de nombreuses hypothèses
- Exemple: mortalité infantile
- Modèle sans restriction: probit mort jumelle femelle age15_19 age35_49 parity1 parity6 bambara primaire secondaire
  - L(UM) = -2315,8856

Test d'hypothèses

- Modèle contraint

Probit regression

Number of obs	= 9793
LR chi2(7)	= 37.13
Prob > chi2	= 0.0000
Pseudo R2	= 0.0079

Log likelihood = -2319.2534

dead	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
twin	.4499418	.0960745	4.68	0.000	.2616393	.6382443
female	-.0218886	.0396241	-0.55	0.581	-.0995504	.0557732
age15_19	.1209075	.059881	2.02	0.043	.003543	.2382721
age35_49	-.0886136	.0776209	-1.14	0.254	-.2407478	.0635205
parity1	.1046958	.0600342	1.74	0.081	-.0129691	.2223608
parity6	.0505236	.0628512	0.80	0.421	-.0726625	.1737097
bambara	.008781	.0416145	0.21	0.833	-.0727818	.0903439
_cons	-1.579112	.0366058	-43.14	0.000	-1.650858	-1.507366

Test d'hypothèses

- Modèle restreint: Modèle restreint: probit mort jumeau femelle age15_19 age35_49 parity1 parity6 bambara
  - L(RM) = -2319,2534
- m = 2

Test d'hypothèses

- LR = -2*(-2319,2534 +2315,8856) = 6,74

Test d'hypothèses

- LR = -2*(-2319,2534 +2315,8856) = 6,74
- Chi-deux lu = 5,99 pour un niveau de signification de 5%

Test d'hypothèses

- LR = -2*(-2319,2534 +2315,8856) = 6,74
- Chi-deux lu = 5,99 pour un niveau de signification de 5%
- Conclusion: Nous rejetons l'hypothèse nulle. L'éducation a un effet significatif sur la mortalité infantile.

Extension

- La variable dépendante comprend plus de deux catégories:
  - Probit / logit ordonné
  - Il existe un classement clair entre les modalités
  - Ex. Quintile de richesse
- Logit / probit multinomial
  - Pas d'ordre, mais groupe distinct
  - Ex. Statut de travail (non, formel, informel)