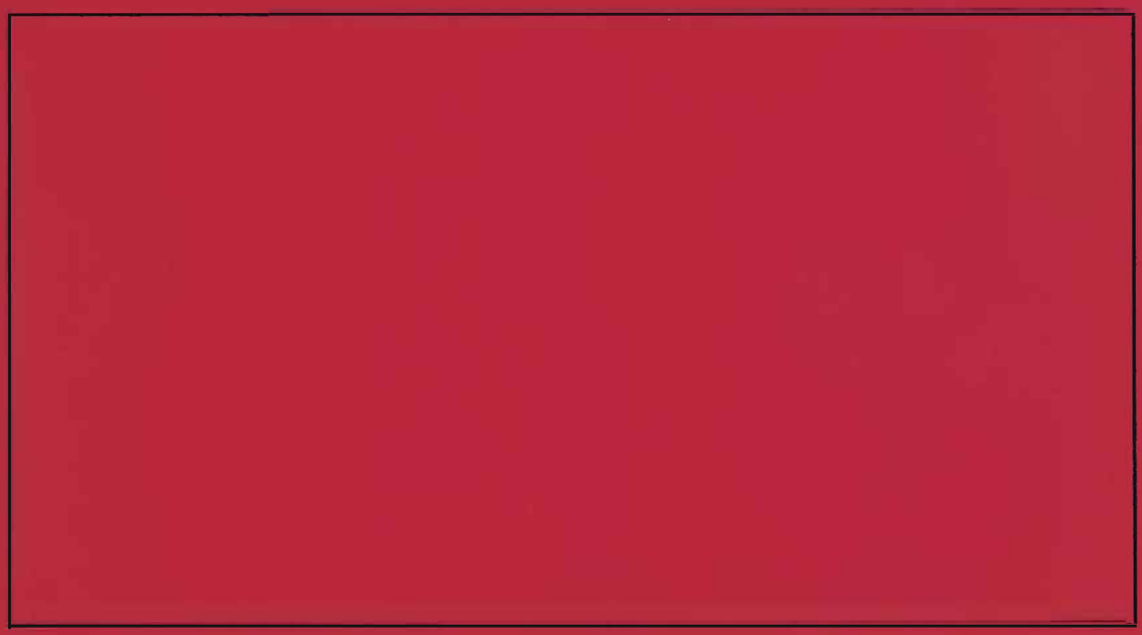
L'estimation de modèles à variable dépendante dichotomique - La sélection universitaire et la réussite en première année d'économie

Alain Mingat, Gérard Lassibille

I.M.E.

EQUIPE DE RECHERCHE ASSOCIEE AU C.N.R.S.

DOCUMENT DE TRAVAIL



INSTITUT DE MATHEMATIQUES ECONOMIQUES

UNIVERSITE DE DIJON

FACULTE DE SCIENCE ECONOMIQUE ET DE GESTION

4, BOULEVARD GABRIEL — 21000 DIJON

N° 20

L'ESTIMATION DE MODELES A VARIABLE DEPENDANTE DICHOTOMIQUE

Gérard LASSIBILLE

LA SELECTION UNIVERSITAIRE ET LA REUSSITE EN PREMIERE ANNEE D'ECONOMIE

Alain MINGAT

Avril 1977

Le but de cette Collection est de diffuser rapidement une première version de travaux afin de provoquer des discussions scientifiques. Les lecteurs désirant entrer en rapport avec un auteur sont priés d'écrire à l'adresse suivante :

INSTITUT DE MATHEMATIQUES ECONOMIQUES
4, Bd Gabriel - 21000 DIJON - France

Cette recherche a été réalisée par l'Institut de Recherche sur l'Economie de l'Education et financée par le Service d'Etudes et d'Informations Statistiques du Secrétariat aux Universités.

TRAVAUX déjà PUBLIES.

N°1 Michel PREVOT: Théorème du point fixe. Une étude topologique générale (juin 1974)

N°2 Daniel LEBLANC: L'introduction des consommations intermédiaires dans le modèle de LEFEBER (juin 1974)

N°3 Colette BOUNON: Spatial Equilibrium of the Sector in Quasi-Perfect Competition (september 1974)

N°4 Claude PONSARD: L'imprécision et son traitement en analyse économique (septembre 1974)

N°5 Claude PONSARD: Economie urbaine et espaces métriques (septembre 1974)

N°6 Michel PREVOT: Convexité (mars 1975)

N°7 Claude PONSARD: Contribution à une théorie des espaces économiques imprécis (avril 1975)

N°8 Aimé VOGT: Analyse factorielle en composantes principales d'un caractère de dimension-n (juin 1975)

N°9 Jacques THISSE et Jacky PERREUR: Relation between the Point of Maximum Profit and the Point of Minimum Total Transportation Cost: A Restatement (juillet 1975)

N°10 Bernard FUSTIER: L'attraction des points de vente dans des espaces précis et imprécis (juillet 1975)

N°11 Regis DELOCHE: Théorie des sous-ensembles flous et classification en analyse économique spatiale (juillet 1975)

N°12 G.LASSIBILLE et C.PARRON: Analyse multicritère dans un contexte imprécis (juillet 1975)

N°13 Claude PONSARD: On the Axiomatization of Fuzzy Subsets Theory (july 1975)

N°14 Michel PREVOT: Probability Calculation and Fuzzy Subsets Theory (August 1975)

N°15 Claude PONSARD: Hiérarchie des places centrales et Graphes -flous (avril 1976)

N°16 Jean-Pierre AURAY et Gérard DURU: Introduction à la théorie des espaces multiflous (avril 1976)

N°17 Roland LANTNER, Bernard PETITJEAN et Marie-Claude PICHERY: Jeu de simulation du circuit économique (Août 1976)

N°18 Claude PONSARD: Esquisse de simulation d'une économie régionale: l'apport de la théorie des systèmes flous (septembre 1976)

N°19 Marie-Claude PICHERY: Les systèmes complets de fonctions de demande (avril 1977)

Dans le cadre d'une recherche financée par le Secrétariat d'Etat aux Universités, l'Institut de Recherche sur l'Economie de l'Education a entrepris depuis la rentrée universitaire 1974-1975 une enquête longitudinale visant à donner une description la plus précise possible des processus de réussite, d'abandon et d'échec à l'Université. Cette enquête a été effectuée à l'Université de Dijon et concerne les disciplines suivantes : Médecine, Deug A de Sciences, Sciences économiques, Lettres classiques, Lettres modernes, Sciences sociales et Philosophie, ainsi que le Département de Gestion des entreprises de l'Institut Universitaire de Technologie de Dijon.

La population ayant servi de base à cette recherche est constituée de l'ensemble des étudiants s'inscrivant dans une des U.E.R. étudiées sans avoir été préalablement inscrit dans la même discipline. Parmi les étudiants de première année, on a donc éliminé les redoublants.

Les résultats concernent majoritairement la première année d'études, sachant que la recherche produira des résultats de façon régulière au cours des périodes à venir.

Globalement, 37,8 % des étudiants de l'échantillon, soit 374 sur les 1 254 inscrits, ont été autorisés à poursuivre en deuxième année. Ce chiffre moyen cache de grandes différences suivant les populations considérées, et il s'est avéré tout à fait nécessaire de rechercher les facteurs de différenciation interindividuelle quant à la sélection. La démarche suivie consiste à construire des modèles de réussite mettant en regard les caractéristiques des étudiants avec leurs "performances" universitaires. Ce point mérite une attention spécifique en raison du caractère particulier de la variable à expliquer la réussite. En effet, celle-ci est dichotomique 1-réussite contre 0-échec. Cette particularité nécessite des moyens d'estimation adaptés. La première partie de ce texte sera consacrée aux problèmes économétriques et aux solutions apportées, alors que la seconde partie donnera des résultats et plus spécialement ceux relatifs à la réussite en Sciences économiques.

L'ESTIMATION DE MODÈLES A VARIABLE DÉPENDANTE DICHOTOMIQUE

Gérard LASSIBILLE*
I.R.E.D.U. - Université de Dijon

RESUME

Ce texte a pour objet l'étude de l'estimation de la probabilité de réalisation d'un événement E, étant donné un certain nombre de caractéristiques associées à cette éventualité. Deux modèles sont envisagés, à savoir le modèle de régression linéaire et le modèle de régression logistique. Le premier, qui revient à estimer une fonction de probabilité linéaire ne vérifie plus les hypothèses classiques des moindres carrés ordinaires. Une première amélioration consiste alors à estimer le modèle par la méthode des moindres carrés généralisés. Cependant, outre le problème des tests de significativité des variables, une autre difficulté subsiste, à savoir que le modèle linéaire est inadéquate pour représenter une probabilité. Pour pallier ces inconvénients, il est nécessaire de recourir à un modèle non linéaire, tel que le modèle logistique, que l'on estimera par la méthode du maximum de vraisemblance.

* Nous exprimons notre reconnaissance envers Monsieur Pietro BALESTRA, Professeur aux Universités de Dijon et de Fribourg (Suisse) pour les discussions que nous avons eu à propos des modèles à variable dépendante qualitative. Nous portons évidemment seul la responsabilité de ce texte.

I - LA FONCTION DE PROBABILITE LINEAIRE

Considérons l'événement :

E = { réussite d'un individu à l'examen de fin d'année d'études }

Notons

y_i= 1, si cet événement se réalise pour l'individu i

et

y_i= 0, sinon

Supposons que la variable y_i, qui est une variable dichotomique, soit déterminée par k variables indépendantes, x (binaires ou non). L'hypothèse la plus simple que nous pouvons formuler lorsqu'une relation est supposée exister entre un certain nombre de variables est l'hypothèse de linéarité. C'est-à-dire que nous avons le modèle :

$$y_i = \beta_o + \beta_1 x_{1i} + \ldots + \beta k\, x_{ki} + \varepsilon_i \qquad i = 1,\ldots, n$$

où ε_i est un terme d'erreur aléatoire additif. Ce que nous pouvons encore écrire :

$$y_i = x^i\beta + \varepsilon_i$$

Où β est le vecteur d'ordre (k+1,1) de paramètres inconnus et x^i est le vecteur (1,k+1) des variables explicatives associées à l'individu i.

La variable dépendante y_i, étant une forme linéaire de l'erreur aléatoire ε_i, est une variable aléatoire. De plus, il s'agit d'une variable indicatrice puisqu'elle ne peut prendre que deux valeurs 0 ou 1. En raison du caractère dichotomique de cette variable, son espérance mathématique n'est rien d'autre que la probabilité conditionnelle de réalisation de l'événement E, étant donné le vecteur des variables exogènes x^1 (d'où le nom de fonction de probabilité linéaire).

---

[1] Voir à ce propos Mac Gillivray, R : "Estimating the Linear Probability Function", Econometrica - Volume 38 - N°5 - 1970 - pp.775-776

Nous allons, à partir d'un problème particulier, montrer d'une part que les hypothèses relatives à la méthode des moindres carrés ordinaires ne sont plus vérifiées dans le cas d'un modèle à variable dépendante dichotomique et d'autre part, qu'une première amélioration consiste à estimer ce modèle par la méthode des moindres carrés généralisés.

Pour ce faire, nous disposons d'un échantillon de 214 étudiants inscrits pour la première fois à l'U.E.R. de Médecine de l'Université de Dijon au début de l'année scolaire 1974-75[2], grâce auquel nous allons estimer la probabilité de réussite. Les variables exogènes, supposées déterminer la réussite scolaire sont les suivantes[3]:

	Moyenne de l'échantillon	Ecart-type de l'échantillon
1 - Taille de la commune de résidence: variable polytomique codée 1. si la commune est < 3000 hab. 2. si la commune est comprise entre 3000 et 20000 hab. 3. si la commune est comprise entre 20000 et 100000 hab. 4. si la commune est > 100000 hab	2,387	1,063
2 - Revenus mensuels des parents	5 000	2 819
3 - Age de l'étudiant	18,791	1,113
4 - Résultat à un test d'aptitudes logiques	29,570	4,178
5 - Résultat à un test de personnalité	35,484	24,474
6 - Moyenne à l'écrit du baccalauréat	10,784	1,810
7 - Etudes précédentes : variable dichotomique codée 1. si l'étudiant était déjà dans le supérieur avant 1974-75 0. sinon	0,056	0,230

---

[2] Cet échantillon fait partie d'une étude réalisée par l'Institut de Recherche sur l'Economie de l'Education et financée par le Service d'Etudes et d'Informations Statistiques du Secrétariat aux Universités.

[3] Voir à ce propos, A. MINGAT : "La première année d'études, la réussite, l'abandon, l'échec" - Cahier de l'I.R.E.D.U. - N°23.

8 - Origine du secondaire : variable dichotomique codée 1. si l'étudiant a effectué ses études secondaires dans un établissement public 0. sinon	0,854	0,361
9 - Baccalauréat scientifique : variable dichotomique codée 1. si l'étudiant posséde un baccalauréat série C 0. sinon	0,350	0,478
10 - Baccalauréat non scientifique : variable dichotomique codée 1. si l'étudiant posséde un baccalauréat série A, ou B, ou F, ou G 0. sinon	0,051	0,221

I-1. Estimation par la méthode des moindres carrés ordinaires

La méthode des moindres carrés ordinaires fournis des estimateurs BLUE, si les erreurs ε_i du modèle satisfont les hypothèses suivantes

$$E(\varepsilon_i) = 0 \quad \forall_i \text{ (hypothèse non restrictive)}$$

$$E(\varepsilon_i\varepsilon_j) = \sigma^2 \text{ Pour } i = j$$

$$= 0 \text{ Pour } i \neq j$$

La dernière hypothèse est celle d'indépendance et d'homoscédasticité des erreurs. ELle indique que le comportement de l'individu i est indépendant du comportement de l'individu j et que la variance de l'erreur est identique quelque soit l'individu.

Si vous estimons le modèle de réussite linéaire par cette méthode nous obtenons les résultats suivants :

Variable	Coefficient	t
Taille de la commune	- 0,043	1,87*
Revenus des parents/1000	0,013	1,42
Age/10	- 0,456	1,60
Test logique	0,038	0,58
Test de personnalité/10	- 0,030	2,71***

Moyenne à l'écrit du bac.	0,078	5,05***
Etudes précédentes	0,232	1,84*
Origine du secondaire	0,050	0,62
Baccalauréat C/D	0,233	3,90***
Baccalauréat A,B,F,G/D	- 0,099	0,79
Constante	0,207	
	R^2 = 0,30	

Tableau 1 - Estimation de la fonction de probabilité linéaire par les moindres carrés ordinaires.

L'interprétation à donner aux coefficients de régression est la suivante : l'augmentation d'une variable dotée d'un coefficient de régression positif entraîne une augmentation de la probabilité de réussite, au contraire, l'augmentation d'une variable dotée d'un coefficient de régression négatif entraîne une augmentation de la probabilité d'échec.

Les tests de significativité des variables sont construits sous l'hypothèse de normalité des erreurs. Les seuils retenus sont les suivants :

* = 10 % ** = 5 % *** = 1 %

Ainsi le modèle linéaire estimé par les moindres carrés ordinaires explique 30 % de la variance de la réussite des étudiants en première année de médecine. Les tests de Student indiquent que seulement 5 variables exogènes sur les 10 initialement retenues sont significatives à un seuil au moins égal à 10 %.

Posons-nous la question de savoir s'il est licite d'estimer ce modèle par les moindres carrés ordinaires et par conséquent si nous pouvons accepter l'hypothèse d'homoscédasticité des erreurs. Pour ce faire, il suffit de construire un test de non-homoscédasticité[4] basé sur l'estimation du modèle pour deux sous-population. Celles-ci ont été tirées de façon aléatoire dans léchantillon initial des 214 individus. La comparaison du rapport entre la somme des carrés des résidus du modèle estimé par les moindres carrés ordinaires sur la première sous-population et la somme des carrés des résidus du modèle estimé par la même méthode sur la seconde sous-population, avec un F de Fisher théorique, indique que nous ne pouvons pas écarter l'hypothèse que les erreurs soient en fait hétéroscédastiques.

---

[4] Pour une description théorique de ce test, voir Theil, H. = Principles of Econometrics , Wiley, New York - 1971 - pp. 196-197

En effet, nous obtenons :

$$\frac{\text{SS de la 1ère sous-population}}{\text{SS de la 2ème sous-population}} = 2,43$$

alors que $F(n_1-11, n_2-11) = 1,39$ (avec n_1 et n_2 les effectifs de chacune des deux sous-populations).

Le test confirme en fait la démonstration théorique de A. Goldberger[5]. En effet, dans un modèle à variable dépendante dichotomique, l'erreur ε_i, ne peut prendre que deux valeurs, à savoir

$$\varepsilon_i = 1-x^i\beta \quad , \quad \text{si } y_i = 1$$
$$\varepsilon_i = -x^i\beta \quad , \quad \text{si } y_i = 0$$

En admettant que $E(\varepsilon_i) = 0$, nous avons

$$E(\varepsilon_i) = \text{Prob}\{\varepsilon_i = -x^i\beta\}(-x^i\beta) + \text{Prob}\{\varepsilon_i=1-x^i\beta\}(1-x^i\beta) = 0$$

et comme

$$\text{Prob}\{\varepsilon_i=-x^i\beta\} + \text{Prob}\{\varepsilon_i=-1-x^i\beta\} = 1$$

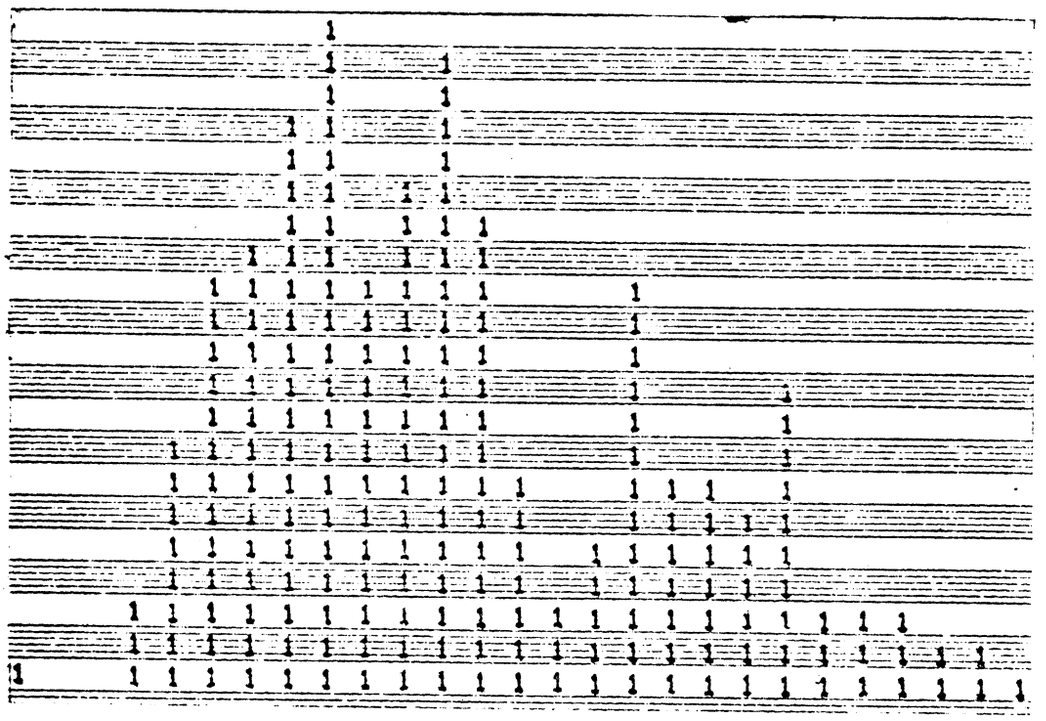
nous en déduisons que

$$\text{Var}(\varepsilon_i) = x^i\beta(1-x^i\beta) \neq \text{Var}(\varepsilon_j)$$

Les erreurs étant hétéroscédastiques, les estimateurs obtenus précédemment s'ils sont centrés et convergents ne sont toutefois pas efficients, c'est-à-dire qu'il est possible de trouver d'autres estimateurs dont la variance est plus petite. La conséquence en est que les tests de significativité des variables sont biaisés. Cependant d'hétéroscédasticité des erreurs n'est pas seule en cause. En effet, nous avons implicitement formulé l'hypothèse de normalité des erreurs pour juger de la significativité des variables. Or, il est bien évident que dans un modèle à variable dépendante dichotomique cette hypothèse n'est plus acceptable. Les tests que nous serions amenés à construire pour juger de la non-normalité (test du X^2, droite de Henry) étant inadaptés puisque les erreurs sont *à la fois* hétéroscédastiques et non normales, seule l'observation de l'histogramme des résidus permet d'indiquer qu'effectivement les erreurs ne suivent pas une loi de Laplace Gauss.

---

[5] Goldberger A. = Econometric Theory, Wiley, New York, 1964, p.249



Histogramme des résidus

Ignorer l'hétéroscédasticité et se fier à des tests inexacts nous conduirait ainsi au modèle simplifié suivant, dans lequel seulement quatre variables sont supposées déterminer la probabilité de réussite :

Variable	Coefficient	t
Taille de la commune	- 0,047	1,88*
Test de personnalité /10	- 0,029	2,65***
Moyenne à l'écrit du bac	0,087	5,82***
Baccalauréat C/D	0,264	4,64***
Constante	- 0,524	
	R^2 = 0,29	

Tableau 2 - Estimation de la fonction de probabilité linéaire "simplifiée" par les moindres carrés ordinaires.

Naturellement, le pouvoir explicatif de ce modèle est sensiblement équivalent au modèle estimé précédemment. Toutefois, il faut remarquer que la variable "Etudes Précédentes" significative dans le premier modèle, ne l'est plus dans le second. Ceci est sans doute dû à un phénomène de multicollinéarité entre les variables "Etudes Précé-

dentes" et "Age de l'Etudiant". En effet, de toutes les variables exogènes, ce sont elles qui sont le plus fortement corrélés (r=0,365).

I-2. Estimation par la méthode des moindres carrés généralisés

Pour pallier le problème de l'hétéroscédasticité, il est nécessaire d'estimer les paramètres inconnus du modèle par la méthode des moindres carrés généralisés. Toutefois, la variance des erreurs étant inconnue, la méthode de Aitken pure n'est pas réalisable, il est donc indispensable de donner auparavant une estimation convergente des variances de chacune des erreurs. Mac Gillivray[6] suggère de prendre comme estimateur de la variance, l'expression suivante :

$$\hat{\text{Var}}\ (\varepsilon_i) = \hat{y}_i\ (1-\hat{y}_i)$$

où $\hat{y}_i$ est la valeur calculée du modèle

$$y_i = x^i\beta+\varepsilon_i$$

estimé par les moindres carrés ordinaires.
Cependant, il n'est pas exclu que certaines variances soient négatives. On peut alors tourner la difficulté, en choisissant de prendre

$$\hat{\text{Var}}\ (\varepsilon_i) = \left|\hat{y}_i\ (1-\hat{y}_i)\right|$$

L'estimation du modèle linéaire de réussite par la méthode de Aitken réalisable est la suivante :

---

[6] Mac Gillivray, R : Estimating the Linear Probability Function, Econometrica, Vol.38, n°5,1970, pp.775-776

Variable	Coefficient	t
Taille de la commune	- 0,028	1,48
Revenus des parents/ 1000	0,013	2,02*
Age/10	- 0,430	2,30**
Test logique	0,001	0,50
Test de personnalité /10	- 0,030	3,66***
Moyenne à l'écrit du bac	0,078	7,86***
Etudes précédentes	0,209	1,81*
Origine du secondaire	0,074	1,37
Baccalauréat C/D	0,210	4,60***
Baccalauréat A,B,F,G/D	- 0,129	2,08***
Constante	0,174	
.	R^2 = 0,29	

Tableau 3 - Estimation de la fonction de probabilité linéaire par les moindres carrés généralisés.

Ce modèle explique 29 % de la variance de la réussite des étudiants. Alors qu'à l'issue de l'estimation du même modèle par les moindres carrés ordinaires, seulement cinq variables sur les dix intialement retenues étaient significatives à un seuil au moins égal à 10 %, avec la méthode des moindres carrés généralisés trois variables supplémentaires sont significatives. Il s'agit des variables "Age de l'Etudiant", "Revenu des parents" et "Baccalauréat non scientifique". Etant donné que les effets marginaux des variables, obtenus par les moindres carrés ordinaires et par les moindres carrés généralisés sont sensiblement équivalent, nous pouvons attribuer cette supériorité de significativité (étant entendu que l'hypothèse de non normalité défavorise aussi bien les tests dans l'une ou l'autre méthode)à une plus grande efficience des estimateurs des moindres carrés généralisés. Il suffit pour s'en convaincre, de comparer les variances des estimateurs obtenus par l'une ou l'autre méthode.

	Variance des estimateurs des MCO	Variance des estimateurs des MCG
Taille de la commune	$0,669.10^{-3}$	$0,371.10^{-3}$
Revenu des parents	$0,936.10^{-10}$	$0,439.10^{-10}$
Age	$0,806.10^{-5}$	$0,352.10^{-5}$
Test logique	$0,442.10^{-4}$	$0,144.10^{-4}$
Test de personnalité	$1,231.10^{-8}$	$0,709.10^{-8}$
Moyenne à l'écrit du bac	$0,241.10^{-5}$	$0,132.10^{-5}$
Etudes précédentes	$0,158.10^{-1}$	$0,133.10^{-1}$
Origine du secondaire	$0,568.10^{-2}$	$0,275.10^{-2}$
Baccalauréat C/D	$0,358.10^{-2}$	$0,209.10^{-2}$
Baccalauréat A,B,F,G/D	$1,59.10^{-2}$	$0,387.10^{-2}$

Si nous estimons à nouveau la probabilité de réussite en ignorant les variables non significatives, nous obtenons le modèle "simplifié" suivant :

Variable	Coefficient	t
Revenu des parents/1000	0,037	1,86*
Age/10	- 0,616	3,08***
Test de personnalité/10	- 0,023	2,93***
Moyenne à l'écrit du bac	0,090	6,93***
Etudes précédentes	0,221	1,89*
Baccalauréat C/D	0,164	3,34***
Baccalauréat A,B,F,G/D	- 0,183	2,71***
Constante	0,487	
	$R^2 = 0,28$	

Tableau 4 - Estimation de la fonction de probabilité linéaire "simplifié" par les moindres carrés généralisés.

Nous donnons ci-après, les prédictions de la probabilité de réussite pour un sous échantillon aléatoire de 40 étudiants. Ces prévisions sont calculées à partir des modèles estimés par les moindres carrés ordinaires (Tableau 1 et 2) et par les moindres carrés généralisés (Tableaux 3 et 4).

(Les numéros en tête de colonnes renvoient aux tableaux dans lesquels figurent les modèles correspondant).

N° d'observation	Valeur observée	1	3	2	4
159	1	0,680	0,665	0,380	0,587
189	0	-0,103	-0,031	-0,057	-0,054
76	0	-0,016	-0,014	0,044	-0,056
167	1	0,475	0,449	0,550	0,566
7	0	-0,033	-0,036	0,064	0,031
34	1	0,958	0,941	0,785	0,934
81	0	0,502	0,461	0,501	0,392
20	1	0,668	0,647	0,632	0,661
66	1	0,657	0,646	0,593	0,634
127	0	0,265	0,278	0,073	0,361
5	0	-0,043	-0,025	0,107	0,075
212	0	0,190	0,200	0,107	0,161
44	0	0,318	0,359	0,361	0,396
184	1	0,458	0,449	0,405	0,312
121	0	0,021	0,094	0,040	0,134
97	1	0,185	0,198	0,238	0,175
118	0	0	-0,002	0,076	0,031
111	1	1,055	1,017	1,038	0,984
21	1	0,046	0,095	0,033	0,048
162	0	-0,318	-0,312	-0,249	-0,224
67	0	0,229	0,242	0,201	0,272
149	0	-0,107	-0,099	0,049	-0,065
163	1	0,592	0,583	0,639	0,559
47	1	0,291	0,300	0,096	0,250
106	0	0,313	0,289	0,336	0,247
102	0	0,448	0,447	0,532	0,388
197	1	0,611	0,614	0,653	0,637
33	0	0,184	0,198	0,166	0,203
29	1	0,443	0,452	0,532	0,465
178	0	0,612	0,608	0,625	0,593
2	0	0,076	0,082	0,094	0,053
158	0	0,147	0,189	0,224	0,332
103	0	0,594	0,573	0,528	0,492
91	1	0,626	0,592	0,582	0,571
179	0	0,028	0,035	0,074	0,091
93	0	0,528	0,585	0,567	0,426
75	1	0,432	0,414	0,390	0,362
101	0	-0,011	-0,023	-0,018	-0,066
169	0	-0,047	-0,023	0,035	0,002
12	0	0,182	0,168	0,131	0,220

Si nous calculons pour chaque modèle, la distance entre les valeurs observées et les valeurs prédites ci-dessus, nous obtenons :

	Modèle complet (1 ou 3)	Modèle simplifié (2 ou 4)
Prédictions des MCO	6,11	6,66
Prédictions des MCG	6,06	6,27

Ainsi pour l'échantillon considéré, et toutes choses égales par ailleurs, les prédictions obtenues par les moindres carrés généralisés sont meilleurs que celles obtenues par les moindres carrés ordinaires. L'écart entre les prévisions du modèle complet et du modèle simplifié (qui diffère selon la méthode d'estimation) est plus faible par les moindres carrés généralisés car ceux-ci permettent de retenir plus de variables significatives parmi un ensemble initial de variables qui le sont en réalité toutes.

Bien que l'estimation de la fonction de probabilité linéaire par la méthode des moindres carrés généralisés constituent une amélioration par rapport à son estimation par les moindres carrés ordinaires, il n'en reste pas moins qu'un problème important subsiste, à savoir que la forme linéaire est inadéquate pour représenter une probabilité qui varie par définition dans l'intervalle $\left[0,1\right]$. Il en est pour preuve les prédictions figurant au tableau précédent et dans lequel certains individus ont une probabilité de réussite négative, voire même supérieure à un. Ceci est un handicap sérieux et plutôt que d'estimer une fonction de probabilité linéaire il est préférable d'estimer une fonction non linéaire prenant ses valeurs dans l'intervalle $\left[0,1\right]$.

II - L'ANALYSE LOGISTIQUE

Cette analyse dénommée par J. Berkson, "Logit Analysis" a été utilisée en premier lieu par les biologistes. Les réponses d'individus à un quelconque stimulus, qu'examinent ceux-ci, sont comparables à certaines réactions d'agents économiques. Ainsi dans le cas qui nous préoccupe, pour chaque étudiant il existe par exemple un certain niveau de revenu en deça duquel il échoue et au delà duquel il réussit. Ce niveau, qu'un biologiste appelerait tolérance est une variable aléatoire et peut donc être caractérisée par une fonction de densité de probabilité. Si un revenu x_o est attribué à un individu et si $f(x)$ représente la densité de la tolérance, alors la probabilité de réalisation de l'événement pour cet individu est donné par[7]:

$$P = \int_o^{x_o} f(x)\ dx$$

---

[7] Voir ASHTON, W.D. : The Logit Transformation, Hafner, New York, 1972

Plusieurs solutions peuvent alors être envisagées pour représenter P par une fonction de répartition logistique du type :

$$P = \frac{1}{1+e^{-(a+bx)}}$$

II-1. La transformation logistique de la fonction de probabilité linéaire

Cette méthode consiste, à partir d'une estimation de la fonction de probabilité linéaire par la méthode des moindres carrés généralisés, à introduire ex-post la fonction logistique. Cette fonction variant dans l'intervalle $[0,1]$ nous n'obtiendrons plus de valeurs aberrantes pour les prédictions des probabilités.

Le modèle linéaire que nous avons estimé par les moindres carrés généralisés (Tableau 3) est de la forme :

$$y_i = x^i\beta+\varepsilon_i$$

La méthode dite de transformation logistique de la fonction de probabilité linéaire consiste à ranger en classes les valeurs prédites obtenues lors de l'estimation de la forme linéaire, et à calculer pour chacune d'elles l'expression :

$$\hat{L} = \ln\left(\frac{\bar{\hat{y}}}{1-\bar{\hat{y}}}\right)$$

où $\bar{\hat{y}}$ représente la moyenne des prédictions de chacun des intervalles, obtenues par la méthode des moindres carrés généralisés.

La régression linéaire de cette expression sur les centres de classes permet de mettre en relation $\hat{L}$ avec $\hat{y}$ et donc indirectement avec le vecteur des variables exogènes.

Les résultats obtenus sur le modèle complet estimé par les moindres carrés généralisés (Tableau 3) sont les suivants :

$$\hat{L} = -2,798 + 5,557\ \hat{y}_i$$

On exprime alors la probabilité de réussite d'un individu i, de la façon suivante :

$$\hat{P}_i = \frac{1}{1+e^{-\hat{L}_i}}$$ [8]

---

[8] Il est à remarquer que si Q_i représente la probabilité d'échec nous avons

$$\ln\frac{\hat{P}_i}{Q_i} = \hat{L}_i = \hat{a} + \hat{b}y_i = \hat{a} + \hat{b}\ (x^i\hat{\beta})$$

Par exemple, si nous considérons le premier individu de notre sous-échantillon aléatoire, qui a une valeur prédite égale à 0,665, il faut pour donner une estimation de sa probabilité de réussite "révisée" par cette méthode, calculer :

$$\hat{L}_i = -2,798 + (5,557 \times 0,665) = 0,897$$

Puis

$$\hat{P}_i = \frac{1}{1+e^{-0,897}} = 0,710$$

Le tableau ci-dessous donne l'estimation des probabilités de réussite révisée pour l'ensemble du sous-échantillon :

N°	Valeur observée	Prédiction	N°	Valeur observée	Prédiction
159	1	0,710	67	0	0,189
189	0	0,048	149	0	0,033
76	0	0,053	163	1	0,608
167	1	0,424	47	1	0,243
7	0	0,047	106	0	0,232
34	1	0,919	102	0	0,422
81	0	0,441	197	1	0,648
20	1	0,689	33	0	0,154
66	1	0,688	29	1	0,428
127	0	0,222	178	0	0,641
5	0	0,050	2	0	0,087
212	0	0,156	158	0	0,148
44	0	0,309	103	0	0,595
184	1	0,424	91	1	0,620
121	0	0,093	179	0	0,068
97	1	0,154	93	0	0,529
118	0	0,056	75	1	0,378
111	1	0,945	101	0	0,050
21	1	0,093	169	0	0,050
162	0	0,010	12	0	0,134

Si cette méthode permet de ne plus obtenir des valeurs prédites négatives ou supérieures à un, elle n'est cependant pas totalement satisfaisante car d'une part elle est basée sur l'estimation de la fonction de probabilité linéaire par les moindres carrés généralisés, or les tests de significativité des variables sont nécessairement biaisés et d'autre part si nous calculons la distance entre les valeurs observées et les valeurs prédites nous obtenons une distance égale à 6,15, donc supérieure à celle obtenue lors de l'estimation du modèle linéaire par les moindres carrés généralisés ou par les moindres carrés ordinaires.

II-2. La Logit Analysis

Nous estimons ici, un modèle du type

$$y_i = \frac{1}{1+e^{-xi\beta}} + \varepsilon i$$

Les hypothèses relatives aux erreurs aléatoires sont les mêmes que précédemment, à savoir :

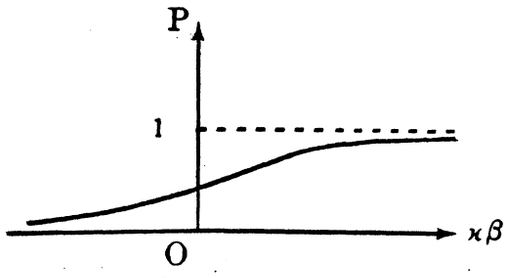
$$E(\varepsilon_i) = 0 \quad \forall_i$$

$$E(\varepsilon_i\varepsilon_j) = y_i(1-y_i) \quad \text{pour} \quad i = j$$

$$= 0 \qquad \text{sinon}$$

Sous l'hypothèse de nullité de l'espérance mathématique de l'erreur aléatoire, la probabilité de réalisation P_i, de l'événement "réussite d'un étudiant médecin à l'examen de fin d'année d'études" est égale à :

$$P_i = E(y_i|x^i\beta) = \frac{1}{1+e^{-x^i\beta}}$$



Comment obtenir les estimateurs $\hat{\beta}$ des paramètres inconnus ?
Une première méthode[9] consisterait à estimer la fonction inverse, c'est-à-dire

$$\ln\frac{Pi}{1-P_i} = x^i\beta$$

par les moindres carrés. A notre avis cette méthode présente plusieurs inconvénients. D'une part, transformer la variable endogène implique que nous transformons également l'erreur aléatoire si bien que sa distribution n'est plus la même que dans le modèle initial, d'autre part l'estimation du modèle transformé, ci-dessus, nécessiterait l'emploi de données groupées afin de définir la nouvelle variable dépendante. Or il est extrêmement difficile de grouper les individus selon les valeurs des variables exogènes et ceci d'autant plus que leur nombre est élevé.

---

[9] Voir Theil H. : Principles of Econometrics, Wiley, New York, 1971, pp.628-63

Une seconde méthode, applicable à des données individuelles consiste à estimer le modèle défini précédemment par la méthode du maximum de vraisemblance. Dans le cas d'un modèle à variable dépendante dichotomique, la fonction de vraisemblance de l'échantillon s'écrit de la manière suivante[10] :

$$L\ (\beta_o,\ldots,\beta_k \mid y_1,\ldots,y_i,\ldots,y_n) = \prod_{i=1}^{n} \left(\frac{1}{1+e^{-x^i\beta}}\right)^{y_i} \left(1-\frac{1}{1+e^{-x^i\beta}}\right)^{1-y_i}$$

Cette fonction n'est rien d'autre que le produit des fonctions de densité de probabilité individuelles, celles-ci s'exprimant par l'une ou l'autre expression située à droite du signe égal selon que la valeur de la variable dépendante de l'individu est 1 ou 0.

L'estimation du modèle par la méthode du maximum de vraisemblance revient à maximiser la fonction L par rapport à tous les paramètres inconnus β_k. La condition pour avoir un maximum est que les dérivés premières de la vraisemblance par rapport aux paramètres inconnus soient nulles. Habituellement, dans le cas linéaire, la résolution du système linéaire d'équations normales permet de déduire ces estimateurs. Il est bien évident que dans le cas qui nous préoccupe, ces équations ne sont pas linéaires dans les paramètres, de ce fait la résolution du système d'équations normales n'est pas simple. Seule une méthode d'optimisation numérique permet alors de découvrir les estimateurs des paramètres inconnus.

Plusieurs techniques peuvent être utilisées, les plus courantes étant la méthode de Newton et celle de la plus grande pente. A ces deux méthodes itératives, nous avons préféré la méthode des variations locales qui présente l'avantage non négligeable, vue la complexité de la fonction, de ne pas recourir au calcul des dérivées. Le principe de cette méthode est le suivant. Supposons que nous cherchions les valeurs x_1^* et x_2^* qui maximisent la fonction $f(x_1, x_2)$. Pour ce faire, nous nous donnons un point de départ (x_1°, x_2°)[11] auquel est associé la valeur $f_1 = f(x_1^\circ, x_2^\circ)$ de la fonction. Soit α la perturbation ou le pas initial que l'on accepte sur l'une

---

[10] Voir Nerlove M-Press, SJ : Univariate and Multivariate Log Linear and Logistic Models. Rand Corporation, Santa Monica, 1973, p.16

[11] Nous avons choisi comme vecteur de départ β_o, le vecteur des paramètres estimés par la méthode des moindres carrés ordinaires.

quelconque des deux variables. Imaginons que nous fassions tout d'abord varier x_1 de $\pm$ α. Il est possible de calculer

$$f_1^+ = f(x_1^o + \alpha, x_2^o) \text{ et } f_1^- = f(x_1^o - \alpha, x_2^o)$$

La méthode des variations locales consiste alors à retenir pour nouvelle valeur de la variable x_1, celle qui réalise le maximum de $\{f_1^-, f, f_1^+\}$. Soit x_1^α cette valeur. Il suffit ensuite de remplacer x_1^o par x_1^α et d'itérer en acceptant cette fois-ci une perturbation sur la variable x_2. Dès que nous trouvons un point stationnaire, c'est-à-dire un point tel qu'il n'est plus possible d'augmenter la valeur de la fonction dans une quelconque direction grâce au pas initial α, nous recommençons le processus en divisant la perturbation par deux. L'optimum est atteint lorsque la différence entre les valeurs de la fonction pour deux points stationnaires consécutifs est inférieur à un seuil donné.

Ayant découvert grâce à cette méthode, l'estimation des paramètres inconnus, $\hat{\beta}$[12], nous ne pouvions pour juger de la significativité d'une variable x_k qu'utiliser le test du rapport de vraisemblance défini par :

$$\lambda = \frac{Lc}{L}$$

où L représente la valeur de la fonction de vraisemblance au point $\hat{\beta}$, et Lc représente la valeur de la fonction de vraisemblance au point

$$\hat{\beta}_c' = (\hat{\beta}_o, \hat{\beta}_1, \ldots, \hat{\beta}_{k-1}, 0).$$

La comparaison de $-2\ln\lambda$ avec un χ^2 théorique permet alors de déterminer la significativité de la variable x_k.

Les résultats obtenus sont les suivants :

---

[12] L'optimisation de la fonction a nécessité 922 itérations, soit 1 H 15 mn d'utilisation de l'ordinatuer PDP 15. A ce propos, nous tenons à remercier Monsieur C. Michelot du Laboratoire d'Analyse Numérique de l'U.E.R. M.I.P.C. de l'Université de Dijon, qui a bien voulu se charger du programme d'optimisation.

Variable	Coefficient	X^2
Taille de la commune	- 0,317	19,78***
Revenu des parents/1000	0,117	13,56***
Age/10	- 3,609	1223,98***
Test logique	0,011	3,58*
Test de personnalité/10	- 0,251	28,34**
Moyenne à l'écrit du bac	0,535	591,55***
Etudes précédentes	1,851	5,36***
Origine du secondaire	0,368	3,27*
Baccalauréat C/D	1,412	25,25***
Baccalauréat A, B, F, G/D	- 15,100	3,29*
Constante	- 0,496	
	% de variance expliquée 0,35	

Tableau 5 - Estimation de la fonction de probabilité logistique par la méthode du maximum de vraisemblance.

L'explication de la réussite des étudiants par le modèle logistique est supérieur de 5 ou de 7 % à celle obtenue par le modèle linéaire estimé par les moindres carrés ordinaires ou par les moindres carrés généralisés. Si ce point est important, un autre l'est encore plus pour le chercheur empiriste, il s'agit du problème de la significativité des variables. Alors qu'au vue de l'estimation de la fonction de probabilité linéaire nous sommes amenés à rejeter l'influence de certaines variables sur la réussite, il n'en est plus de même dans le cas du modèle logistique. La raison en est qu'il était abusif d'admettre que les estimateurs des paramètres inconnus suivent une loi de Student. L'avantage de l'estimation du modèle logistique réside dans la qualité des estimateurs du maximum de vraisemblance. En effet, la fonction de vraisemblance étant convexe, ceux-ci sont asymptotiquement efficients et asymptotiquement normaux.

Non seulement certaines variables ne sont pas significatives à l'issue de l'estimation de la fonction de probabilité linéaire mais de plus la structure ordinale des variables significatives influant sur la probabilité de réussite diffère considérablement selon que l'on adopte le modèle linéaire ou le modèle logistique.

Pour nous permettre d'évaluer les différences entre les prédictions obtenues par le modèle linéaire et par le modèle logistique nous donnnons ci-dessous, les valeurs calculées de la probabilité de réussite pour chacun des individus constituant le sous-échantillon aléatoire défini précédemment. Ces valeurs prédites sont obtenues à partir du Tableau 3 pour la fonction de probabilité linéaire et du Tableau 5 pour la fonction de probabilité logistique. (Les numéros en tête de colonnes renvoient aux tableaux dans lesquels figurent les modèles correspondants).

N° d'observation	Valeur observée	3	5
159	1	0,665	0,849
189	0	- 0,031	0,021
76	0	- 0,014	0,030
167	1	0,449	0,474
7	0	- 0,036	0
34	1	0,941	0,965
81	0	0,461	0,439
20	1	0,647	0,782
66	1	0,646	0,728
187	0	0,278	0,192
5	0	- 0,025	0,021
212	0	0,200	0,133
44	0	0,359	0,222
184	1	0,449	0,443
121	0	0,094	0,048
97	1	0,198	0,109
118	0	- 0,002	0,026
111	1	1,017	0,978
21	1	0,095	0,060
162	0	- 0,312	0,002
67	0	0,242	0,146
149	0	- 0,099	0
163	1	0,583	0,639
47	1	0,300	0,210
106	0	0,289	0,196
102	0	0,447	0,414
197	1	0,614	0,728
33	0	0,198	0,111
29	1	0,452	0,352
178	0	0,608	0,699
2	0	0,082	0,055
158	0	0,189	0,105
103	0	0,573	0,661
91	1	0,592	0,670
179	0	0,035	0,035
93	0	0,525	0,567
75	1	0,414	0,368
101	0	- 0,023	0,026
169	0	- 0,023	0,021
12	0	0,168	0,142

La distance entre les valeurs observées et les valeurs prédites pour le modèle logistique est égale à 5,99 alors qu'elle est de 6,06 pour le modèle linéaire estimé par les moindres carrés généralisés. La comparaison des prédictions indique que par rapport au modèle logistique, le modèle linéaire surestime la probabilité de réussite des individus dont le variable endogène est égale à un, dans 50 % des cas, alors qu'il surestime la probabilité d'échec des individus dont la variable endogène est égale à zéro, dans 48 % des cas.

La comparaison des prédictions obtenues par la méthode de transformation logistique de la fonction de probabilité linéaire (page 14) et par le modèle logistique indique quant à elle que par rapport à ce dernier la transformation logistique de la fonction de probabilité linéaire surestime la probabilité d'échec et sous-estime la probabilité de réussite.

Le tableau ci-dessous donne l'élasticité de la probabilité de réussite (calculée au point moyen) par rapport à chacune des variables dans le but de faciliter la comparaison des résultats fournis par le modèle linéaire estimé par les moindres carrés généralisés (Tableau 3) et par le modèle logistique estimé par la méthode du maximum de vraisemblance (Tableau 5). L'avantage qu'il y a à comparer les élasticités plutôt que les effets marginaux tient au fait que dans le modèle logistique ceux-ci ne sont pas constants comme dans le modèle linéaire, mais varie en fonction du niveau de probabilité auquel on se situe.

Variables	Modèle logistique	Modèle linéaire
Taille de la commune	- 0,54	- 0,23
Revenus des parents/1000	0,42	0,23
Age/10	- 4,88	- 2,88
Test logique	0,23	0,10
Test de personnalité/10	- 0,64	- 0,38
Moyenne à l'écrit du bac	4,15	3,00
Etudes précédentes	0,07	0,04
Origine du secondaire	0,22	0,22
Baccalauréat C/D	0,35	0,26
Baccalauréat A,B,F,G/D	- 0,55	- 0,02

Tableau 6 - Elasticités de la probabilité de réussite par rapport à chacune des variables.

Les variables influant le plus sur la probabilité de réussite (du point de vue des élasticités) sont dans l'un et l'autre modèle les variables "Age" et "Moyenne à l'écrit du baccalauréat". Toutefois les élasticités de la probabilité de réussite par rapport à ces variables sont beaucoup plus faibles dans le modèle linéaire comme le sont d'ailleurs toutes les autres élasticités. Alors qu'une augmentation identique de chacune des variables exogènes aurait pour effet de laisser pratiquement inchangée la probabilité de réussite du modèle linéaire, elle diminuerait de plus de 1 % la probabilité du modèle logistique.

LA SÉLECTION UNIVERSITAIRE ET LA RÉUSSITE EN PREMIÈRE ANNÉE D'ÉCONOMIE

Alain MINGAT.

La seconde partie de ce texte sera consacrée à l'utilisation des outils économétriques exposés dans la première partie et d'une façon plus générale à donner des résultats empiriques sur la réussite universitaire. Les résultats dont nous disposons à l'heure actuelle concernent majoritairement la première année d'études. Dans la présentation, nous donnerons une place privilégiée aux études de sciences économiques sachant, qu'au niveau des conclusions, nous élargirons le champ des résultats.

A titre introductif à la recherche des variables déterminantes dans le processus de réussite en sciences économiques, nous nous attacherons tout d'abord à l'examen des caractéristiques spécifiques de la population étudiante inscrite dans cette discipline.

I - CARACTERISTIQUES DE LA POPULATION ETUDIANTE

Il est possible de caractériser l'étudiant suivant plusieurs dimensions : suivant ses caractéristiques démographiques propres (âge, sexe...) suivant son origine scolaire (série du bac, année du bac...) suivant son origine socio-professionnelle (catégorie sociale des parents, revenu des parents...), suivant son origine géographique ou suivant ses performances individuelles à des tests psychologiques.

Nous examinerons successivement ces différents points.

1.La faculté de sciences économiques a une structure des séries au baccalauréat de ses étudiants relativement ouverte.

Bien que les titulaires du baccalauréat puissent formellement s'inscrire dans n'importe quelle discipline quelle que soit la série de leur baccalauréat d'origine, on observe que les U.E.R. sont relativement bien typées quant à leur structure suivant la série du bac. Ainsi, aux deux extrémités on trouve deux disciplines très caractérisées ; les lettres classiques d'une part avec uniquement des étudiants ayant un bac littéraire, les études de Deug A de sciences (mathématiques et physiques) d'autre part avec une très grande majorité de bacheliers de la série C. Entre ces deux extrêmes, l'économie a une position assez centrale avec un recrutement très varié.

Série du bac	A	B	C-E	D	F-G	TOTAL
%	3,9	24,7	27,3	29,9	14,2	100,0

2. La faculté de sciences économiques comprend une part importante d'étudiants qui ne sont pas des bacheliers de l'année mais qui effectuent une réorientation généralement après un échec. (Souvent dans les disciplines scientifiques)

On peut grossièrement distinguer deux groupes de disciplines à l'intérieur de l'Université. Celui qui comprend les disciplines dans lesquelles presque tous les étudiants sont des bacheliers de l'année et celui qui comprend les disciplines dans lesquelles une part non négligeable des effectifs n'est pas constituée de bacheliers de l'année mais d'années antérieures. Les lettres classiques, la médecine et le Deug A de sciences sont dans le premier groupe, alors que l'économie, est dans le second avec la psychologie, la philosophie et les lettres modernes. En fait seulement 63,7 % des étudiants qui s'inscrivent pour la première fois en économie sont des bacheliers de l'année. Si on examine séparément la structure des séries du baccalauréat dans les deux sous-populations (bacheliers de l'année, bacheliers d'années antérieures) en économie, on trouve une situation assez différente avec un taux beaucoup plus faible de bacheliers de la série C dans la population des bacheliers de l'année que dans la seconde population.

Il apparait donc que les bacheliers C choisissent assez rarement économie en premier choix, mais se replient volontiers sur cette discipline avec un échec dans les disciplines scientifiques.

Ce tableau est également renforcé si on sait que les bacheliers C et D qui s'inscrivent en économie l'année de l'obtention de leur baccalauréat ont eu des performances modestes au baccalauréat parmi les titulaires du bac dans les mêmes séries (ex : 7,6 en maths au bac pour la série C en économie, contre 9,7 pour l'ensemble des bacheliers C - ou bien 9,95 à l'écrit du bac pour la série D en économie contre 10,54 pour l'ensemble des bacheliers D.)

3. La population des étudiants en économie est caractérisée par une structure sociale moyenne

Alors que certaines disciplines sont caractérisées par une structure sociale élevée telle que la médecine ou modeste telle que psychologie ou lettres modernes, l'économie se trouve dans une situation intermédiaire. Toutefois, on observe que la population des étudiants en réorientation en économie est caractérisée par une structure sociale plus élevée que la population des bacheliers de l'année.

4. A part la structure par sexe (avec une majorité de garçons : 68,8 %), l'ensemble des autres caractéristiques (âge, résultats aux tests, origines géographiques...) est tel que la faculté d'économie occupe une position moyenne dans l'ensemble de l'Université

Après avoir examiné les caractéristiques essentielles de la population en économie, nous pouvons maintenant nous attacher à décrire comment s'est déroulée la première année d'études et à chercher quelles variables sont décisives dans la réussite.

II - LA REUSSITE DE 1ERE ANNEE

1. Résultats globaux

Sur les 152 étudiants en première inscription, 65 ont été autorisés à s'inscrire en deuxième année après une année d'études, soit 42,8 % de l'effectif initial. Ce taux n'est pas très élevé, mais il est toutefois légèrement supérieur à la moyenne des disciplines universitaires étudiées.

Le tableau ci-dessous donne les taux de réussite dans les différentes disciplines étudiées ainsi que la façon avec laquelle la réussite a été obtenue.

	Eco.	Méd.	MIPC	ψ	Φ	IUT	L.C.	L.M.	TOTAL
Rien passé * (1)	18,4	9,3	17,6	37,4	43,3	0,7	6,7	29,3	17,4
Seulement 1 partiel (2)	14,5	0	7,4	5,5	3,3	24,8	0	1,0	7,3
Juin et/ou sept. échec (3)	24,3	68,7	46,1	20,9	23,3	2,8	26,7	21,2	37,6
Réussite (4)	42,8	22,1	28,9	36,2	30,0	71,7	66,6	48,5	37,8
TOTAL (5)	100	100	100	100	100	100	100	100	100

* *Ces étudiants n'ont passé aucun partiel ni examen.*

Numériquement, les taux de réussite sont très différents d'une discipline à l'autre, mais peut-être plus importantes encore sont les différences quant à la manière avec laquelle la réussite a été obtenue.

Les taux de réussite sont particulièrement faibles en médecine et en M.I.P.C., qui attirent majoritairement pourtant des étudiants originaires des séries du baccalauréat les plus sélectives (C et D). De plus, dans ces disciplines, les non-réussites ont principalement pour cause ce que nous pourrions appeler des "échecs pédagogiques", c'est-à-dire des échecs consécutifs à de mauvaises notes à un contrôle de connaissances effectivement subi. A l'opposé, en lettres (à l'exception de lettres classiques), même si

les taux ne sont pas très élevés (bien que supérieurs à ceux de médecine ou M.I.P.C.), ils sont, pour une part beaucoup plus importante, obtenus avec un taux de "rien passé" relativement plus élevé, ce qui signifie que l'"élimination pédagogique" sur examen joue un rôle moindre. En effet, ces disciplines sont aussi caractérisées par un taux "d'absence" de la scène universitaire assez important.

Quant à l'économie elle se trouve globalement dans la moyenne des disciplines tant par la réussite sur l'ensemble des inscrits que par le taux des étudiants n'ayant passé aucune épreuve. Toutefois, le nombre d'étudiants ayant abandonné les études à la suite du premier partiel semble relativement important.

L'observation des modalités de la non réussite doit nous conduire à éviter de mélanger deux types de phénomènes de nature différente : celui de l'abandon, immédiatement ou rapidement après l'inscription universitaire d'une part, et celui de l'échec "pédagogique" sur contrôles de connaissances subis et résultats insuffisants, d'autre part. C'est pourquoi, bien qu'ayant opté pour une description "modélisée" du cursus de la première année, nous avons estimés des modèles de réussite sur quatre types de population.

- Modèle I - "Explication" de la réussite par rapport à l'échec sur la totalité des inscrits.
- Modèle II - "Explication" de la "présence" sur la scène universitaire (a passé au moins une épreuve) par rapport à "l'absence" (rien passé) sur la totalité des inscrits.
- Modèle III - "Explication" de la réussite par rapport à l'échec sur la population des étudiants ayant passé au moins un partiel ou un examen.
- Modèle IV - "Explication" de la note moyenne finale d'écrit pour les étudiants ayant passé l'ensemble des épreuves de contrôle.

2. Les résultats des modèles de réussite en économie

Deux modèles ont été estimés : le modèle type I qui cherche à "expliquer" la réussite (1) par rapport à l'échec (0) sur l'ensemble de la population en "première inscription", et le modèle type III qui est

Réussite Moyenne

42,8 %

Série du bac %

Série du bac	A	B	C.E.	D	F. G
Echec	66,7	52,8	38,5	55,8	95,4
Réussite	33,3	47,2	61,5	44.2	4,6

Année d'obtention du bac %

Année du bac	74	73
Echec		
Réussite	42,3	62,5

Moyenne d'écrit au bac %

Moyenne d'écrit	<10	10 - 12	>12
Echec	61,6	52,5	38,1
Réussite	38,4	47,5	61,9

Sexe %

Sexe	Masculin	Féminin
Echec	64,7	42,5
Réussite	35,3	57,5

Origine du Secondaire %

Statut de l'Etablissement Secondaire d'Origine	Public	Privé
Echec	53,6	60,9
Réussite	46,4	39,1

Age de l'Etudiant %

Age en Septembre 74	⩽18ans	19-20ans	⩾21ans
Echec	34,1	60,9	80,6
Réussite	65,9	39,1	19,4

Mode de logement. %

mode de Logement	chez les Parents	Logement Ville	CROUS
Echec	33,3	46,7	54,8
Réussite	66,7	53,3	45,2

Catégorie Socio-Professionnelle d'Origine %

Catégorie Socio-Professionnelle	Agriculteur	Artisan Commerçant	Cadre Sup. Prof. Libale	Cadre Moyen	Employé	Ouvrier
Echec	65,2	40,0	59,5	33,3	63,6	63,6
Réussite	34,8	60,0	40,5	66,7	36,4	36,4

semblable, mais limité à la population des étudiants ayant passé au moins une épreuve de contrôle des connaissances (partiel ou examen).

a) "Explication" de la réussite sur l'ensemble de la population. (Variables de base).

VARIABLES X_j	COEFFICIENTS	χ^2 (2)	VARIATIONS "MARGINALES" (1)	
			sur X_j ⟶	Sur Y réussite
Catégorie socio-professionnelle des parents	+ 0,664	3,8*	Cadre / non cadre	+ 16,3 %
"Etudes précédentes"	+ 2,418	31,4***	Non terminale (sup)/terminale	+ 59,2 %
Autre inscription	- 1,027	4,4**	Autre inscript./ pas d'autre inscript.	- 25,1 %
Test BV.17	+ 0,018	51,4***	10 points	+ 4,4 %
Test D. 48	- 0,048	43,4***	5 points	- 5,9 %
Moyenne d'écrit au bac	+ 0,322	192,4***	1 point	+ 7,9 %
Sexe	+ 1,188	10,8***	féminin/ masculin	+ 29,8 %
Bac C.E / Autres séries	+ 1,711	15,0***	Bac C.E/ autres séries	+ 41,9 %
Origine du secondaire	+ 0,598	6,2**	Public/privé	+ 14,6 %
E P I.-N |écart à la moyenne|	- 0,089	3,8*	3 points	- 6,4 %
E P I.-E |écart à la moyenne|	+ 0,216	12,2***	3 points	+ 15,8 %
Age	- 0,272	456,4***	1 an	- 6,7 %
Ressources des parents ---1.000-F.	- 0,158	9,8***	1.000 F.	- 3,9 %
Bac ⩽ 72/Bac 73.74	- 1,781	7,0***	Bac ⩽ 72/ Bac 73.74	- 43,6 %
Constante	+ 0,398	3,2*	-	-

(1) Les estimations sont effectuées au point correspondant à la probabilité moyenne (42,8 %).

(2) * : significatif au seuil de 10 %
** : significatif au seuil de 5 %
*** : significatif au seuil de 1 %

Cette discipline est caractérisée par l'influence significative d'un nombre important de variables, avec des possibilités nombreuses de se trouver dans une zone de fortes probabilités ou de faibles probabilités de réussir. Dans cet ensemble de variables, le groupe qui apparaît avoir les effets les plus importants et les plus significatifs, est constitué des variables de type scolaire : moyenne au baccalauréat, série du baccalauréat, année du baccalauréat et "études précédentes".

Les bacheliers de la série C (ou E) réussissent significativement mieux que ceux des autres séries (avec peu d'écarts entre séries non C à l'exception des séries F et G qui sont marquées par des taux extrêmement faibles). La différence entre les bacs C et les autres est très importante (41,9 %). Elle est plus forte que celle qui apparaît dans un tableau de réussite par série de bac car les bacs C en Sciences économiques ont en moyenne une note d'écrit au bac plus faible, ont plus souvent une autre inscription et sont plus souvent originaires de l'enseignement secondaire privé. Si l'effet de la série du bac est important, la façon avec laquelle le bac a été obtenu l'est aussi. Ainsi, avec 7,9 % de différence dans la probabilité de réussir, par point à la moyenne d'écrit, on arrive à un écart de 39,5 % entre un étudiant ayant 9 à l'écrit du bac et un étudiant ayant eu 14.

De plus, en Sciences économiques, une partie non négligeable (1/3) des "premiers inscrits" dans cette discipline n'est pas un bachelier "de l'année". On observe alors en prenant en compte simultanément la variable "études précédentes", qui distingue les bacheliers de l'année des autres, la variable année du bac et la variable Age, que les étudiants qui ne viennent pas de terminale (mais majoritairement de classes préparatoires et de la faculté des sciences) ont un avantage très important (+ 59,2 %). Toutefois, cet avantage est normalement amputé de l'effet de l'âge (- 6,7 % par année) et surtout de l'effet de la variable "année du bac". On arrive alors à la conclusion que les bacheliers de 73 (bac 73 n'est pas significativement différent de bac 74) qui ont fait des études scientifiques et qui se réorientent en économie ont de très grandes chances de réussir. Toutefois, cet avantage est annulé si deux ans ou plus (généralement deux échecs ou plus) séparent le baccalauréat de l'inscription en économie.

Cette mauvaise réussite, sur la population globale, s'explique sans doute par l'effet psychologique des échecs multiples antérieurs et par le découragement. En effet, ces étudiants échouent majoritairement par l'absence aux épreuves de contrôle des connaissances.

A côté de ces variables, nous trouvons le sexe, avec une réussite significativement plus élevée pour les filles que pour les garçons, toutes choses égales par ailleurs. A noter que les garçons échouent spécialement plus par abandon puisque 60 % des garçons passent les épreuves de juin ou septembre, alors que ce même chiffre est supérieur à 80 % pour les filles. Nous trouvons également comme variables significatives les résultats aux tests avec des écarts relativement peu élevés.

Enfin, le coefficient de la variable "autre inscription" indique que, toutes choses égales par ailleurs, les étudiants qui prennent une autre inscription simultanément à celle prise en économie, ont une probabilité plus faible de réussir. Cette notion de probabilité ayant "uniquement un sens statistique descriptif externe" car la majorité des étudiants ayant une autre inscription réussit lorsqu'elle passe les épreuves alors que sa façon d'échouer est essentiellement "par l'absence".

Le tableau ci-après donne les résultats de quelques simulations portant principalement sur la série du bac, la moyenne au bac et l'âge pour des étudiants bacheliers de l'année, d'origine modeste, de sexe masculin, originaires de l'enseignement public, l'ensemble des autres variables étant assignées à leurs valeurs moyennes.

%

AGE ENTREE		18 ANS			19 ANS			20 ANS		
Moyenne écrit bac.		8,0	10,0	14,0	8,0	10,0	14,0	8,0	10,0	14,0
BAC	Série C.	53,5	68,6	88,8	48,0	62,5	86,5	41,3	57,3	82,9
	Série non C.	17,2	28,4	58,9	14,3	23,2	53,6	11,3	19,5	46,8

Ces 2 chiffres calculés pour des garçons deviennent respectivement 87,7 % et 56,5 % pour les filles.

Ces 2 chiffres calculés pour des bacheliers de l'année deviennent respectivement 94,9 % et 77,2 % pour des bacheliers de 73 ayant fait des études scientifiques l'année précédente.

Après ce modèle de réussite sur la population totale des premiers inscrits, examinons les résultats d'un modèle semblable sur la population limitée à ceux des premiers inscrits qui ont passé au moins une épreuve de contrôle des connaissances.

b) *"Explication" de la réussite sur la population des étudiants ayant passé au moins une épreuve de contrôle des connaissances. (Variables de base + conditions de vie).*

VARIABLES X_j	COEFFICIENTS	χ^2	VARIATIONS "MARGINALES" (1)	
			Sur Xj ⟶	Sur Y réussite
"Etudes précédentes"	+ 3,321	128,2***	non terminale/ terminale	+ 82,8 %
Test BV.17	+ 0,027	55,9***	10 points	+ 6,8 %
Moyenne d'écrit au bac	+ 0,550	257,7***	1 point	+ 13,7 %
Sexe	+ 0,695	1,8	Féminin/ masculin	-
Bac C.E/Autres séries	+ 3,559	23,1***	Bac C.E/ autres séries	+ 88,8 %
Heures de cours, TP, TD / semaine	+ 0,120	48,2***	1 heure/Semaine	+ 3,0 %
Heures de travail universitaire hors cours/semaine	- 0,073	16,1***	1 heure/Semaine	- 1,8 %
Travail en bibliothèque	+ 0,613	11,2***	20 % du travail univers.hors cours en bibliothèque	+ 15,2 %
Ressources parents 1000 F	- 0,380	31,9***	1000 F. / mois	- 9,5 %
Bac 73/74	- 1,519	21,4***	Bac 73/74	- 37,7 %
Appartement ville/parents	- 1,543	10,7***	Appartement ville / Autres	- 38,5 %
Cité universitaire/parents	- 0,664	1,7	Cité Universi./ parents	-
Heures de loisirs organisés par semaine	- 0,169	15,7***	1 heure/semaine	- 4,2 %
Constante	- 6,066	362,8***	-	-

(1) Les estimations ont été effectuées au point correspondant à la probabilité moyenne (52,4 %). La variable âge a été omise dans l'analyse.

On observe par rapport au modèle précédent un renforcement de l'effet scolaire par l'intermédiaire des variables série du baccalauréat, moyenne à l'écrit du baccalauréat et "études précédentes".

Ce renforcement vient du fait des différences dans les populations de référence avec une origine à un moindre degré scolaire des causes de l'abandon sans notes. On remarquera aussi la disparition de la variable sexe avec des réussites semblables pour les garçons et pour les filles. (La différence significative au niveau de la population globale s'explique exclusivement par des abandons beaucoup plus fréquents chez les garçons).

Les autres variables n'ont pas d'effets inattendus ni d'effets très marquants, à l'exception de la variable "heures de travail universitaire hors cours, T.P., T.D./semaine" qui a un coefficient négatif. Nous avons déjà abordé ce problème dans la note introductive et nous n'y reviendrons pas. Les conditions de logement, et notamment le logement dans une chambre en ville, semblent avoir un effet sur la probabilité de réussite. Ce type de logement "libéral" est éventuellement susceptible de "distraire" l'étudiant de ses études. En effet, le type de logement est connu avec une bonne précision alors que l'utilisation du temps supporte une large imprécision, ce qui est susceptible de faire apparaître cette première variable très significative et de "perturber" les coefficients relatifs aux variables d'utilisation du temps.

Enfin notons que le travail en bibliothèque (généralement bibliothèque de section) est récompensé avec un coefficient relativement élevé (+ 15 % sur la probabilité lorsqu'on substitue un travail en bibliothèque à un travail en dehors des locaux universitaires pour une quantité représentant 20 % du travail universitaire hors cours, T.P., T.D. par semaine) et très significatif.

c) *"Explication" de la note d'écrit en juin pour les étudiants ayant passé l'ensemble des épreuves. (Variable âge omise)*

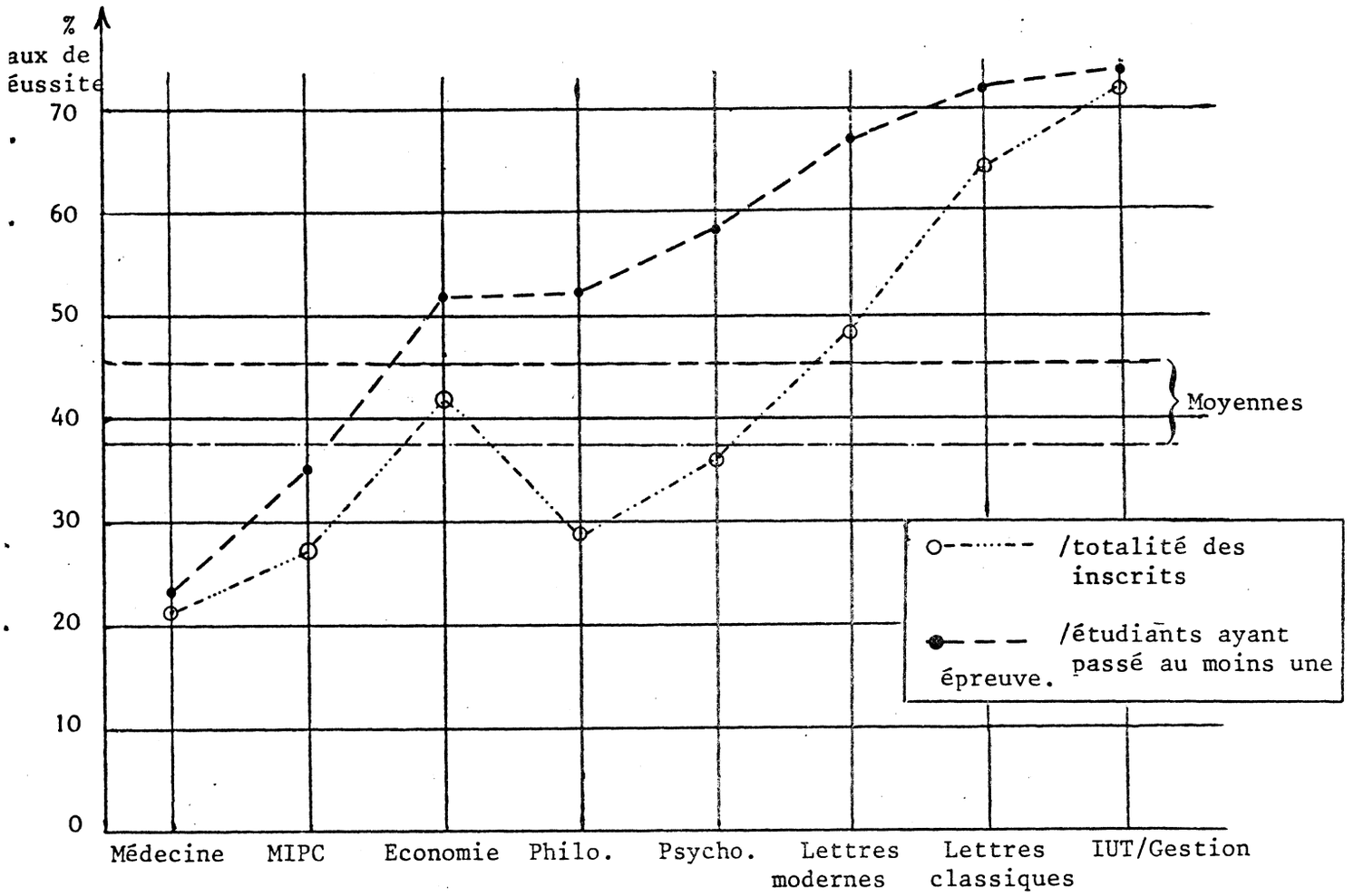
Outre la variable âge, qui est absente de l'analyse, et qui apporterait manifestement une contribution à l'explication statistique du phénomène, trois variables rendent compte d'une part relativement importante de la variance de la note moyenne obtenue. Ces variables sont celles que nous avons relevées dans les deux modèles précédents, à savoir, la série du baccalauréat, la moyenne à l'écrit du bac et la variable "études précédentes".

VARIABLES	COEFFICIENT	|t| DE STUDENT
Bac C/non C	+ 2,89	5,4 ***
Moyenne à l'écrit du bac	+ 0,56	3,8 ***
Activité 73/74 non terminale/terminale	+ 1,93	3,7 ***
Constante	+ 3,29	

R^2 = 0,426

III - QUELQUES CONCLUSIONS GENERALES QUANT AUX PROCESSUS DE SELECTION DANS LES DIFFERENTES DISCIPLINES ETUDIEES.

La première conclusion qu'on peut tirer de l'analyse est que l'Université n'est pas une institution monolithique dans ses processus de sélection. Le graphique suivant rappelle les différences globales entre disciplines.



Ces taux de réussite sont différents, mais les caractéristiques des populations inscrites dans les différentes disciplines le sont aussi, ainsi que les variables qui en rendent compte. Si on s'attache un moment à l'examen de la réussite suivant les origines-série du baccalauréat, on arrive toutes disciplines confondues, au tableau suivant :

(totalité des inscrits) %

SERIE DU BAC	A	B	C E	D	F G	Autres - Equivalence
"rien passé"	27,0	11,6	14,7	13,0	24,1	31,6
Seulement 1 partiel	3,0	23,2	4,8	4,2	19,5	26,3
juin et/ou sept.= Echec	24,1	20,5	37,7	55,9	28,7	15,8
Réussite	46,0	44,6	42,8	28,0	27,6	26,3
TOTAL	100,0	100,0	100,0	100,0	100,0	100,0

Ce tableau qui donne une hiérarchie de la réussite "dans l'ordre de l'alphabet" avec des taux élevés pour A puis B et C et des taux relativement faibles pour D, F et G renvoie pour partie à la "clientèle" des disciplines et pour partie à la structure de la réussite par discipline. Si maintenant, on examine la "hiérarchie" des séries de baccalauréat à l'intérieur des différentes U.E.R. de l'échantillon, on arrive à une image tout à fait différente.

SERIE DU BAC	A	B	C E	D	F G	Autres - Equivalence
Economie	33,3	47,2	61,5	44,2	4,6	-
Médecine	0	0	39,8	15,6	0	0
Deug A - MIPC	-	-	35,3	0	0	-
Psychologie	41,9	6,6*	40,0	37,8	35,3	20,0*
I.U.T./ G.E.	85,7*	61,5	85,0	83,3	63,0	66,7*
Lettres modernes	53,7	0*	66,7*	20,0*	-	25,0*
ENSEMBLE	46,0	44,6	42,8	28,0	27,6	26,3

* Petits effectifs.

Au contraire du tableau précédent, les taux de réussite les plus élevés sont obtenus par la série C puis ensuite avec un écart important par la série D. De plus, on observe que l'effet de la série du baccalauréat est surtout important dans les disciplines scientifiques (surtout M.I.P.C. puis médecine) pour être relativement très réduit dans les disciplines littéraires. Ceci atteste du caractère "ouvert" de ces disciplines (à un moindre degré économie) et permet de comprendre les procédures d'accès aux différentes filières et de réorientattions (structures de "choix descendants").

Toutefois s'il paraissait évident, ou pour le moins probable, que la série du baccalauréat serait un "prédicteur efficace" de la réussite universitaire, il était aussi clair que d'autres caractéristiques avaient une part dans l'"explication statistique" de la réussite. Des analyses multivariées du type de celle présentée pour les sciences économiques permettent alors d'aboutir aux résultats globaux suivants :

1. Il y a un certain nombre de spécificités dans les modes de réussite suivant les différentes disciplines, mais dans chacune d'entre elles on peut noter un effet notable des variables d'ordre scolaire.

- La série du baccalauréat est très discriminante dans les disciplines scientifiques et moins (peu) dans les disciplines littéraires.

- La moyenne d'écrit au bac est également très significative avec des résultats d'autant meilleurs à l'Université que le bac a été obtenu de façon plus brillante.

Entre un bac obtenu avec 9 à l'écrit et un bac obtenu avec 13 également à l'écrit, les écarts des probabilités de réussite à l'avantage des seconds s'établissent, toutes choses égales par ailleurs, comme suit :

Médecine	I.U.T./GE	Lettres mod.	Psycholo.	Economie	Deug A. M.I.P.C.
+ 36,8 %	+ 17,2 %	+ 31,6 %	+ 26,8 %	+ 31,6 %	+ 28,0 %

2. Dans toutes disciplines, les plus jeunes réussissent mieux. (A l'exception d'I.U.T.) les écarts de probabilité de réussite entre un étudiant de 20 ans et un étudiant de 18 ans, à l'avantage du second s'établissent, toutes choses égales par ailleurs, comme suit :

Médecine	I.U.T./GE	Lettres mod.	Psycholo.	Economie	Deug A. M.I.P.C
+ 12,4 %	-	+ 28,8 %	+ 5,8 %	+ 13,4 %	+ 12,8 %

3. Influence sélective de la catégorie socio-professionnelle des parents et du statut (public/privé) de l'établissement secondaire d'origine. Si on met de côté l'effet indirect de la catégorie socio-professionnelle d'origine sur la série du baccalauréat, et par conséquent sur la réussite, on observe un effet net direct de l'origine sociale seulement dans les disciplines littéraires (psychologie et lettres modernes). - Comme si l'inefficacité des séries du bac à prédire la réussite était remplacée par l'origine sociale-. Cette différence qui tient au type des matières enseignées est également renforcée par l'effet du statut (public/privé) de l'établissement secondaire d'origine. En effet, dans toutes les disciplines non littéraires, les étudiants originaires du privé réussissant significativement moins bien, toutes choses égales par ailleurs, que ceux originaires du public. Par contre cette relation est inexistante en psychologie et s'inverse en lettres modernes où les étudiants du privé réussissent significativement mieux.

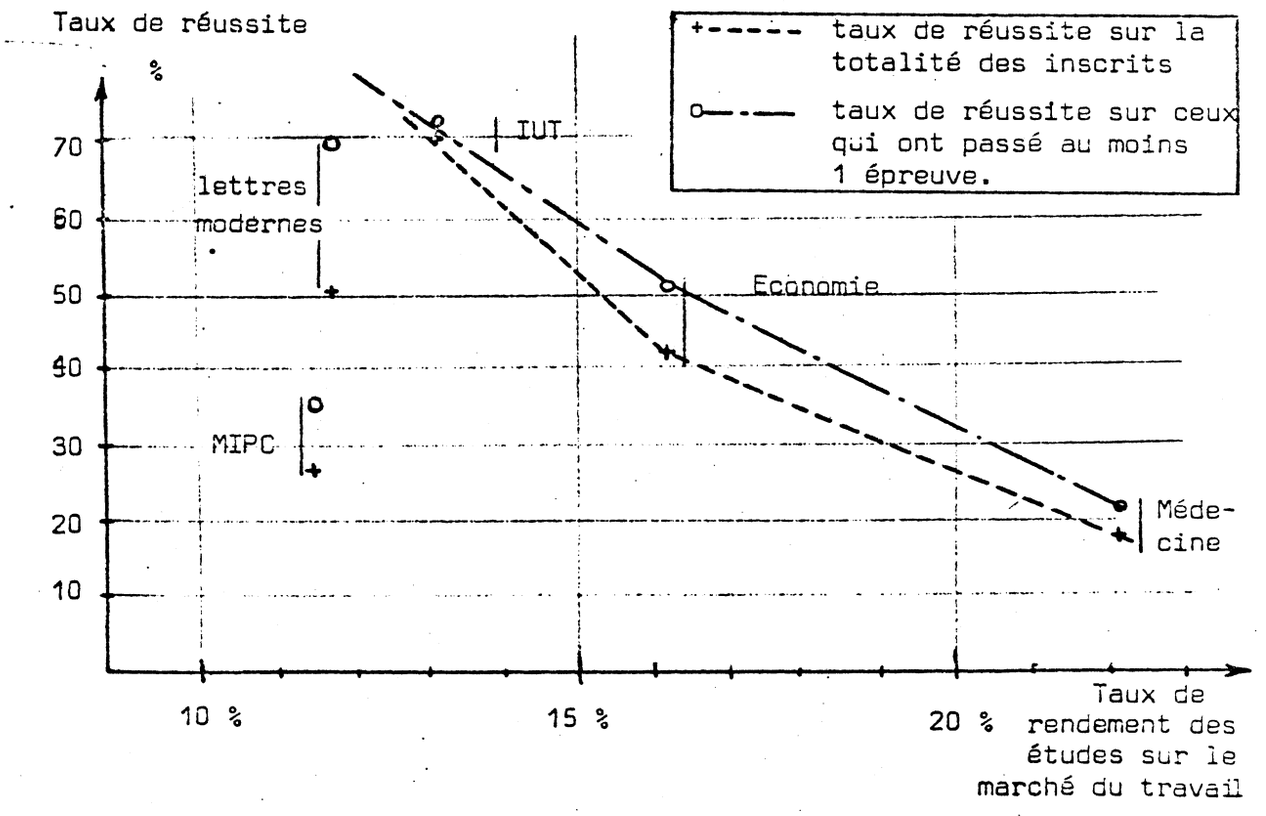
4. Les résultats aux tests (logique, verbal, personnalité) n'apportent que peu d'éclairages à la réussite universitaire sauf en ce qui concerne le test de personnalité avec des résultats meilleurs pour les étudiants introvertis que pour les étudiants extravertis (I.U.T.-psychologie-M.I.P.C.).

5. Les variables de conditions de vie (qui ne sont pas exemptes d'imprécision quant à leur mesure) apportent une contribution assez modeste à l'explication. Les seules variables qui paraissent importantes sont le logement -avec des réussites meilleures pour les étudiants qui habitent chez leurs parents plutôt qu'en cité universitaire ou en chambre en ville- et le mode de prix des repas -avec un effet significativement positif du restaurant universitaire sur la réussite- (hypothèse sous-jacente est que plus on utilise le restaurant universitaire plus on est présent à l'université et moins on perd de temps).

Pour achever ces quelques remarques de conclusion, nous nous contenterons de poser la question de savoir dans quelle mesure les procédures d'accès à l'Université et les procédures de réussite sont liées. En effet, il apparaît évident que ces deux niveaux sont réciproquement interdépendants.

Dans une certaine mesure, les étudiants sont conscients de la sélection et "arbitrent" entre les différentes disciplines en en tenant compte, et, comme une image de cette situation, la sélection s'exerce vraisemblablement en fonction des caractéristiques de la population inscrite dans chacune des disciplines. Toutefois, dans la mesure où l'enseignement supérieur n'est pas uniquement une consommation mais présente aussi très nettement un caractère d'investissement, les étudiants attachent aussi une valeur à l'emploi qu'ils occupent, ou pensent occuper à l'issue de leurs études.

Le graphique ci-dessous met en relation le taux de rendement monétaire des différentes disciplines du supérieur[1] (en 1970) avec le taux de réussite en première année d'études universitaires.



[1] L. LEVY-GARBOUA et A. MINGAT : "Les taux de rendement privés et sociaux de l'éducation en France". (année 1970).

Il apparaît nettement que les études de M.I.P.C. sont atypiques. En effet, les étudiants y cumulent un taux de sélection très élevé, avec des espérances de gains relativement modestes (concurrence des écoles) sachant que ces étudiants sont très majoritairement titulaires de baccalauréats de la série la plus sélective C. Cette accumulation est-elle responsable des échecs des différents plans visant à développer les études scientifiques ? Toujours est-il que la réussite en science est très aléatoire, surtout face aux espérances de gains des diplômés et à la qualité des étudiants qui s'engagent dans cette filière.