



Chapitre 3

Les informations
qualitatives

Dans les modèles économétriques, on distingue traditionnellement 2 types de variables explicatives :

- les variables quantitatives
- les variables qualitatives : utilisées pour signaler une caractéristique non mesurable.



Les variables qualitatives introduites dans les modèles doivent toujours être binaires (indicatrices)

Bien évidemment, ces deux types de variables peuvent être simultanément présents dans un même modèle

3. I Les variables indicatrices

Une variable *indicatrice (dummy)* est une variable qui ne prend que deux valeurs possibles : 1 ou 0.

On peut également parler de *variable binaire*.

Exemples :

- sexe : 0 pour les hommes, 1 pour les femmes
- *statut matrimonial* : 1 si indiv. marié, 0 sinon
- *lieu de résidence* : 1 pour Ile-de-France, 0 sinon
- *actif* : 1 si en emploi, 0 sinon

3. I Les variables indicatrices

Considérons un modèle simple avec une variable continue (x) et une variable indicatrice (d)

$$y = \beta_0 + \beta_1 \cdot x + \delta_0 \cdot d + u$$

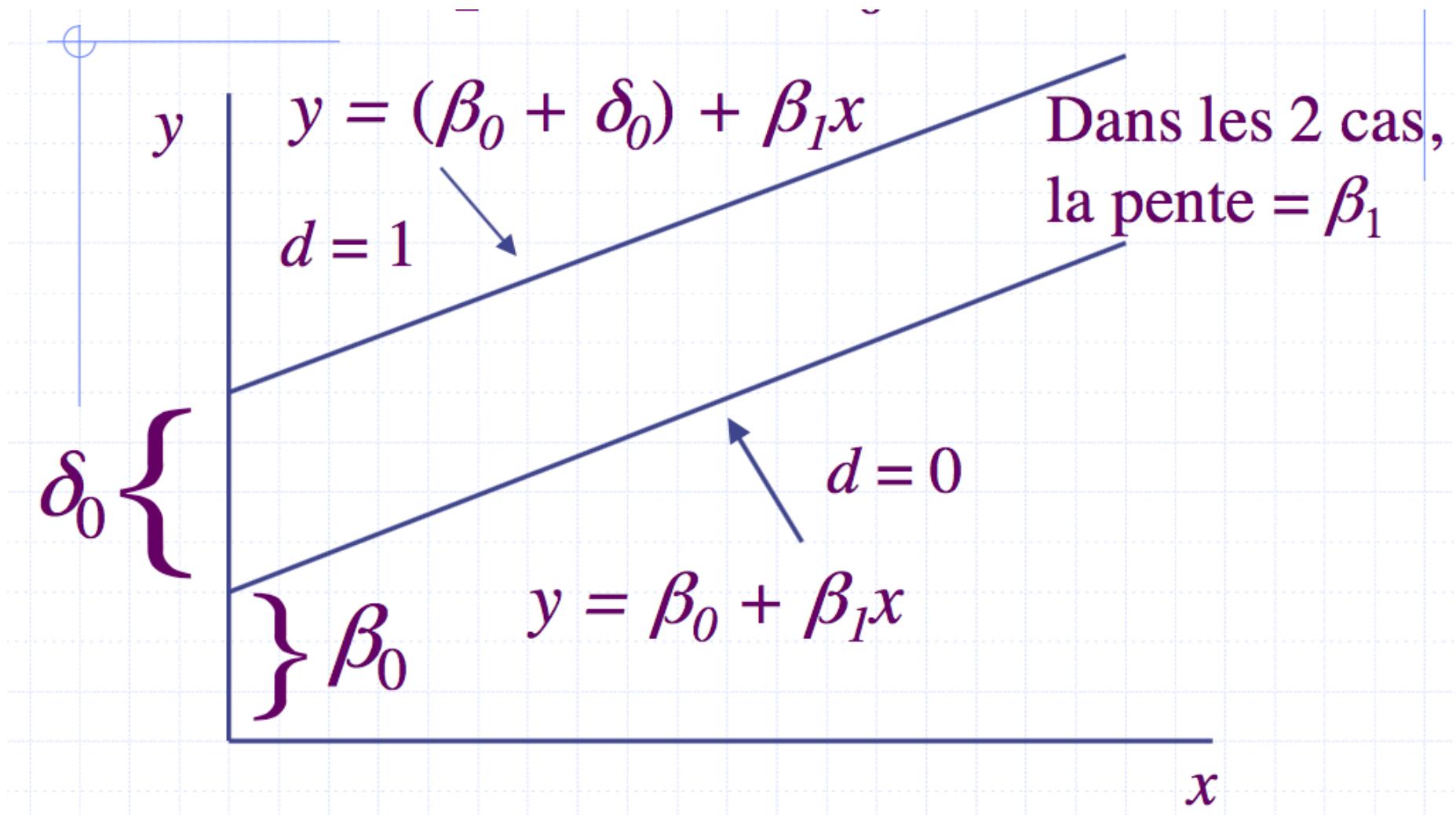
Le coefficient (MCO) de la variable d peut être interprété comme une translation du modèle (modification du terme constant)

- si $d = 0, y = \beta_0 + \beta_1 \cdot x + u$
- si $d = 1, y = \beta_0 + \beta_1 \cdot x + \delta_0 + u = (\beta_0 + \delta_0) + \beta_1 \cdot x + u$

Le cas $d = 0$ est **le groupe de référence**

3. I Les variables indicatrices

Si $\delta_0 > 0$

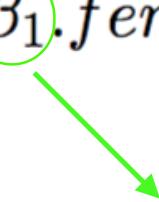


3. I Les variables indicatrices

Exemple : Supposons que l'on pense que les salaires des femmes tendent à être inférieurs à ceux des hommes à niveau d'éducation donné.

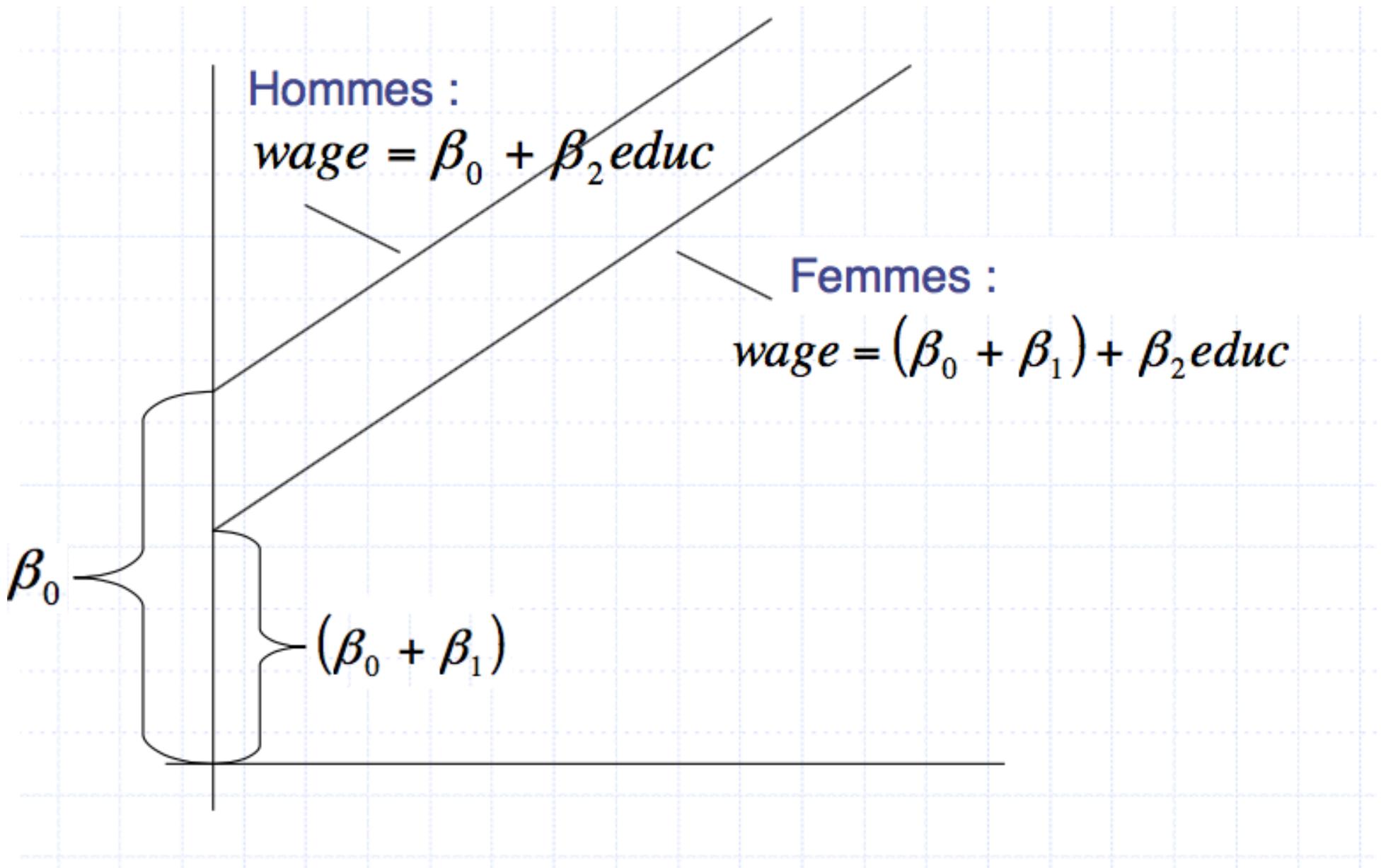
Cette idée pourra simplement être modélisée en considérant une variable *female* valant 1 si l'individu est une femme et 0 sinon.

$$wage = \beta_0 + \beta_1 \cdot female + \beta_2 \cdot educ + u$$



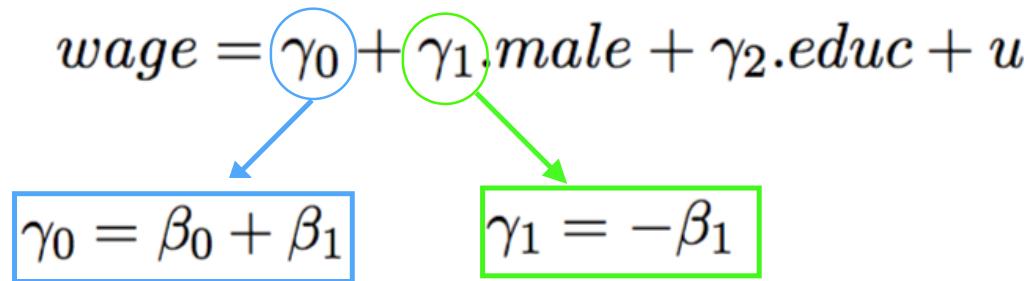
$$\begin{aligned} &= E(wage | female = 1, educ) - E(wage | female = 0, educ) \\ &= E(wage | female, educ) - E(wage | male, educ) \end{aligned}$$

3. I Les variables indicatrices



3. I Les variables indicatrices

Le même modèle aurait pu être estimé en introduisant une variable dichotomique *male* qui vaut 1 si l'indiv. est un homme et 0 sinon (les femmes auraient alors été le groupe de référence)

$$wage = \gamma_0 + \gamma_1 \cdot male + \gamma_2 \cdot educ + u$$

$$\gamma_0 = \beta_0 + \beta_1$$
$$\gamma_1 = -\beta_1$$

On ne doit pas inclure conjointement les variables *female* et *male* dans la régression. Sinon, on aurait un problème de colinéarité parfaite.

```
. reg incearn female education tenure businesses
```

Source	SS	df	MS	Number of obs	=	994
Model	3.6343e+09	4	908569775	F(4, 989)	=	60.06
Residual	1.4962e+10	989	15128816.2	Prob > F	=	0.0000
Total	1.8597e+10	993	18727772.7	R-squared	=	0.1954

Adj R-squared = 0.1922
Root MSE = 3889.6

incearn	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
female	-862.1825	251.8908	-3.42	0.001	-1356.484 -367.8808
education	376.4086	40.88503	9.21	0.000	296.1772 456.64
tenure	12.54285	1.332865	9.41	0.000	9.927278 15.15841
businesses	1638.655	335.5927	4.88	0.000	980.0993 2297.211
_cons	-2010.119	568.4594	-3.54	0.000	-3125.644 -894.5934

```
. gen male = 1- female
```

```
. reg incearn male education tenure businesses
```

Source	SS	df	MS	Number of obs	=	994
Model	3.6343e+09	4	908569775	F(4, 989)	=	60.06
Residual	1.4962e+10	989	15128816.2	Prob > F	=	0.0000
Total	1.8597e+10	993	18727772.7	R-squared	=	0.1954

Adj R-squared = 0.1922
Root MSE = 3889.6

incearn	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
male	862.1825	251.8908	3.42	0.001	367.8808 1356.484
education	376.4086	40.88503	9.21	0.000	296.1772 456.64
tenure	12.54285	1.332865	9.41	0.000	9.927278 15.15841
businesses	1638.655	335.5927	4.88	0.000	980.0993 2297.211
_cons	-2872.301	541.9166	-5.30	0.000	-3935.74 -1808.863

```
. display -2010.119 - 862.1825
-2872.3015
```

```
. display -2872.301 + 862.1825
-2010.1185
```

peu importe le groupe mis en référence : on aboutira aux mêmes conclusions « littéraires »

3.2 Variables indicatrices et catégories multiples

© Théo Jalabert



On doit utiliser des variables indicatrices pour contrôler une information discrète multiple.

Exemple : prenons l'exemple de la variable *diplôme le plus élevé*, codée initialement comme suit :

- 0 : Aucun diplôme
- 1 : BEPC
- 2 : BEP-CAP
- 3 : Bac
- 4 : Bac+2 ou plus

On va créer 4 variables indicatrices et mettre une catégorie en référence

Toute variable catégorielle peut (et doit) être transformée en un jeu de variables indicatrices.

Le groupe de référence est représenté par le terme constant.

Si la variable catégorielle contient n catégories, on va introduire seulement ($n-1$) variables indicatrices.

On peut rapidement se retrouver avec un nombre important de variables. D'où des regroupements de catégories (ex: CSP, secteurs d'activité, etc.) (Règle pratique des « 5% »).

Exemple : équation de salaire

Tableau 6 : équation de salaire pour les emplois occupés en 2000 : influence des formations structurée et informelle

Variables	Paramètre	P. value
Constante	1,900	0,000
Sexe (homme=1)	0,145	0,000
Probabilité de suivre une formation structurée	0,394	0,000
Probabilité de suivre une formation structurée*sexé	-0,055	-0,535
Probabilité de suivre une formation en cours d'emploi	0,264	-0,072
Probabilité de suivre une formation en cours d'emploi*sexé	0,041	-0,751
Probabilité d'obtenir une promotion	0,159	-0,004
Probabilité d'obtenir une promotion*sexé	0,016	-0,831
Marié	0,084	0,000
En couple	0,084	0,000
Enfants à charges	0,000	-0,998
Ancienneté avec l'employeur	0,001	0,000
Ancienneté au carré /100	0,000	-0,110
Expérience à temps plein	0,019	0,000
Expérience au carré/100	-0,032	0,000
Niveau de scolarité (Réf. = Pas de diplômes d'études secondaires)		
- Diplômes d'études secondaires	0,076	0,000
- Certificat	0,108	0,000
- Diplôme collégial	0,118	0,000
- Diplôme universitaire	0,259	0,000
Heures travaillées		
- Temps partiel	0,017	-0,324
- Horaire de travail flexible	0,014	-0,259
- Travail entre 6h et 18h	-0,013	-0,444
- Heures habituelles non rémunérées	0,008	0,000
Statut d'emploi		
- Emploi permanent	0,002	-0,926
- Emploi de supervision	0,042	-0,001
- Emploi syndiqué	0,049	-0,002
- Procédure d'évaluation du rendement	-0,018	-0,186
Etablissement à but non lucratif	-0,007	-0,760
Etablissement à propriété étrangère	0,064	-0,002

Profession (Réf. = Personnel technique/métiers)		
- Gestionnaires	0,187	0,000
- Professionnels	0,176	0,000
- Commercialisation ou ventes	-0,092	-0,003
- Personnel de bureau	-0,077	0,000
- Personnel non qualifié	-0,109	0,000
Branche d'activité (Réf. = Commerce de détail)		
- Exploitation des ressources naturelles	0,397	0,000
- Industries de la fabrication	0,245	0,000
- Constructions	0,359	0,000
- Transport, entreposage et commerce de gros	0,233	0,000
- Communications et autres services publics	0,270	0,000
- Finance et assurances	0,223	0,000
- Services immobiliers et de location	0,274	0,000
- Services aux entreprises	0,251	0,000
- Enseignement et services de soins de santé	0,213	0,000
- Information et industries culturelles	0,264	0,000
Taille de l'établissement (Réf. = 500 employés et plus)		
- Moins de 20 employés	-0,124	0,000
- Entre 20 et 99 employés	-0,147	0,000
- Entre 100 et 499 employés	-0,078	0,000
Région (Réf. = Ontario)		
- Colombie-Britannique	0,024	-0,282
- Alberta	-0,090	0,000
- Provinces des Prairies	-0,153	0,000
- Québec	-0,036	-0,174
- Provinces de l'Atlantique	-0,223	0,000
<i>R</i> ²	54,91%	
Nombre d'observations	18 870	

On dispose parfois d'informations qualitatives ordonnées (niveaux de satisfaction, classements, niveaux de risque, etc.). Or, souvent une augmentation d'une unité n'a pas de raison d'avoir un effet constant. D'où le recours aux variables indicatrices.

Exemples :

Niveau de satisfaction :

- 1 : très faible
- 2 : faible
- 3 : moyen
- 4 : élevé
- 5 : très élevé

Classement de clients par niveaux de risques pour une assurance :

- 1 : risque très faible
- 2 : risque faible
- 3 : risque incertain
- 4 : risque élevé
- 5 : risque très élevé

On peut parfois créer des variables indicatrices à partir de variables quantitatives pour capter des non-linéarités.

Exemples :

- classes d'âge
- classes de revenu
- niveaux d'éducation construits à partir d'un nombre d'années

Cela permet des spécifications plus flexibles mais au prix d'un nombre de paramètres à estimer plus important.

3.3 Variables indicatrices et interactions

L'utilisation qui vient d'être faite des variables indicatrices peut naturellement se généraliser à tous les coefficients d'une régression : ce n'est pas nécessairement le terme constant qui seul se modifie lorsque l'on considère des sous-populations d'individus.

L'utilisation de termes croisés entre une variable continue et une variable indicatrice va permettre de capter des différences de pentes entre groupes.

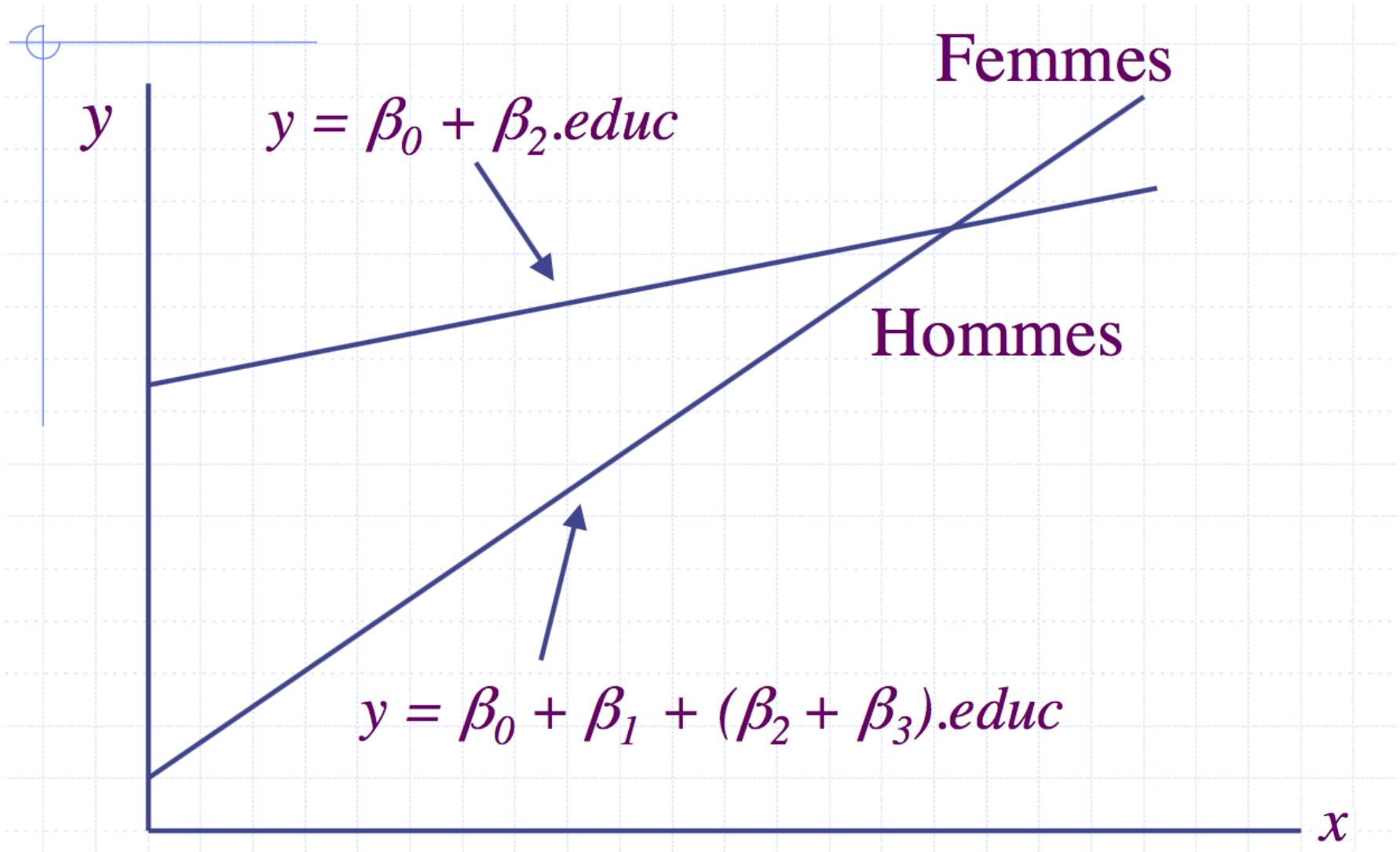
En pratique, lorsque l'on suppose que le coefficient d'une variable varie selon une caractéristique propre à une sous-partie de la population étudiée, il suffit de construire une variable indicatrice de sous-population et d'ajouter à la liste des variables explicatives de départ la variable obtenue en multipliant pour chaque individu les valeurs de la variable indicatrice et de la variable dont le coefficient est supposé varier.

$$wage = \beta_0 + \beta_1.female + \beta_2.educ + \beta_3.female.educ + u$$

rendement de
l'éducation des hommes

$\beta_2 + \beta_3$
rendement de
l'éducation des femmes

Tester la significativité de β_3 revient à tester l'égalité des rendements de l'éducation entre hommes et femmes.



Exemple :

Distance domicile-travail

- SEXE=1 pour homme et 0 pour femme ; tps=1 pour temps plein et 0 pour temps partiel ; age2 = âge au carré ;
- Les CSP : employé =1 si employé, 0 sinon ; profint=1 si profession intermédiaire et 0 sinon ; cadplib=1 si cadres et professions libérales et 0 sinon ; Autres = Autres CSP, 0 sinon. Pour toutes les CSP, la référence est Ouvrier.
- *Permis_disp2* est une indicatrice qui vaut 1 si l'individu a le permis mais pas de voiture à disposition et 0 sinon ; *permis_disp3* vaut 1 si l'individu a le permis et une voiture à sa disposition et 0 sinon (La référence pour ces deux variables est un individu qui n'a pas le permis).
- ZUS = 1 si l'individu habite dans un quartier ZUS et 0 sinon ;
- DistCentreVo : distance au centre de son lieu d'habitation à vol d'oiseau ;
- Lieu d'habitation : *ZoneHGL* =1 si l'individu habite hors Grand Lyon et 0 sinon ; *GL_Metro* : vaut 1 si l'individu habite dans un quartier du Grand Lyon ayant une station de métro et 0 sinon ; *GL_smetro* = 1 si l'individu habite dans un quartier du Grand Lyon mais sans station de métro et 0 sinon ; la référence pour ces variables est HyperCentre.

Source	SS	df	MS	Number of obs	= 6576
Model	724160.999	16	45260.0624	F(16, 6559)	= 84.65
Residual	3507037.13	6559	534.690826	Prob > F	= 0.0000
Total	4231198.13	6575	643.528232	R-squared	= 0.1711 = $\frac{SCE}{SCT} = \frac{724160.999}{4231198.13}$

Adj R-squared = 0.1691 = $1 - \frac{SCE/(T-k)}{SCT(T-1)} = 1 - \frac{534.690826}{643.528232}$

Root MSE = 23.123

DistTotale~v	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
1.SEXE	7.61441	.7018908	10.85	0.000	6.238476 8.990345
1.ZUS	1.100919	1.177947	0.93	0.350	-1.20824 3.410078
SEXE#ZUS					
1 1	-3.649904	1.51964	-2.40	0.016	-6.628892 -.6709151
tps	2.30811	.8691691	2.66	0.008	.6042557 4.011965
AGE	.4205017	.2084309	2.02	0.044	.0119092 .8290941
age2	-.0059504	.0025392	-2.34	0.019	-.010928 -.0009727
employe	-.8346921	.9022544	-0.93	0.355	-2.603405 .9340204
profint	3.949531	.8694305	4.54	0.000	2.245164 5.653898
cadplib	6.619653	.9123529	7.26	0.000	4.831144 8.408162
Autres	2.765077	1.581004	1.75	0.080	-.3342062 5.86436
permis_disp2	.2807874	1.576125	0.18	0.859	-2.80893 3.370505
permis_disp3	6.662492	1.364884	4.88	0.000	3.987033 9.337951
zoneHGL	8.332358	2.762881	3.02	0.003	2.916211 13.74851
GL_metro	-2.540658	.9495929	-2.68	0.007	-4.402169 -.6791464
GL_smetro	-6.215189	2.489729	-2.50	0.013	-11.09587 -1.334509
DistCentreV0	.7089818	.0595777	11.90	0.000	.5921902 .8257734
_cons	-1.988831	4.249963	-0.47	0.640	-10.32014 6.342481

Autre exemple : Déterminants de l'indice de masse corporelle

Afin de mieux cibler les politiques de prévention contre l'obésité chez les adultes, on cherche à déterminer les caractéristiques socio-économiques qui influencent l'indice de masse corporelle (IMC). Cet indice ($poids/taille^2$) a été défini en 1997 par l'Organisation Mondiale de la Santé comme le standard pour évaluer les risques liés au surpoids : si un adulte a un IMC supérieur à 25, il est considéré en surpoids et s'il a un IMC supérieur à 30, il est considéré comme obèse. A partir des données de l'*Etude Individuelle Nationale des Consommations Alimentaires* (2006-2007) pour la population adulte en emploi, on a effectué une régression par moindres carrés ordinaires qui avait comme variable à expliquer l'indice de masse corporelle et comme variables explicatives :

- *femme* : variable indicatrice de genre (1 femme, 0 homme)
- *age* : variable continue indiquant l'âge de la personne
- *nbenf* : variable continue indiquant le nombre d'enfants de la personne
- *heurtrav* : variable continue indiquant le nombre d'heures de travail effectuées la semaine précédent l'enquête
- *ordi* : Temps moyen quotidien passé devant un ordinateur (en min)
- *tele* : Temps moyen quotidien passé devant la télé (en min)

et 3 variables croisées : *femme * ordi*, *femme * heurtrav*, *femme * nbenf*. Les résultats de l'estimation sont donnés dans le tableau 1.

Tableau 1

	Coefficient	Ecart-type	t de student
age	.0614782	.0061198	10.05
femme	-.7485824	.4505336	-1.66
nbenf	-.2061546	.0984261	-2.09
heurtrav	.014308	.0068698	2.08
ordi	.0034577	.0012724	2.72
tele	.0035719	.0012852	2.78

Variables croisées

femme*heurtrav	-.0159522	.0093731	-1.70
femme*nbenf	.3532757	.1303062	2.71
femme*ordi	-.0027627	.0016983	-1.63

Obs. 2 584

$R^2=0.3974$

3.4 Test de structures différentes par groupes

© Théo Jalabert 

Pour tester si un modèle est différent entre deux groupes (hommes/femmes, ZUS/non ZUS, etc.), on peut simplement croiser chaque variable avec une variable indicatrice de groupe et tester la significativité jointe des termes croisés.

Or, on peut se retrouver très rapidement avec un nombre considérable de coefficients à estimer.

Une démarche alternative lorsqu'il y a trop de variables explicatives est d'effectuer le test de Chow.

Le test de Chow

On est ici intéressé par une possible modification des coefficients du modèle $Y = X\beta + u$ lorsque l'on passe d'un groupe à l'autre

Exemples :

Sur données temporelles, on cherche à déterminer si les paramètres du modèle se sont modifiés au cours du temps : Est-ce que les coefficients sont identiques avant et après une date de rupture ?

Sur données individuelles, on cherche à déterminer si des groupes d'individus sont homogènes ou hétérogènes : est-ce que les coefficients sont identiques pour les hommes et les femmes ? Pour les petites et les grandes entreprises ?

Il s'agit d'effectuer l'estimation séparément sur les deux groupes et de tester l'égalité des coefficients entre les deux régressions.

Le test de Chow

Le test de Chow se ramène à la question suivante : existe-t-il une différence significative entre la somme des carrés des résidus (SCR) de l'ensemble de l'échantillon et l'addition de la somme des carrés des résidus calculée à partir des deux sous-échantillons (SCR1+SCR2) ? Le fait de scinder l'échantillon en 2 améliore-t-il la qualité du modèle ? Pour faire ce test :

- 1) Estimation du modèle sur chacun des 2 sous-échantillons et calcul des sommes des carrés des résidus.
- 2) Estimation du **même** modèle sur l'ensemble de l'échantillon.
- 3) Calcul de la statistique de test :

$$F^* = \frac{[SCR - (SCR1 + SCR2)]/k}{(SCR1 + SCR2)/(T - 2k)} \sim F(k, T - 2k)$$

Si $F^* > F_\alpha$: on rejette H_0

=> structure différente entre les groupes

. reg lincearn education tenure businesses

© Théo Lalabert

Source	SS	df	MS	Number of obs = 994
Model	2228.68624	3	2579.56225	F(3, 990) = 98.61
Residual	25897.8863	990	26.1594811	Prob > F = 0.0000
Total	33636.573	993	33.8736888	R-squared = 0.2301 Adj R-squared = 0.2277 Root MSE = 5.1146

$k=4$

lincearn	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
education	.3354688	.0535933	6.26	0.000	.2302992 .4406384
tenure	.0246301	.0017343	14.20	0.000	.0212268 .0280334
businesses	3.481862	.4378783	7.95	0.000	2.622586 4.341139
_cons	-1.71791	.7070579	-2.43	0.015	-3.105414 -.3304055

. reg lincearn education tenure businesses if female == 1

Source	SS	df	MS	Number of obs = 501
Model	4247.68801	3	1449.22934	F(3, 497) = 50.84
Residual	14167.5286	497	28.5060938	Prob > F = 0.0000
Total	18515.2166	500	37.0304333	R-squared = 0.2348 Adj R-squared = 0.2302 Root MSE = 5.3391

lincearn	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
education	.4503757	.0840316	5.36	0.000	.2852746 .6154767
tenure	.029138	.0030001	9.71	0.000	.0232434 .0350325
businesses	3.553221	.7238004	4.91	0.000	2.131135 4.975307
_cons	-3.839609	1.07801	-3.56	0.000	-5.957628 -1.721589

. reg lincearn education tenure businesses if female == 0

Source	SS	df	MS	Number of obs = 493
Model	2823.06208	3	957.689027	F(3, 489) = 41.89
Residual	11180.7683	489	22.8645568	Prob > F = 0.0000
Total	14053.8353	492	28.564706	R-squared = 0.2044 Adj R-squared = 0.1996 Root MSE = 4.7817

lincearn	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
education	.209006	.0676879	3.09	0.002	.076011 .3420009
tenure	.0204169	.0020748	9.84	0.000	.0163403 .0244935
businesses	3.128716	.5359526	5.84	0.000	2.075662 4.181771
_cons	.7811232	.9226064	0.85	0.398	-1.031639 2.593885

$$f^* = \frac{25897.8863 - (14167.5286 + 11180.7683)}{(14167.5286 + 11180.7683)} * ((994 - 2 * 4) / 4)$$

$$5.3444927$$

```

. gen femeducation = female * education
. gen femtenure = female * tenure
. gen fembusinesses = female*businesses
. reg lincearn female education femeducation tenure femtenure businesses fembusinesses

```

Source	SS	df	MS	Number of obs	=	994
Model	8288.27612	7	1184.03945	F(7, 986)	=	46.06
Residual	25348.2969	986	25.7082119	Prob > F	=	0.0000
Total	33636.573	993	33.8736888	R-squared	=	0.2464

lincearn	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
female	-4.620732	1.41602	-3.26	0.001	-7.399491	-1.841973
education	.209006	.0717737	2.91	0.004	.0681592	.3498527
femeducation	.2413697	.1073299	2.25	0.025	.0307484	.451991
tenure	.0204169	.0022	9.28	0.000	.0160996	.0247341
femtenure	.0087211	.0035997	2.42	0.016	.0016572	.015785
businesses	3.128716	.5683042	5.51	0.000	2.013492	4.243941
fembusines^s	.4245046	.8918728	0.48	0.634	-1.325682	2.174691
_cons	.7811232	.9782977	0.80	0.425	-1.138662	2.700908

```
. test female femeducation femtenure fembusinesses
```

```

< 1> female = 0
< 2> femeducation = 0
< 3> femtenure = 0
< 4> fembusinesses = 0

F( 4, 986) = 5.34
Prob > F = 0.0003

```

3.5 Une variable indicatrice comme variable expliquée

Lorsque la variable à expliquer est une indicatrice, l'approximation linéaire est peu adaptée comme modélisation.

On aura recours à des modèles dichotomiques/binaires : *probits et logits* (simples)

Il s'agit généralement d'expliquer « la survenue ou la non-survenue d'un événement » en fonction d'un certain nombre de caractéristiques observées

$$y_i = \begin{cases} 1 & \text{si l'événement s'est réalisé pour l'individu } i \\ 0 & \text{si l'événement ne s'est pas réalisé pour l'individu } i \end{cases}$$

Exemples :

Idée originelle :

les médecins administraient aux patients une dose de médicaments et ils observaient si les patients supportaient bien ou non cette dose.

$$y_i = \begin{cases} 1 & \text{si le patient } i \text{ supporte bien la dose} \\ 0 & \text{si le patient } i \text{ ne supporte pas la dose} \end{cases}$$

L'expérience dépend des caractéristiques du patient X_i

Exemples :

Autres applications :

- étude des choix d'éducation [Radner et Miller, 1970]
- modélisation des risques de défaillance dans une relation de prêt, ou dans toute autre forme de contrat d'engagement (bases des méthodes de scoring en finance, assurance, télécommunications)

Modélisations :

Les modèles dichotomiques admettent pour variable expliquée, non pas un codage quantitatif associé à la réalisation d'un événement (comme dans le cas de régression linéaire) mais la **probabilité d'apparition de cet événement**, conditionnellement aux variables explicatives. Ainsi, on considère le modèle suivant :

$$p_i = \text{Prob} (y_i = 1 | x_i) = F(x_i \beta)$$

où $F(\cdot)$ désigne une fonction de répartition.

Les 2 lois les plus couramment utilisées dans la pratique sont la loi logistique (logit) et la loi normale (probit).

Modélisations :

Fonctions de répartition

loi logistique

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

loi normale centrée, réduite

$$\Phi(u) = \int_{-\infty}^u \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

Modèles

logit

$$p_i = \Lambda(x_i \beta) = \frac{1}{1 + e^{-x_i \beta}}$$

probit

$$p_i = \Phi(x_i \beta) = \int_{-\infty}^{x_i \beta} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$

=> il faut trouver un estimateur de β

Technique d'estimation :

Estimation par maximum de vraisemblance et non par moindres carrés ordinaires.

Si les données sont indépendamment distribuées (iid), la fonction de vraisemblance s'écrit comme le produit des probabilités individuelles :

$$p_i = P(y_i=1|x_i) = F(x_i\beta)$$

$$L(y, \beta) = \prod_{i=1}^N p_i^{y_i} (1 - p_i)^{1-y_i} = \prod_{i=1}^N [F(x_i\beta)]^{y_i} [1 - F(x_i\beta)]^{1-y_i}$$

La log-vraisemblance

$$\ln L(y, \beta) = \sum_{i=1}^N y_i \cdot \ln[F(x_i\beta)] + (1 - y_i) \cdot [1 - F(x_i\beta)]$$

L'estimateur du maximum de vraisemblance vérifie :

$$\sum_{i=1}^N y_i \frac{f(x_i\beta)}{F(x_i\beta)} x'_i + (y_i - 1) \frac{f(x_i\beta)}{1 - F(x_i\beta)} x'_i = 0$$

soit après simplification :

$$\sum_{i=1}^N \frac{[y_i - F(x_i\beta)] f(x_i\beta)}{F(x_i\beta) [1 - F(x_i\beta)]} x'_i = 0$$

Le vecteur des paramètres ne peut pas être estimé directement, d'où le recours à un algorithme d'optimisation.

Interprétation des résultats :

Les paramètres β ne sont identifiés qu'à une constante additive et une constante multiplicative près.

La valeur des coefficients n'est pas interprétable directement.

La seule information utilisable directement est le signe des paramètres, indiquant si la variable associée influence la probabilité à la hausse ou à la baisse.

Interprétation des résultats :

Pour quantifier les effets propres, on doit calculer les effets marginaux

Effet marginal associé à la j ème variable explicative :

$$\frac{\partial p_i}{\partial x_i^{[j]}} = \frac{\partial F(x_i\beta)}{\partial x_i^{[j]}} = \frac{\partial F(x_i\beta)}{\partial (x_i\beta)} \frac{\partial (x_i\beta)}{\partial x_i^{[j]}} = \frac{\partial F(x_i\beta)}{\partial (x_i\beta)} \beta_j$$

$$\frac{\partial p_i}{\partial x_i^{[j]}} = f(x_i\beta) \cdot \beta_j$$

n'est pas constant
souvent évalué au point moyen

avec $x_i\beta = \sum_{k=1}^K x_i^{[k]} \beta_k$.

Logistic regression

Number of obs = 44705
 © Théo Jalabert
 LR chi2(49) = 6673.06
 Prob > chi2 = 0.0000
 Pseudo R2 = 0.2127



Log likelihood = -12348.826

expo	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
age	-.016916	.0019435	-8.70	0.000	-.0207251 -.0131068
2.sexé	-.7125173	.0508737	-14.01	0.000	-.8122279 -.6128068
1.noct	.273152	.0453531	6.02	0.000	.1842615 .3620426
1.equip	.2273376	.0444861	5.11	0.000	.1401465 .3145288
natd	.0911977	.0936553	0.97	0.330	-.0923634 .2747588
statut123	.1313596	.1320187	1.00	0.320	-.1273923 .3901114
statut4	.0812004	.1140992	0.71	0.477	-.14243 .3048308
statut5	.2208903	.100556	2.20	0.028	.0238041 .4179764
statut6	.1394878	.119092	1.17	0.241	-.0939283 .3729039
ancien					
2	.2640287	.0729058	3.62	0.000	.121136 .4069214
3	.2981609	.072732	4.10	0.000	.1556088 .4407131
4	.4303763	.0779104	5.52	0.000	.2776747 .583078
tetab					
2	-.2803355	.05251	-5.34	0.000	-.3832532 -.1774177
3	-.3319378	.0684258	-4.85	0.000	-.4660499 -.1978256
4	-.4238416	.0800485	-5.29	0.000	-.5807337 -.2669494
5	-.4096693	.0777711	-5.27	0.000	-.5620978 -.2572407
1.tps	.4013243	.0740311	5.42	0.000	.256226 .5464226
fonct					
0	-.5056537	.0595533	-8.49	0.000	-.6223759 -.3889314
2	.6503129	.0482277	13.48	0.000	.5557884 .7448375
3	-.786435	.1059787	-7.42	0.000	-.9941494 -.5787206
4	-1.265519	.0825742	-15.33	0.000	-1.427362 -1.103677

Indicateur de la qualité de l'ajustement

Nombre de fausses prédictions $\sum_{i=1}^N (y_i - \hat{y}_i)^2$

où $\hat{y}_i = 1$ si $\hat{F}(x_i\beta) \geq 1/2$ et $\hat{y}_i = 0$ si $\hat{F}(x_i\beta) < 1/2$.

Probabilité prédite

Illustration :

Currency crashes in emerging markets: an empirical treatment

Frankel et Rose (1996)

Journal of International Economics

Ils utilisent les données annuelles d'environ 100 pays en développement de 1971 à 1992 pour caractériser les crises de change. Ils se sont intéressés à 4 groupes de variables explicatives :

- variables étrangères
- indicateurs macroéconomiques du pays
- variables extérieures
- composition de la dette

Table 1
Probit estimates

© Théo Jalabert

	Default		Predictive	
	$\delta F(x)/\delta x$	z	$\delta F(x)/\delta x$	z
Commercial Bank/Debt	-0.07	0.57	0.03	0.21
Concessional	-0.10	1.74	-0.14	2.10
Variable Rate	0.03	0.21	-0.03	0.22
Short Term	0.04	0.34	0.23	1.97
FDI/Debt	-0.33	2.88	-0.31	2.47
Public Sector/Debt	0.11	1.32	0.19	2.18
Multilateral/Debt	-0.03	0.46	-0.06	0.81
Debt/GNP	0.03	1.33	-0.04	1.71
Reserves/Imports	-0.01	1.99	-0.01	3.39
Current Account	0.10	1.03	0.02	0.22
Over-Valuation	0.05	1.51	0.08	2.53
Government Budget	0.27	1.90	0.16	1.06
Domestic Credit	0.13	4.78	0.10	3.24
Growth Rate	-0.38	3.13	-0.16	1.29
Northern Growth	0.55	0.98	-0.85	1.50
Foreign Interest	1.27	4.50	0.80	2.60
Sample Size	803		780	
Pseudo-R2	0.20	P-Val	0.17	P-Val
Ho: Slopes = 0; $\chi^2(16)$	93.6	0.00	81.2	0.00
Ho: Debt Effects = 0; $\chi^2(7)$	14.2	0.05	25.5	0.00
Ho: External Effects = 0; $\chi^2(4)$	8.8	0.07	16.5	0.00
Ho: Macro Effects = 0; $\chi^2(3)$	32.9	0.00	12.3	0.01
Ho: Foreign Effects=0; $\chi^2(2)$	21.5	0.00	15.4	0.00

Default model: Goodness of fit

	Tranquility	Crash	Total
Predicted tranquility	727	65	792
Predicted crash	6	5	11
Total	733	70	803

Predictive model: Goodness of fit

	Tranquility	Crash	Total
Predicted tranquility	707	64	771
Predicted crash	4	5	9
Total	711	69	780
