Université Panthéon-Assas

Master 2

Ingénierie Statistique et Financière

Insurance Econometrics
Logistic regression

Joseph Lanfranchi, 2021-22

Introduction

In this lecture, we want to deal with two issues:

- **How to deal with qualitative variables modelling in the SAS System**

First, we will focus on applications of your lecture with Ali Skalli about qualitative variables econometrics.

These applications will cover various forms of choice: binary, categorical and non categorical, ordered evaluation of choices, countable events. Also, Sas will be used as an econometric tool for modelling these choices.

- **Highlight these methods in the field of insurance economics**
- Also, these applications will be chosen from the field of insurance economics at large. The scope of study will cover the factors explaining the decisions between various forms of insurance, of complementary insurance and the effect of social insurance on individual's health behaviour.

Outline of the lectures

INGÉNIERIE STATISTIQUE FINANCIÈRE
UNIVERSITE PANTHEON ASSAS

- **Discrete choice modelling:**
  - Logit (or probit) modelling of choice of life insurance
  - Multinomial logit modelling of long-term care insurance
  - Ordered Logit or Probit modelling of choice of the extent of insurance cover
- **Count data**
  - Poisson and negative binomial modelling of the number of doctor visits in relation with the access to Medicaid or Medicare Insurance systems

Discrete choice models: The PATER Survey

- Our first econometric modelling of insurance choice is an attempt to explain the French households' choices in terms of contract with insurance companies. We will deal with their decisions to sign a death insurance contract and/or to sign life insurance contract.
- Unparalleled database: The Pater surveys (PATrimoine et Préférences vis-à-vis du Temps et du Risque), have been carried out in May 2007, June 2009, November 2011 and 2014, among more than 3 600 households representative of the French population.
- The survey includes a socio-demographic description of the household, the valuation, composition and management of its assets, a census of its income, and the existence of intergenerational transfers received or paid.
- As of 2007, the survey includes a greater number of questions aimed at measuring individual preferences with respect to risk, time and lineage, based on lottery choices, Likert scales but also on attitudes, opinions or behaviours in different areas of life.

PATER Survey

- Every questionnaire includes a series of preference scales obtained using different revelation methods.
- In addition to the « usual » methods proposed by the empirical literature, these scales come from an original approach, based on a scoring procedure, developed in 1998 and perfected since then:
- questions sweeping through various domains of life make it possible for the authors of the survey to evaluate ordinal, synthetic and coherent indicators concerning each respondent's attitudes towards risk and uncertainty, his or her degree of family altruism and his or her preference for present consumption.

Theoretical foundations (I)

- Household wealth accumulation is basically explained by the life cycle model. In this saving model, households derive satisfaction only from the consumption achieved in each period of their life, which in turn determines the savings chosen over the life-cycle.
- Furthermore, in addition to their life-cycle savings, explained by the distribution of consumption over time, households have a desire to pass on their wealth, a function of their degree of family altruism, θ.
- The model assumes the maximization of an intertemporal utility function U, additive in instantaneous utilities $u_t(C_t)$, whose present value being discounted at an exponential rate of preference for the present δ, a rate that depends on age but but not on distance from the present.
- .

Theoretical foundations (II)

- Instantaneous utility has a constant generic form over time:

$$u(C) = {C^{1-\gamma}}/{(1-\gamma)}$$

Where γ represents the degree of concavity of the function u. Inverting γ gives the intertemporal elasticity of substitution. That one measures the elasticity of C_{t+1}/C_t to the relative price of present consumption in terms of future consumption (u_{t+1}/u_t) .

- When the relative price of future consumption increases by 1%, the future consumption decreases by 1/ γ%.
- it can be interpreted as the desire to smooth consumption over time.
- In uncertainty, the household maximizes his expected utility and this parameter is viewed as the constant relative risk aversion degree, which influences the household prudence (precautionary saving).

Theoretical foundations (III)

- γ conditions the intertemporal smoothing of consumption and attitude in face of risk. The level of wealth increases with it, for precautionary reasons.
- Moreover, the shared of risky assets decreases with risk aversion.
- For a given lifetime, δ reduces the foresight degree of the households and their horizon for decision. Hence, rise of time preference decreases the value of wealth as it diminishes the life cycle saving and retirement saving.
- Finally, θ increases wealth as the degree of family altruism raises the saving devoted to intergenerational transfers.

Theoretical foundations (IV)

- This basic model does not predict accurately insufficient retirement saving and rationalizes more diversified portfolios than the ones observed in reality. For example, life annuity rent is too rare and so are stock options.
- Non standard but more realistic models take into account optimism, loss aversion, ambiguity aversion, all necessitating further new parameters in the utility function.
- Agents are assumed to have limited rationality: they are impatient, subject to emotions and sensitive to suffered costs.
- These theories require to take into account a high number of parameters for the preferences, requiring an empirical estimation that is almost impossible and with a highly variable.

Scoring in PATer

- Considering how difficult and doubtful the estimation of these parameters could have been, the survey managers have preferred an evaluation of the three types of preferences (risk, present and altruism) and did not try to asses the extent of risk aversion, prudence, loss aversion, optimism and so on…
- Multiplication of questions gave rise for each type of preferences to a large number of indicators that could be used to assess the tastes of the households.
- A large scope of domains is covered by the indicators: leisure, financial assets, family, health, retirement. Their wording is simple and relative to everyday life, allowing for an easy understanding. Some are devoted to behaviours, others to opinions or intentions.
- Statistical validation gave rise to final score, selecting the questions correlated to a common unknown factor.

Tableau 3
Questions parmi les plus contributives pour le score d'attitude vis-à-vis du risque

Top ten 2009	Classement			Tolérance au risque (en %)		
	2007	2009	2011	Élevée (- 1)	Faible (+ 1)	Moyenne (0)
Précaution contre une météo incertaine	1	1	1	44,7 45,8 41,5	53,3 52,8 56,6	2,0 1,4 1,8
Dépasse la vitesse auto-risée, ne met pas sa ceinture, passe au feu orange	2	2	2	17,5 14,9 15,7	32,3 37,0 37,2	50,2 48,1 47,1
Gare son véhicule en dehors des zones autorisées	3	3	3	8,1 8,2 8,1	53,1 36,9 37,2	38,8 55,0 54,7
Désir de se priver pour vivre plus longtemps	6	4	4	7,1 7,6 7,2	16,3 16,2 14,8	76,7 76,2 78,0
Pense que l'homogamie est un critère de longévité pour le couple (même revenu, même milieu social, même sensibilité politique, même religion…)	7	5	5	23,5 26,3 27,8	33,6 31,3 30,2	42,9 42,4 42,0
« Le mariage est une assurance »	8	6	6	17,7 20,9 20,0	10,5 9,0 8,5	71,8 70,1 71,5
Pense qu' « être propriétaire, c'est l'assurance d'avoir toujours un toit au-dessus de sa tête »	4	7	7	16,9 14,8 15,6	35,7 34,3 37,7	47,4 51,0 49,7
Conseille aux proches de prendre des risques professionnels	12	8	8	10,0 16,0 16,3	6,1 4,8 4,9	83,1 79,3 78,8
A pris des risques dans son comportement professionnel et/ou ses pratiques sportives et/ou ses pratiques sexuelles	9	9	9	24,4 22,2 21,7	33,9 36,7 36,9	41,7 41,1 41,4
Être plutôt du genre à aller se faire vacciner même quand la vaccination n'est pas obligatoire	11	10	10	24,9 26,3 27,9	20,6 18,3 19,3	54,5 55,4 52,8

Lecture: 44,7 % des ménages ne prennent pas de précautions quand la météo est incertaine en 2007 (45,8 % en 2009 et 41,5% en 2011). Ils sont donc classés parmi les plus tolérants au risque (note : - 1 dans le score).
Champ : échantillon total représentatif de la population française.
Source : enquêtes Pater 2007, 2009 et 2011.

Tableau 4
Questions parmi les plus contributives pour le score de préférence temporelle

Top ten 2009	Classement			Vision de long terme		
	2007	2009	2011	Courte (+ 1)	Longue (- 1)	Moyenne (0)
« La retraite c'est quelque chose qui se prépare longtemps à l'avance »	2	1		21,3	21,7	57,0
				24,4	21,2	54,4
			3	21,6	23,7	54,7
Souci du maintien de la forme	3	2		13,8	21,6	64,6
				15,5	21,0	63,5
			2	16,3	18,9	64,8
Il faut inculquer à ses enfants le goût de l'épargne	1	3		4,9	43,4	51,7
				4,3	45,2	50,5
			1	3,9	47,6	48,5
Prépare ses vacances longtemps à l'avance	5	4		12,6	22,3	65,1
				14,0	20,0	66,0
			4	11,2	22,2	66,7
Désir de se priver pour vivre plus longtemps	7	5		7,1	16,3	76,7
				7,6	16,2	76,2
			11	7,2	14,8	78,0
Approuve des enfants qui privilégient leurs loisirs par rapport à leurs études	4	6		4,2	37,2	58,7
				4,8	35,0	60,2
			6	3,6	37,1	59,3
Est quelqu'un qui fait généralement des projets	11	7		31,6	64,9	3,5
				37,9	59,6	2,5
			9	37,5	59,7	2,8
Intéressé par un retrait précoce du marché du travail contre une pension réduite après 65 ans	10	8		27,1	18,7	54,3
				25,4	23,2	51,4
			12	22,0	21,4	56,6
Préoccupé par le risque de finir sa vie dans une maison de retraite	12	9		_	42,9	57,1
				_	44,2	55,8
			4	_	40,9	59,1
Prend ses billets à l'avance et arrive à l'avance pour prendre le train ou l'avion	14	10		8,8	27,7	63,6
				8,3	29,0	62,7
			5	6,8	32,4	60,8

Lecture : 21,7 % des ménages pensent que « la retraite, c'est quelque chose qui se prépare longtemps à l'avance » en 2007 (21,2 % en 2009 et 23,7 % en 2011). Ils sont classés comme des ménages voyant le plus à long terme (note : - 1 dans le score).
Champ : échantillon total représentatif de la population française.
Source : enquêtes Pater *2007, 2009 et 2011.*

Tableau 5
Questions parmi les plus contributives pour le score d'altruisme familial

Top ten 2009	Classement			Altruisme familial (en %)		
	2007	2009	2011	Faible (-1)	Fort (+1)	Neutre (0)
« Transmettre à ses descendants » est une raison d'épargner importante	1	1		19,0	24,7	56,3
				22,0	22,6	55,5
			1	23,7	22,2	54,1
« Une fois les enfants élevés, les parents peuvent dépenser l'argent qu'ils possèdent comme bon leur semble, quitte à ne pas laisser d'héritage »	2	2		53,9	44,0	2,1
				54,0	44,2	1,0
			2	54,0	44,2	1,8
Favorable à un allègement des droits de succession pour les enfants	9	3		_	90,9	9,1
				_	91,1	8,9
			4	_	88,8	11,2
Cherche prioritairement à transmettre à ses enfants le sens de la famille	7	4		_	29,8	70,2
				_	29,2	70,8
			7	_	30,2	69,8
« Le mariage, c'est pour le meilleur et pour le pire »	5	5		17,5	80,5	2,1
				17,5	80,8	1,7
			5	18,3	79,9	1,8
Il faut inculquer à ses enfants le goût de l'épargne	4	6		4,7	44,1	51,2
				4,3	45,2	50,5
			6	3,9	47,6	48,5
Sa famille profiterait de son gain éventuel de 300 000 €	3	7		_	37,4	62,6
				_	48,6	51,4
			3	_	46,1	53,9
Considère qu'il faut aider ses enfants tout au long de leur vie	8	8		16,8	20,1	63,1
				22,2	21,6	56,3
			8	20,0	22,7	57,3
« Avoir des enfants, c'est s'engager pour la vie »	6	9		5,5	92,4	2,2
				5,0	93,5	1,4
			9	5,1	93,1	1,7
Gestion plus prudente du patrimoine hérité	10	10		7,6	89,9	2,5
				7,8	90,3	1,9
			10	9,7	88,4	1,9

Lecture : 24,7 % des ménages pensent que « Transmettre à ses descendants » est une raison d'épargner importante en 2007 (22,6 % en 2009 et 22,2 % en 2011). Ils sont classés comme des ménages à fort altruisme (note : + 1 dans le score).
Champ : échantillon total représentatif de la population française.
Source : enquêtes Pater *2007, 2009 et 2011.*

Some comments about the scores

- When analyzing the individual determinants of the scores, they seem generally consistent with previous knowledge: men are more tolerant to risk than women, and so are the young compared to old; also, married people and children of business owners are less risk averse. However, the level of education favours ore risky behavior only in 2 out of 5 of the surveys.

- Also, the respondent is always more forward looking (low preference for present) if he is older, more educated and lives as a couple. However, the higher foresight of women is only true in the last three waves of the survey.

C10_ter. Parmi les types de placements financiers suivants, quels sont ceux que vous détenez ou que détient l'un des membres de votre foyer ?

Sous variable	Dénomination	Type
C10_ter_1	Aucun 125ociet du foyer ne détient de placements financiers	0/1
C10_ter_2	Compte ou plan épargne logement (CEL, PEL)	0/1
C10_ter_3	LivretA, LivretBleu	0/1
C10_ter_4	Autres comptes ou livrets d'épargne	0/1
C10_ter_5	Plan d'épargne retraite populaire (PERP)	0/1
C10_ter_6	Plan d'épargne retraite (PEP-PER)	0/1
C10_ter_7	Épargne retraite complémentaire volontaire	0/1
C10_ter_8	Epargne Salariale (Plan Epargne Entreprise –PEE, FCPE).	0/1
C10_ter_9	Contrats d'assurance-vie en euros (investis sur des supports non risqués et garantis en capital)	0/1
C10_ter_10	Contrats d'assurance-vie en Unité de Compte (investis sur des actions ou SICAV/FCP)	0/1
C10_ter_11	Contrats d'assurance-décès volontaires	0/1

Age

Valeurs	Dénomination
1	Moins de 25 ans
2	25-34 ans
3	35-44 ans
4	45-54 ans
5	55-64 ans
6	65-74 ans
7	75 ans et plus

Sexe

Valeurs	Dénomination
1	Homme
2	Femme

Niveak

Valeurs	Dénomination
1	N'a jamais fait d'etudes
2	Etudes primaires
3	Enseignement secondaire (6e a 3e)
4	Technique court (CAP,BEP,...)
5	2e, 1ere, Niveau Bac ou Brevet Profes
6	Technique superieur (IUT,BTS
7	Superieur 1er cycle
8	Superieur 2eme cycle
9	Superieur 3eme cycle

B1. Combien d'enfants avez-vous... ?

Valeurs	Dénomination
0-11	Nombre d'enfant
-1	Non réponse

B2. Et combien vivent encore au domicile... ?

Valeurs	Dénomination
0-6	Nombre d'enfant
-1	Non réponse

C21. Actuellement, quelle est approximativement la valeur du patrimoine global (financier ou non, incluant le cas échéant votre logement...) que vous possédez seul ou en commun avec un membre de votre foyer, sans en déduire votre endettement ?

Valeurs	Dénomination
1	Moins de 8 000 €
2	De 8 000 à 14.999 €
3	De 15 000 à 39 999 €
4	De 40 000 à 74 999 €
5	De 75 000 à 149 999 €
6	De 150 000 à 224 999 €
7	De 225.000 € à 299.999 €
8	De 300 000 à 449 999 €
9	450 000 à 749 999 €
10	750 000 € et plus
-1	Non réponse

Revenu_net

Valeurs	Dénomination
1	Moins de 300 euros
2	300 à 600 euros
3	601 à 900 Euros
4	901 à 1 200 Euro
5	1.201 à 1.500 Euros
6	1.501 à 1 900 Euros
7	1 901 à 2 300 Euros
8	2 301 à 2 700 Euros
9	2 701 à 3 000 Euros
10	3 100 à 3 800 Euros
11	3 801 à 5 300 Euros
12	5 301 Euros et plus
13	Non renseigné

Linear Probability Model (LPM)

- The Linear Probability Model (LPM) is an econometric model where the explained variable is the probability that a binary event happens ($y_i = 1$) and is a linear function of the dependent variables. The LPM predicts the probability of an event occurring.
- Here, we can model the holding of a life insurance contract as a linear regression model:

$$y_i = \beta' x_i + u_i \qquad with \qquad E(u_i) = 0.$$

- Basically, five assumptions underlie Ordinary Least Square method. Assumption 1 says that y is a linear function of x plus a random disturbance term u, for all members of the sample. while assumption 2 is that $E(u) = 0$, meaning that the expected value of u does not vary with x, implying that x and u are uncorrelated.

- If those two assumptions hold, ordinary least squares will produce unbiased estimates of β.

- A first problem with using LPM to explain a binary event, coded with a dichotomous variable, is the apparent difficulty of interpretation of the parameters β. In fact, a 1-unit change in x produces a change of β in the probability that $y = 1$.

Linear Probability Model (LPM)

- The remaining basic hypothesis are:
- The homoscedasticity assumption says that the variance of u is the same for all observations:

$$var(u_i) = \sigma^2$$

- The no correlation assumption between error terms says that the random disturbance for one observation is uncorrelated with the random disturbance for any other observation:

$$Cov(u_i, u_j)=0$$

- The last assumption says that the random disturbance is normally distributed.
- If assumptions 1 and 2 are true, then the error terms can only take on two values and are not normally distributed. The normality assumption is false!
- The normality assumption is not needed if the sample is large. The central limit theorem assures that coefficient estimates will have an approximately normal distribution even when u is not normally distributed. That means that we can still use a normal table to calculate p-values and confidence intervals. If the sample is small, these approximations could be poor.

Heteroscedasticity

- As you know from Ali Skalli's lecture, the variance of the error term is equal to the variance of y, and the variance of a dummy is the product :

$$var(u_i) = E(y_i)[1 - E(y_i)] = \beta' x_i(1 - \beta' x_i)$$

Therefore, this proves that the variance of u_i will change with observations and the values of x_i. Because of heteroscedasticity, the OLS estimators are no longer efficient. This means that there are alternative methods of estimation with smaller standard errors.

Besides, the standard error estimates are no longer consistent estimates of the true standard errors. That means that the estimated standard errors could be biased to unknown degrees and the test statistics could also be biased.

Heteroscedasticity

- Fortunately, this problem is solved in the REG procedure in SAS, as we can include the option HCC in the MODEL instruction. Then, following the correction proposed in Huber (1967) and White (1980), the variance-covariance matrix will be corrected from heteroscedasticity.

```
PROC REG DATA=paterscore;
MODEL model ass_vie = age2534 age3544 age4554 age5564 age6574 age75sup
homme diplome pres_enfant patfin patrfintmiss scorar scopt scoralt/ HCC;
RUN;
```

- The option SPEC performs a specification test of the model. The null hypothesis maintains the homoscedasticity assumption and the independence of regressors and errors. When errors are independent from explanatory variables, rejection of the null hypothesis confirms heteroscedasticity.

Odds

- Before considering the logit model, let us examine one specific concept, the odds of an event. The odds of an event is the ratio of the expected number of times that an event will occur to the expected number of times it will not occur.
- An odds of 4 means we expect 4 times as many occurrences as non-occurrences. An odds of 1/5 means that we expect only one-fifth as many occurrences as non-occurrences.
- A simple relationship relies P the probability of an event and O the odds of the event :

$$O = \frac{P}{(1-P)} \text{ and so } P = \frac{O}{(1+O)}$$

- Note that odds less than 1 correspond to probabilities below .5, while odds greater than 1 correspond to probabilities greater than .5. Like probabilities, odds have a lower bound of 0, but unlike probabilities there is no upper bound on the odds. Hence, you can easily double the odds for example. A probability of .60 corresponds to odds of .60/.40=1.5. Doubling that yields odds of 3 and a probability of.75.

Odds examples

	Age 55		
Ass_vie	0	1	Total
0	1292	894	2186
1	663	767	1430
Total	1955	1661	3616

For all households, the probability of life insurance is given by :
P=(663+767)/(1292+894+663+767)=0,396

The odds of the event life insurance is:
O=(663+767)/(1292+894)=1430/2186=0,65

For heads of households older than 55, the odds of the event life insurance is
O=767/894=0,86

For heads of households younger than 55, the same odds is O=663/1292=0,51

Odds ratios is a widely used measure of the relationship between two dichotomous variables.

The ratio of the old odds to the young odds is 0,86/0,51=1,67 (odds ratio). The odds of a life insurance contract in an "old" household are 67% greater than for "young" households.

Logit and Probit Models

- The assumption of the LPM that there will be a straight linear relationship between Independent and dependent Variables is very questionable when the dependent variable is a binary one. It is more reasonable to assume a nonlinear function of X, one which approaches 1 at slower and slower rates as Xi gets larger and larger.
- The models probit and logit explain the probability of an event, conditionally to the values of observed exogenous variables. Hence the following model :

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

where $F(.)$ is a cumulative distribution function.

- Choosing the cumulative distribution function $F(.)$ is free and the most commonly used, the cumulative distributive function of the logistic distribution and the cumulative distributive function of the standard normal distribution, give birth respectively to logit and probit models:

$$\text{logit model: } p_i = \Lambda(x_i\beta) = \frac{1}{1+e^{-x_i\beta}}, \text{for all } i$$

$$\text{probit model: } p_i = \Phi(x_i\beta) = \int_{-\infty}^{x_i\beta} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz, \text{for all } i$$

Logit and Probit Models (II)

- It is possible to compare of the estimated values of parameters obtained with the LPM, logit and probit models.
- Amemiya proposed approximated relationships between those estimated values:

$\hat{\beta}_{LPM} = 0{,}25\hat{\beta}_L$ for the slopes and $\hat{\beta}_{LPM} = 0{,}25\hat{\beta}_L + 0{,}5$ for the constant term.

$$\hat{\beta}_L = 1{,}6\ \hat{\beta}_P$$

- Those approximated relationships are quite precise when the sample does not include outliers or extreme values (that is when the average values of $xi\beta$ is close to zero or if the average values of probabilities are close to 0,5).

LOGISTIC procedure

- The Proc LOGISTIC is one among four of the procedures estimating dichotomous models, allowing for different distribution. It can perform logit, probit and complementary log-log models (the last being associated with the Gumbell or double exponential distribution). The default in LOGISTIC is the logit model, but it is sufficient to change the LINK option to change the model (LINK=probit). It has a large set of features used by data analysts.
- A central option is the option EVENT='1' in the MODEL statement, after the dependent variable. By default, the proc LOGISTIC estimates a model that predicts the weakest value for the explained variable. Hence, this means that for every binary variable, the proc LOGISTIC estimates a model that predicts the probability that the explained event does not realize.
- We can also use the DESCENDING option in the model statement which tells LOGISTIC to model the "higher" value of the binary variable.

Comparison of estimated parameters

Variable	Logit	Probit	LPM
Age55	0,492***	0,475***	0,404***
Homme	0,128*	0,124*	0,103*
Diplôme	0,122***	0,118***	0,101***
Patrimoine	0,0030***	0,0029***	0,0025***
Enfants à domicile	-0,099**	-0,089**	-0,075**
Enfants indépendants	-0,064*	-0,062*	-0,048
Aversion risque	0,0089	-0, 0089	-0,0075
Préférence présent	-0,062***	-0,060***	-0,051***
Altruisme familial	0,037**	0,036**	0,030**

LOGISTIC procedure (II)

- The proc LOGISTIC realizes linear tests of the estimated coefficients. The code of those tests requires to give numbers to the tests and then to write the « to be tested » linear formula.
- It is also possible to save within the original file the model predicted probability and the odds for all the individuals in the sample (households). Hence, we can evaluate all statistical characteristics of these new variables.
- Finally, it is also possible to run the regressions within subsamples. In our program, we can estimate the model for all households whose heads are older and younger than 55.

Distinctive characteristics of Logit model (I)

- **The logisitic distribution tends to attribute to extreme events (rare) a higher probability than the normal distribution.**

- The logistic density function has thicker tails than the tails of the normal distribution. Even if those two distributions belong to the same family of exponential distributions, the profile of those two distributions differs for the extreme values of the support: extreme values are less likely for the normal distribution.

- Economically speaking, choosing the logit model is equivalent to give a higher probability to « extreme » events, in comparison with a probit model based upon a normal distribution.

Distinctive characteristics of Logit model (II)

- **The logit model implies a specific interpretation of the parameters β associated with explanatory variables.**
- From the logistic distribution, we can deduce the log-odds form of the model:

$$\log\left(\frac{p_i}{1-p_i}\right) = x_i\beta \text{ ou } e^{x_i\beta} = \frac{p_i}{1-p_i} = O_i$$

- Then the odds O_i is equal to the quantity $e^{x_i\beta}$. Hence, we can calculate the estimated odds for each observation and check if this corresponds to the realization of the explained event.
- Also, we can express the marginal effect of any covariate *j* on the odds:

$$O_i = \frac{p_i}{1-p_i} = \exp\left(\sum_{k=1}^{K} xi^k\,\beta_k\right) = \prod_{k=1}^{K} \exp\left(xi^k\,\beta_k\right)$$

Distinctive characteristics of Logit model(III)

- Hence the share of the change due to a unitary change in the j th explanatory variable is:

$$\overline{O_i} = \exp\left[\left(xi^j + 1\right)\beta_j\right]\prod_{k\neq j}^{K}\exp(xi^k\beta_k) = \exp(\beta_j)\prod_{k=1}^{K}\exp(xi^k\beta_k)$$

- That means that the new value of the odds is equal to the product of the previous odds by the exponential of the coefficient associated with the explanatory variable *j:*

$$\overline{O_i} = \exp(\beta_j)O_i \text{ or } \overline{O_i} - O_i = \exp(\beta_j)\,O_i - O_i = \left[\exp(\beta_j) - 1\right]O_i$$

- Finally, we show that if f is the probability density function of the residuals of the model, the marginal effect associated with the j-th explanatory variable is defined as:

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

As the density is always positive, the sign of this marginal effect is the same as the sign of the estimated coefficient β_j. The other way to write this marginal effect:

$$\frac{\partial p_i}{\partial xi^j} = p_i(1 - p_i)\beta_j$$

Interpretation of the coefficients (I)

- One of the peculiarities of logistic regression is the intuitive meaning of estimations. They're not as easy to interpret as coefficients in the linear probability model. For the linear probability model, a coefficient of .25 tells you that the predicted probability of the event increases by .25 for every 1-unit increase in the explanatory variable.
- By contrast, a logit coefficient of .25 tells you that the log-odds increases by .25 for every 1-unit increase in the explanatory variable. But what does a .25 increase in the log-odds means?
- The basic problem is that the logistic model assumes a nonlinear relationship between the probability and the explanatory variables. The change in the probability for a 1-unit increase in an independent variable varies according to where you start, i.e. the value of $x's$. Things become much simpler, however, if we think in terms of odds rather than probabilities.

Estimated coefficients on the odds of a life insurance contract

Variable	Logit	Intervalle de confiance	
Age55	1,636	1,330	2,011
Male	1,137	0,979	1,319
Diploma	1,130	1,081	1,181
Wealth	1,003	1,003	1,003
Children at home	0,906	0,828	0,991
Independent Children	0,938	0,872	1,009
Risk Aversion	1,009	0,995	1,023
Preference for present	0,940	0,917	0,962
Family Altruism	1,038	1,008	1,068

Non standardized estimated coefficients of the life insurance contract model

Variable	Logit	Probit	LPM
Age55	0,492***	0,297***	0,101***
Male	0,128*	0,078*	0,026*
Diploma	0,122***	0,074***	0,025***
Wealth	0,0030***	0,0018***	0,0006***
Children at home	-0,099**	-0,056**	-0,019**
Independent Children	-0,064*	-0,039*	-0,012
Risk Aversion	0,0089	-0, 0056	-0,0019
Preference for present	-0,062***	-0,037***	-0,013***
Family Altruism	0,037**	0,023**	0,0074**

Interpretation of the coefficients (II)

- We should look at the numbers in the "Odds Ratio Estimates" table, which are obtained from the parameter estimates by computing exp(estimated coefficient in the logit model). These can be seen as estimated ceteris paribus odds ratio because they control for other variables in the model.

- The estimated odds ratio of 1,636 tells us that the predicted odds of a life insurance contract for « more than 55 years old » households, are 1,636 times the odds for younger households.

- In other words, the odds of a life insurance contract for "old" heads of household are 64% higher than the odds for other households. This is the best estimate of the effect of this variable.

Interpretation of the coefficients (III)

- How interpreting the coefficient for the variable SCOPT (time preference), which is statistically significant at the .01 level? Recall that this variable is a score measured on a real scale.
- For quantitative variables, it's helpful to take the formula from the marginal change for odds, subtract 1 from the odds ratio and multiply by 100, that is, calculate $100(exp\beta - 1)$. This tells us the **percent change** in the odds for each 1-unit increase in the independent variable. In this case, we find that a 1-unit increase in the SCOPT score is associated with a 6% decrease in the predicted odds of possession of life insurance contract.
- Note that if an estimated coefficient is significantly different from 0, then the corresponding odds ratio is significantly different from 1. There is no need for a separate test for the odds ratio.

Marginal effects (I)

- If one seeks to interpret logistic models in terms of probability, one should calculate marginal effects. To do this, we have to use the marginal effect formula

$$\frac{\partial p_i}{\partial xi^j} = p_i(1 - p_i)\beta_j$$

- This equation says that the change in the probability for a 1-unit increase in x^j depends on the logistic regression coefficient for x^j, as well as on the value of the probability itself. For this to be useful, we need to know what probability we are starting from.
- All possible choices are valid, but if we have to choose one value, the most natural is the overall proportion of cases that have the event. Hence, we will evaluate what is the change in the average choice when one of the explanatory variable is changed.

Marginal effects (II)

- In our example, 1430 out of 3616 households possess a life insurance contract, so the overall proportion is 39.55%. Taking p*(1-p)=0.3955*(1-0.3955), we get 0.239. Hence, we can multiply each of the coefficients in the output table by 0,239.
- For the variable "patrimoine", the estimated coefficient is 0,030. We then multiply this coefficient to obtain the marginal effect of estate ("patrimoine") 0.030*0.239= 0.00717. This means that on average, the probability of an household to possess a life insurance contract is increased by 0,7 percentage point when the estate is increased by a thousand euros. The probability will therefore be equal to 0,3955+0,0072=0,4027.

Marginal effects (II)

- For the dummy variable age55, the raise in probability is obtained by multiplying the associated coefficient 0,492 by 0,239, that is 0,117. We can then say that, on average, the probability of a life insurance contract is .117 higher if the head of the household is older than 55 compared with younger than 55 years old head of household.
- Note that this effect is not very far from the results of the Linear Probability Model, as the estimated coefficient was equal to 0.101.
- The same is true for the coefficient associated with "homme". The marginal effect estimated with the logistic model is around 0,030, while the estimated coefficient in the LPM is 0.026.

Marginal effects (III)

- Using the option MARGINAL in the proc QLIM proposed another calculation of marginal effects. Instead of choosing a single value for the probability, we can calculate a predicted probability for each individual. Then we can use the derivative formula to generate marginal effects for each individual.
- The proc QLIM will indeed obtain the same coefficients, standard errors and statistical tests as those obtained with proc LOGISTIC. Including an option OUT will save the marginal effects for each household.
- PROC PRINT produces a table for the first n (here 10) observations in the output data set. For each variable, we get the predicted change in the probability of possessing a life insurance contract for a 1-unit increase in that variable, for a particular individual based on that individual's predicted probability.
- So, we may calculate the average value of those individual marginal effects. This a kind of average treatment effect, different from the treatment effect evaluated at the average of the sample.

Categorical explanatory variables

- PROC LOGISTIC has a CLASS statement that allows to specify that a variable should be treated as categorical. When a CLASS variable is included as an explanatory variable in the MODEL statement, LOGISTIC automatically creates a set of "design variables" to represent the levels of that variable.
- Example: recoding the variables associated with the level of diplomas in 5 classes. This new variable DIPBIS takes integer values from 1 to 5, 5 being the highest diploma or « at least master degree ».
- The instruction CLASS identifies DIPBIS as a categorical variable. The PARAM=REF option tells LOGISTIC to create a set of four dummy (indicator) variables, one for each value of DPBIS except the highest one (DIPBIS=5) as a reference.
- LOGISTIC estimates four coefficients associated with four created dummy variables, one for each of the values 1 through 4. Thus, each of the four coefficients for DIPBIS is a comparison between that particular value and the highest value. More specifically, each coefficient can be interpreted as the log-odds of the life insurance contract for that particular value of DPBIS minus the log-odds for DPBIS=5, controlling for other variables in the model.

Estimation du rapport de cotes			
Effet	Estimation du point	Intervalle de confiance de Wald à95%	
age55	1.603	1.304	1.970
homme	1.148	0.989	1.333
dipbis 1 vs 5	0.557	0.443	0.702
dipbis 2 vs 5	0.744	0.588	0.943
dipbis 3 vs 5	0.805	0.616	1.051
dipbis 4 vs 5	0.965	0.650	1.433
patrimoine	1.003	1.003	1.003
nbr_enfdom	0.905	0.827	0.990
nbr_enfind	0.937	0.871	1.008
scorar	1.060	0.970	1.159
scopt	0.793	0.726	0.867
scoralt	1.107	1.022	1.198

dipbis	5	1	0.5846	0.1176	24.6970	<.0001
dipbis	4	1	0.5493	0.1929	8.1033	0.0044
dipbis	3	1	0.3672	0.1247	8.6758	0.0032
dipbis	2	1	0.2892	0.0972	8.8472	0.0029

Categorical explanatory variables (II)

- When you have a CLASS variable in a model, LOGISTIC provides an additional table, labeled "Type 3 Analysis of Effects." For CLASS variables, on the other hand, it gives a test of the null hypothesis that all of the coefficients pertaining to this variable are 0. In other words, it gives us a test of whether DPBIS has any impact on the probability of the possession of life insurance contract. In this case, we clearly have strong evidence that DPBIS makes a difference. Note that this test is invariant to the choice of the omitted category.
- The second table gives the coefficients and the tests of equality to zero. The pattern for the four coefficients is just what one might expect. Heads of households with DPBIS=1 are much less likely to possess a life insurance contract than those with DPBIS=5. Each increase of DPBIS is associated with an increase in the probability of a life contract.

Categorical explanatory variables (III)

- To illustrate the differences between levels of diploma, we can modify the reference category. If we want it to be the lowest value of DIPBIS, just use CLASS DIPBIS / PARAM=REF DESCENDING; If you want it to be some particular value, say 3, use CLASS DIPBIS (REF='3') / PARAM=REF;
- We may assess the effects of two categories of a variable like DIPBIS=2 and DIPBIS=3. It is easy to use a TEST statement or a CONTRAST statement. The following statements tests the null hypothesis, that there is no difference in the coefficients for DIPBIS=2 and DIPBIS=3: **contrast 'dipbis2 vs. dipbis3' dipbis 0 1 -1 0;**

- If we include the DESCENDING option, the reference catgory is DIPBIS=1, the weakest of all categories. For the same test, it is then necessary to test the equality between the fifth and fourth category in descending order, dipbis5, dipbis4, dipbis3, dipbis2:

  contrast 'dipbis2 vs. dipbis3' dipbis 0 0 -1 1;

- The instruction TEST is more straightforward: to refer to an estimated coefficient, we add to the variable name the number of the category to be tested:

  test dipbis2=dipbis3;

Test deux à deux

dipbis	1 vs 3	0.693	0.542	0.884
dipbis	2 vs 3	0.925	0.719	1.190
dipbis	4 vs 3	1.200	0.800	1.800
dipbis	5 vs 3	1.243	0.952	1.623

Exemple 1

Résultats des tests de contraste

Contraste	DDL	Khi-2 de Wald	Pr > Khi-2
dipbis2 vs. dipbis3	1	0.3692	0.5434

Résultats des tests des hypothèses linéaires

Libellé	Khi-2 de Wald	DDL	Pr > Khi-2
Test 1	0.3692	1	0.5434

Exemple 3

Résultats des tests de contraste

Contraste	DDL	Khi-2 de Wald	Pr > Khi-2
dipbis2 vs. dipbis5	1	6.0008	0.0143

Résultats des tests des hypothèses linéaires

Libellé	Khi-2 de Wald	DDL	Pr > Khi-2
Test 1	6.0008	1	0.0143

Multiplicative Terms in the MODEL Statement

- Interaction terms in econometric models helps to test if the effect of one variable depends on the level of another variable. The most popular way of doing this is to include a new explanatory variable in the model, one that is the product of the two original variables.
- With PROC LOGISTIC, rather than creating a new variable in a DATA step, you can specify the product directly in the MODEL statement. For example, some authors consider that the portfolio composition depends both of the time horizon, and then of age, and of the number of descendants. Besides, the age effect may be more serious with more children in the household.
- We can test that hypothesis for the Pater data with this program that allows to specify both the interaction and the two main effects:

```
model ass_vie (event='1') = homme dipbis patrimoine nbr_enfdom scorar scopt
scoralt age55 | nbr_enfind;
```

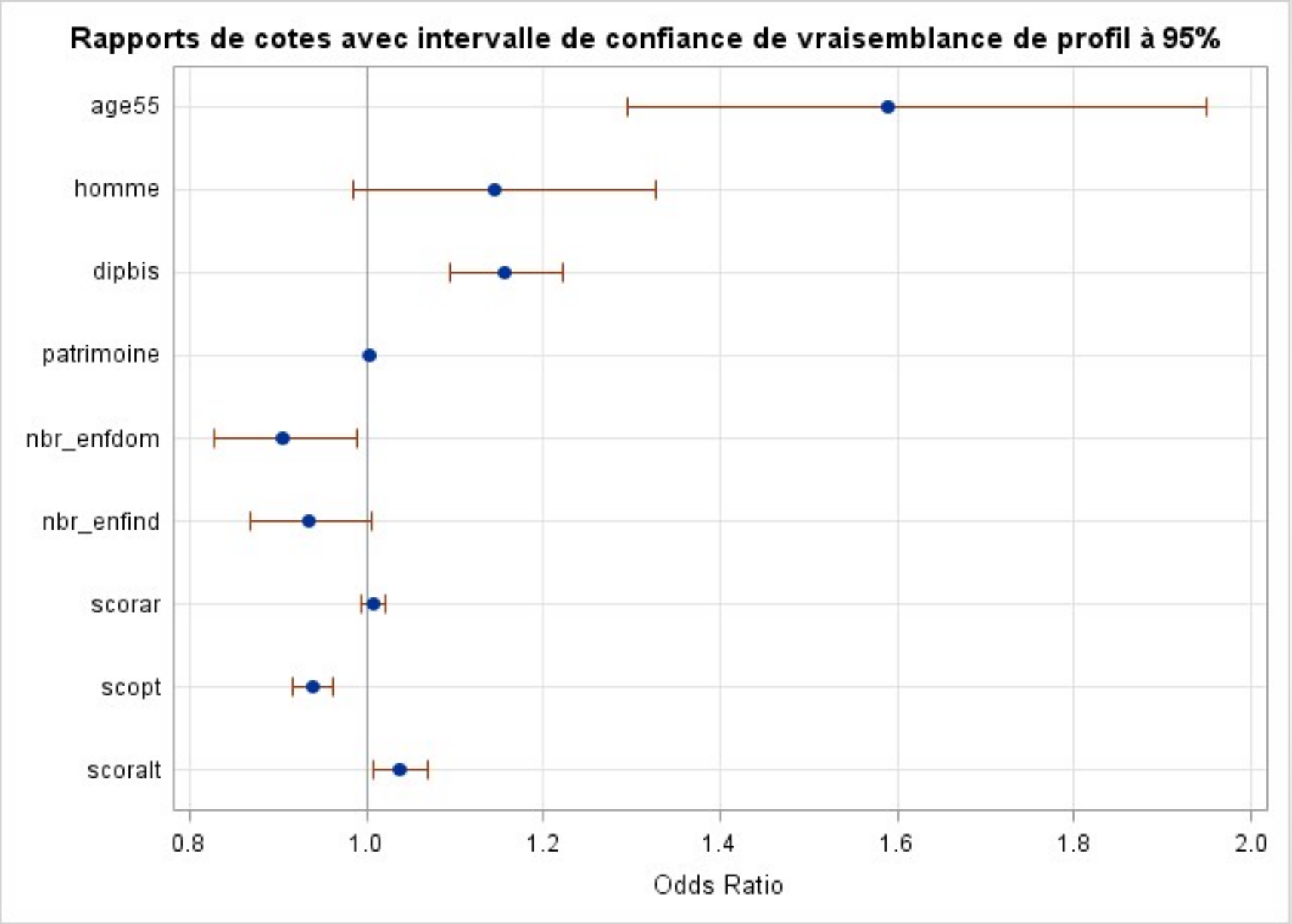
age55	1	0.4656	0.1104	17.7964	<.0001
homme	1	0.1347	0.0761	3.1363	0.0766
dipbis	1	0.1457	0.0282	26.6516	<.0001
patrimoine	1	0.00307	0.000205	225.0334	<.0001
nbr_enfdom	1	-0.0980	0.0480	4.1667	0.0412
nbr_enfind	1	-0.0682	0.0374	3.3364	0.0678
scorar	1	0.00851	0.00688	1.5307	0.2160
scopt	1	-0.0630	0.0122	26.7068	<.0001
scoralt	1	0.0374	0.0149	6.3106	0.0120
age55*nbr_enfdom	1	-0.0124	0.1528	0.0066	0.9353

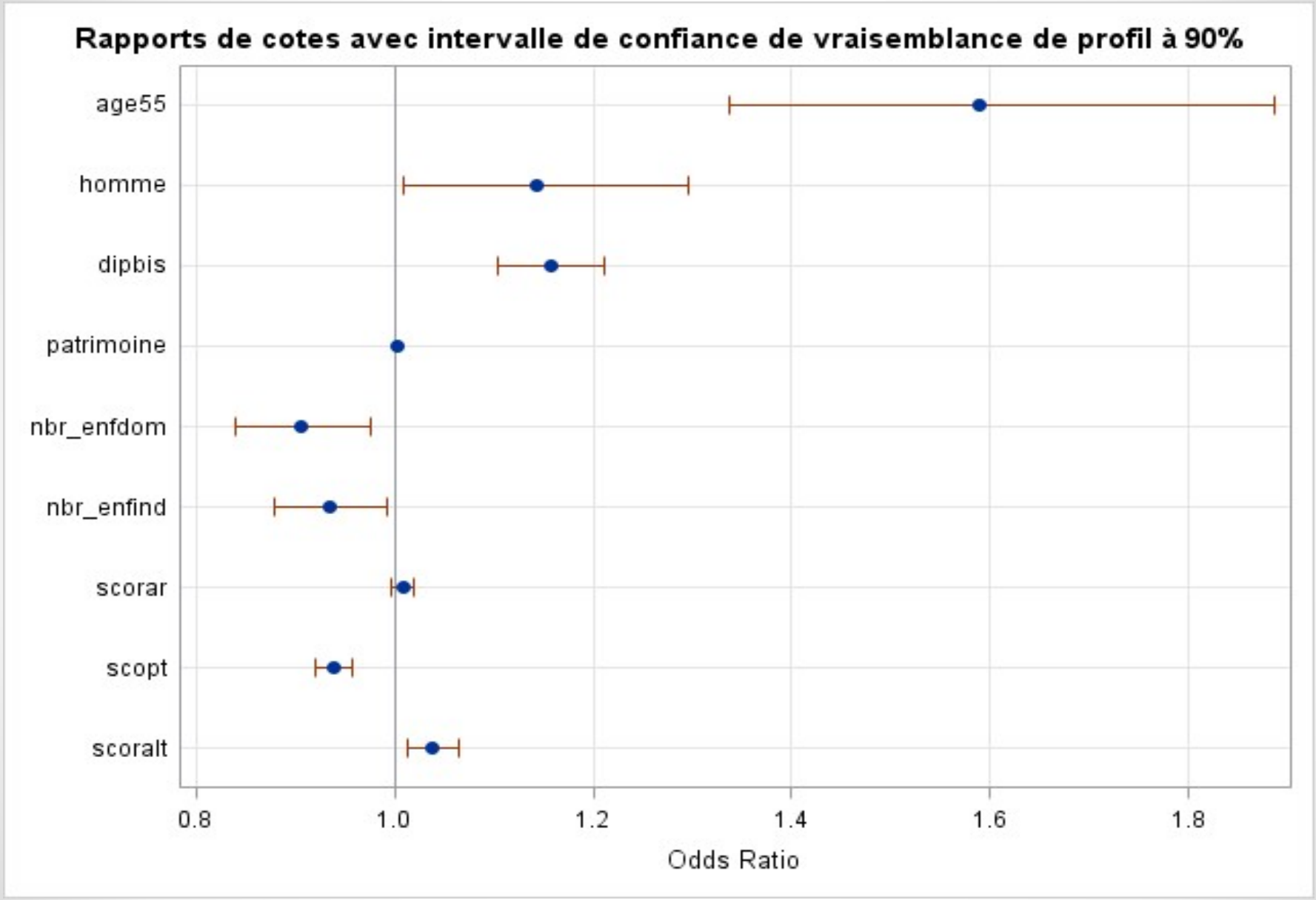
age55	1	0.4135	0.1189	12.0976	0.0005
homme	1	0.1335	0.0761	3.0796	0.0793
dipbis	1	0.1420	0.0285	24.7799	<.0001
patrimoine	1	0.00309	0.000206	224.9716	<.0001
nbr_enfdom	1	-0.1047	0.0462	5.1271	0.0236
nbr_enfind	1	-0.1360	0.0873	2.4248	0.1194
scorar	1	0.00844	0.00687	1.5078	0.2195
scopt	1	-0.0635	0.0122	27.0480	<.0001
scoralt	1	0.0368	0.0149	6.1108	0.0134
age55*nbr_enfind	1	0.0828	0.0957	0.7486	0.3869

Confidence intervals

- Confidence intervals give a better picture of the sampling variability of the estimates. PROC LOGISTIC automatically produces 95% confidence intervals for the odds ratios, but you may also want them for the regression coefficients.
- In LOGISTIC, the option in the MODEL statement for conventional (Wald) confidence intervals is CLPARM=WALD. To change the statistical threshold in order to evaluate a 90% interval, put the option ALPHA=.10 on the MODEL statement.
- The procedure LOGISTIC has another method, called profile likelihood confidence intervals, that may produce better approximations in smaller samples. This method involves an iterative evaluation of the likelihood function. In LOGISTIC, the model option is CLPARM=PL (for profile likelihood). The profile likelihood method is computationally intensive so you may need to use it sparingly for large samples.
- If we want both Wald and profile likelihood confidence intervals, we can use the option CLPARM=BOTH.
- Finally, the proc allows us to change the unit change to get the odds ratio. For example, to get the odds ratio for a 2-unit increase in DPBIS: units dipbis=2 / default=1;

Estimated values of parameters and Wald confidence intervals			
Parameter	Estimation	95% Confidence Interval	
Intercept	-2.0030	-2.2628	-1.7432
age55	0.4627	0.2579	0.6675
homme	**0.1346**	**-0.0145**	**0.2836**
dipbis	0.1456	0.0903	0.2009
patrimoine	0.00307	0.00267	0.00347
nbr_enfdom	-0.0992	-0.1889	-0.00944
nbr_enfind	**-0.0680**	**-0.1409**	**0.00497**
scorar	0.00853	-0.00493	0.0220
scopt	-0.0630	-0.0869	-0.0391
scoralt	0.0374	0.00822	0.0666


Rapports de cotes avec intervalle de confiance de vraisemblance de profil à 95%
age55
homme
dipbis
patrimoine
nbr_enfdom
nbr_enfind
scorar
scopt
scoralt
0.8
1.0
1.2
1.4
1.6
1.8
2.0
Odds Ratio


Rapports de cotes avec intervalle de confiance de vraisemblance de profil à 90%
age55
homme
dipbis
patrimoine
nbr_enfdom
nbr_enfind
scorar
scopt
scoralt
0.8
1.0
1.2
1.4
1.6
1.8
Odds Ratio

Estimates of the odds ratio and confidence interval				
Effect	unit	Estimated value	95% confidence interval	
Age55	1.0000	1.588	1.294	1.950
Homme	1.0000	1.144	0.986	1.328
Dipbis	**2.0000**	**1.338**	**1.198**	**1.495**
Patrimoine	1.0000	1.003	1.003	1.003
nbr_enfdom	1.0000	0.906	0.827	0.990
nbr_enfind	1.0000	0.934	0.868	1.005
scorar	1.0000	1.009	0.995	1.022
scopt	1.0000	0.939	0.917	0.962
scoralt	1.0000	1.038	1.008	1.069

Multicollinearity (I)

- The logistic regression shares with the linear regression pleasant and unpleasant features. One of these is multicollinearity, which occurs when there are strong linear dependencies among the explanatory variables.
- if two or more variables are highly correlated with one another, it's hard to get good estimates of their distinct effects. Although multicollinearity doesn't bias the coefficients, it does make them more unstable. Standard errors may get large, and variables that appear to have weak effects may actually have quite strong effects.
- How do you diagnose multicollinearity? Examining the correlation matrix produced by PROC CORR may be helpful but is not sufficient. It's quite possible to have data in which no pair of variables has a high correlation, but several variables together may be highly interdependent.
- Good diagnostics are produced by PROC REG with the options TOL and VIF. But PROC LOGISTIC doesn't have these options. Remember is that multicollinearity is a property of the explanatory variables, not the dependent variable. So whenever you suspect multicollinearity in a logit model, just estimate the equivalent model in PROC REG and request the collinearity options.

Multicollinearity (II)

- Two indicators, tolerance and variance inflation, are available in the REG procedure. The tolerance is computed by regressing each variable on all the other explanatory variables, calculating the R2, then subtracting that from 1. Low tolerances correspond to high multicollinearity. While there's no strict cutpoint, we maybe concerned when the tolerances are below .40.
- The variance inflation factor is simply the reciprocal of the tolerance. Hence we may be concerned with multicolinearity if it is superior to 2.5. It tells how "inflated" the variance of the coefficient is, compared to what it would be if the variable were uncorrelated with any other variable in the model.
- In the next example, there is no obvious problem of multicollinearity. Nonetheless, trying to introduce two variables measuring the same thing, diplôme and dipbis, we show that both tolerance and inflation of variance will explode.

Estimated values of parameters and colinearity							
Variable	DDL	Estimated values of parameters	Standard error	T test value	Pr > |t|	Tolerance	Inflation of variance
Intercept	1	0.03085	0.03152	0.98	0.3278	.	0
age55	1	0.10180	0.02135	4.77	<.0001	0.49596	2.01631
diplome	1	0.02490	0.00459	5.43	<.0001	0.75506	1.32440
patrimoine	1	0.0006289 9	0.0000367 4	17.12	<.0001	0.84116	1.18883
nbr_enfdom	1	-0.01905	0.00901	-2.12	0.0345	0.74498	1.34232
nbr_enfind	1	-0.01262	0.00744	-1.70	0.0899	0.61425	1.62799
scorar	1	0.00147	0.00137	1.07	0.2840	0.67959	1.47148
scopt	1	-0.01281	0.00244	-5.24	<.0001	0.68715	1.45528
scoralt	1	0.00741	0.00298	2.48	0.0130	0.84993	1.17657

Estimated values of parameters and colinearity							
Variable	DDL	Estimated values of parameters	Standard error	T test value	Pr > |t|	Tolerance	Inflation of variance
Intercept	1	0.01760	0.04136	0.43	0.6704	.	0
age55	1	0.10090	0.02145	4.70	<.0001	0.49131	2.03539
diplome	**1**	**0.02453**	**0.01333**	**1.84**	**0.0658**	**0.08949**	**11.17412**
homme	1	0.02583	0.01545	1.67	0.0947	0.94481	1.05841
dipbis	**1**	**0.00098548**	**0.01666**	**0.06**	**0.9528**	**0.09470**	**10.55930**
patrimoine	1	0.00062133	0.0000370 2	16.78	<.0001	0.82814	1.20753
nbr_enfdom	1	-0.01868	0.00901	-2.07	0.0383	0.74404	1.34402
nbr_enfind	1	-0.01192	0.00745	-1.60	0.1099	0.61205	1.63386
scorar	1	0.00187	0.00139	1.35	0.1786	0.65916	1.51708
scopt	1	-0.01264	0.00245	-5.16	<.0001	0.68578	1.45819
scoralt	1	0.00742	0.00299	2.48	0.0131	0.84773	1.17961

Model Goodness-of-Fit Statistics

- The first table to judge the performance of the model is the table of the goodness-of-fit statistics.
- The table reports three different "Model Fit Statistics": AIC, SC, and -2 Log L. The values of these fit statistics are displayed for two different models, a model with an intercept but no predictors, and a model that includes all the specified predictors.
- The most fundamental of the fit statistics, -2 Log L, is simply the **maximized value of the logarithm of the likelihood function multiplied by –2**. Higher values of -2 Log L mean a worse fit to the data but there is no absolute standard for what's a good fit, so one can only use this statistic to compare different models fit to the same data set.
- We should also keep in mind that the overall magnitude of this statistic is heavily dependent on the number of observations. The other two fit statistics avoid this problem by penalizing models that have more covariates
- **Akaike's Information Criterion (AIC)** : $AIC = -2\log L + 2k$, where k is the number of parameters (including the intercept).
- **The Schwarz Criterion (SC)**: this criteria gives a more severe penalization for additional parameters, by multiplying their number by the number of observations;

  $SC = -2\log L + nk$ where n is equal to the number of observations.

- Both of the penalized statistics can be used to compare models with different sets of covariates. The models being compared do not have to be nested in the sense of one model being a special case of another.

Testing Global Null Hypothesis

- LOGISTIC's global chi-square addresses the question, "Is this model better than nothing?" A significant chi-square signals a "yes" answer, suggesting that the model is acceptable.

- Output of the model gives three χ^2 statistics for a similar test of the the same null hypothesis—that all the explanatory variables have coefficients of 0. The first column displays the value of the χ^2 test statistic, the second the number of degrees of freedom, equal to the number of variables minus the constant and the third the p-value.

- The first statistic is the **likelihood ratio** obtained by comparing the log-likelihood for the fitted model with the log-likelihood for a model with no explanatory variables (intercept only). The proc LOGISTIC reports –2 Log L for each of those models, and the chi-square is just the difference between those two numbers.

- **The score statistic** is a function of the first and second derivatives of the log-likelihood function that follows a χ^2 distribution under the null hypothesis.

- **The Wald statistic** is a function of the coefficients and their covariance matrix.

- In large samples, there is no reason to prefer any one of these statistics. However, In small samples or samples with extreme data patterns, there is some evidence that the likelihood ratio chi-square is superior.

Analysis of Maximum Likelihood estimates (I)

- We will examine now the deviance χ^2, which answers the question, "Is there a better model than this one?" Again, a significant χ^2 corresponds to a "yes" answer, and that leads to rejection of the model.

- The deviance statistic is described as a goodness-of-fit statistic. Such statistics involve a comparison between the model of interest and a "maximal" model. The maximal model is often referred to as the saturated model. A saturated model is one in which there are as many estimated parameters as data points. By definition, this will lead to a perfect fit to the data.

- The question is whether the difference in fit could be explained by chance.

- As a likelihood ratio statistic, the deviance is equal to twice the positive difference between the log-likelihood for the fitted model and the log-likelihood for the saturated model. With individual-level data, the log-likelihood for the saturated model is necessarily 0, so the deviance is just –2 times the log-likelihood for the fitted model.

Analysis of Maximum Likelihood estimates (II)

- This deviance does not always follow a χ^2 distribution, as the number of parameters in the saturated model increases with the number of observations. Anyhow, If the number of explanatory variables is small and each variable has a small number of values, you can use the AGGREGATE and SCALE options in LOGISTIC to get a deviance that does have a chi-square distribution:

```
proc logistic data=paterscore;
model ass_vie (event='1') = age55 homme dipbis patrimoine nbr_enfdom
nbr_enfind scorar scopt scoralt / AGGREGATE SCALE=NONE;
run;
```

- In our model, the presence of both the « patrimoine » variable and especially of the three continuous scores means that we have a lot of different values for this last variables. Hence, the deviance test is not likely to be of a high quality.

Statistique d'adéquation de la déviance et de Pearson				
Critère	Valeur	DDL	Valeur/DDL	Pr > Khi-2
Ecart (deviance)	4261.7811	3589	1.1875	<.0001
Pearson	3660.1766	3589	1.0198	0.1998

Analysis of Maximum Likelihood estimates (III)

- As it is common in economics to deal with continuous variables, neither the deviance nor the Pearson chi-square will have true chi-square distributions. We may prefer to use the Hosmer and Lemeshow test (2000) that has rapidly gained widespread use.
- It may be implemented in LOGISTIC with the LACKFIT option in the MODEL statement.
- The Hosmer-Lemeshow (HL) statistic is calculated by generating the predicted probabilities by the model for all observations. These are sorted by size, and then grouped into approximately 10 intervals. Within each interval, the expected number of events is obtained by adding up the predicted probabilities.

Analysis of Maximum Likelihood estimates (III)

- These expected frequencies are compared with observed frequencies by the conventional Pearson chi-square statistic. The degrees of freedom is the number of intervals minus 2.
- A high p-value indicates that the fitted model cannot be rejected and leads to the conclusion that the model fits well. That is, it can't be significantly improved by adding non-linearities and/or interactions.

Test d'adéquation de Hosmer et de Lemeshow		
Khi-2	DDL	Pr > Khi-2
12.0150	8	0.1505

Statistics Measuring **Predictive Power**

- Another class of statistics describes how well you can predict the dependent variable based on the values of the independent variables. This is a very different criterion from the goodness-of-fit measures that we've just been considering.

- It is entirely possible to have a model that predicts the dependent variable very well, yet has a terrible fit as evaluated by the deviance or the HL statistic. Nor is it uncommon to have a model that fits well, as judged by either of those goodness-of-fit statistics, yet has very low predictive power.

- Proc LOGISTIC only calculates one of the many R^2 measures. It is based on the likelihood ratio chi-square for testing the null hypothesis that all the coefficients are 0 which is the statistic reported by LOGISTIC under the heading "Testing Global Null Hypothesis: BETA=0."
- If we denote that statistic by L2 and let n be the sample size, the generalized R^2 is:

$$R^2 = 1 - exp\left\{-\frac{L^2}{n}\right\}$$

- Although this is easy to compute with a hand calculator, LOGISTIC will do it for you if you put the RSQ option on the MODEL statement.

Statistics Measuring **Predictive Power (III)**

- It is a generalization of the conventional R^2 and has several things going for it:
  - It is based on the quantity being maximized, namely the log-likelihood.
  - It's readily obtained with virtually all computer programs because the loglikelihood is nearly always reported by default.
  - It never diminishes when you add variables to a model.
  - The calculated values are usually quite similar to the R^2 obtained from fitting a linear probability model to dichotomous data by ordinary least squares.
- This R^2 has the possible drawback that its upper bound is strictly less than 1 because the dependent variable is discrete. To fix this, one divides this R^2 by its upper bound and obtains the "Max-rescaled Rsquare,".

R squared	0.1503	Max Rescaled R square	0.2035

Statistics Measuring **Predictive Power** (IV)

- PROC LOGISTIC reports four ordinal measures of association by default whenever you run a binary (or ordinal) logistic regression.
- The measures are calculated the following way: there are 3616(3615)/2= 6535920 different ways to pair up the observations. Of these, we keep the pairs with different observed value for the dependent variable.
- When the household who contracted a life insurance contract has a predicted value higher than the predicted value for the other household without life insurance contract, this pair of observed households is said to be concordant. In the contrary, the pair is said to be discordant. If the two cases have the same predicted value, we call it a tie.

Association des probabilités prédites et des réponses observées			
Concordant percentage	72.7	D de Somers	0.456
Discordant Percentage	27.0	Gamma	0.458
Percentage of ties	0.0	Tau-a	0.218
Pairs	3125980	c	0.728

Statistics Measuring **Predictive Power** (V)

- the number of concordant pairs *C*.
- the number of discordant pairs *D*.
- the number of ties *T*.
- the total number of pairs *N*.

- The four measures of association are then defined as follows:
- Somers' D: D de Somers ou indice de la justesse des prédictions $= \frac{C-D}{C+D+T}$
- Gamma: Γ de Goodman-Kruskal $= \frac{C-D}{C+D}$
- Tau-a: τ_a de Kendall= $\frac{C-D}{N}$
- c: c de Hanley and McNeil 1982 = 0,5 (1+ D de Somer)
- Those four measures play an identical role. Their value lies between 0 and 1 and the higher their value, the better the predictive power of the model.

Quality of prediction and ROC curve

- Another approach to evaluating the predictive power of models for binary outcomes is the ROC curve. To understand it, we must understand classification tables.

- Let $\widehat{p_i}$ be the predicted probability that $y_i = 1$ for individual i, based on some model that we have estimated. If we want to use the predicted probabilities to generate actual predictions of whether or not $y_i = 1$, we need some cutpoint value. A natural choice would be .5. If $\widehat{p_i} \geq 0{,}5$, we predict $y_i = 1$. If $\widehat{p_i} < 0{,}5$, we predict $y_i = 0$.

Quality of prediction and ROC curve

Tableau de classification									
Niveau de proba.	Correct		Incorrect		Pourcentages				
	Evénement	Non-événement	Evénement	Non-événement	Correct	Sensi-bilité	Spéci-ficité	Faux POS	Faux NEG
0.500	606	1897	289	824	69.2	42.4	86.8	32.3	30.3

- The reading is the following: we predict 606 cases correctly:

	$\widehat{p_i} \geq 0.5$	$\widehat{p_i} < 0.5$	Total
$y_i = 1$	606	824	1430
$y_i = 0$	289	1897	2186

Quality of prediction and ROC curve (II)

- The overall proportion of predictions that are correct seems good: (1897+606)/(1430+2186)=69,2. But a correct estimation can be misleading:

- Suppose that a data set has 100 events and 900 non-events. A model with no predictors will generate predicted values that are all .10, and thus all the cases would be predicted as non-events. This model would be right 90 percent of the time, but is not especially predictive!

- To do better, we first define *sensitivity* and *specificity:*

- **Definition of sensitivity** : proportion of events correctly predicted, in this case 606/1430=42,38%

- **Definition of specificity** : proportion of non events correctly predicted, in this case 1897/2186=86,78%.

- The first goal is for both of these proportions to be high. In our hypothetical example with no predictor, the specificity would be 1 but the sensitivity would be 0, a non acceptable result !

Quality of prediction and ROC curve (III)

- If you use the CTABLE option without the PPROB option, you get classification tables for a wide range of possible cutpoint values, where the cutpoint probabilities (in the first column) range from 0 to .98, incrementing by .02 for each table.
- The column labeled "Correct" is the overall percentage of predictions that were correct. Interestingly, this reaches its highest value of 69,6 percent for cutpoint of 0,48.
- The next two columns report the sensitivity and specificity for each cutpoint, reported as percentages.
- There is an inverse relationship between specificity and sensitivity: the closer the cutpoint is to zero, the higher probability that an observation will be predicted to be an event. So most events will be predicted to be events, and sensitivity will be high.

Quality of prediction and ROC curve (IV)

- However, most non-events will also be predicted to be events, and specificity will be low.
- When the cutpoint is high, you get the opposite pattern with a low sensitivity but high specificity.
- "What's the best cutpoint value?" Some authors recommend choosing a cutpoint that produces approximately equal values of sensitivity and specificity. For our model, that result occurs with a cutpoint of 0,36-0,38.
- But, ideally, the choice would depend on your assessment of the relative costs of the two kinds of errors. Failing to detect a disease might be regarded as a substantially more costly error than diagnosing someone with the disease who doesn't really have it.
- Here, a cut point of 0.3 would give a sensitivity almost equal to 0.8, that is a correct prediction of the event in 4 cases out of 5.

ROC Curves

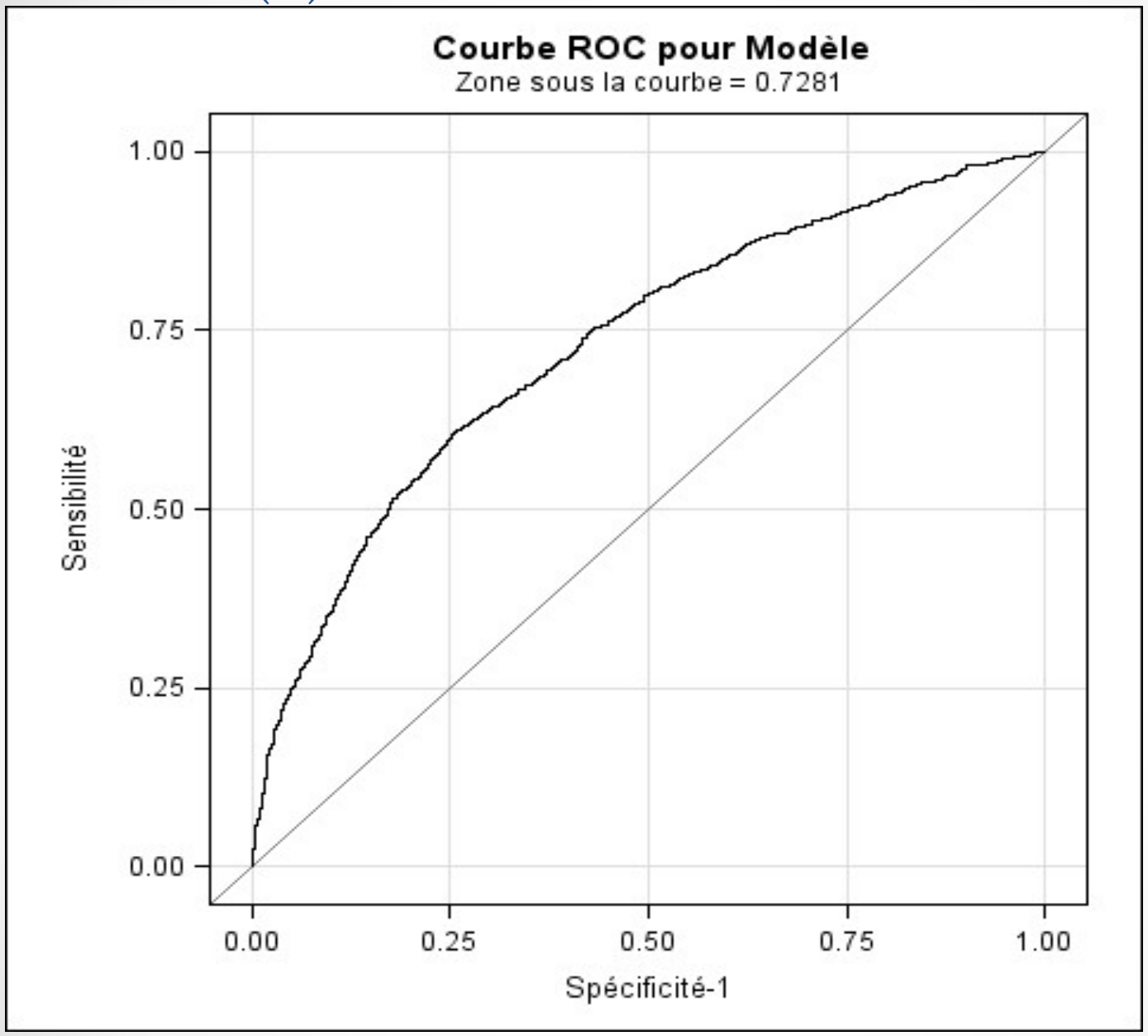
- The ROC curve gives us a way of graphically summarizing the information in the output and also provides the basis for calculating a single statistic that assesses the predictive power of the model and does not depend on the cutpoint value.

- The ROC curve is simply a graph with sensitivity on the vertical axis and 1 minus specificity on the horizontal axis, both of which increase as the cutpoint decreases from 1 to 0.

- PROC LOGISTIC DATA=paterscore PLOTS(ONLY)=ROC;

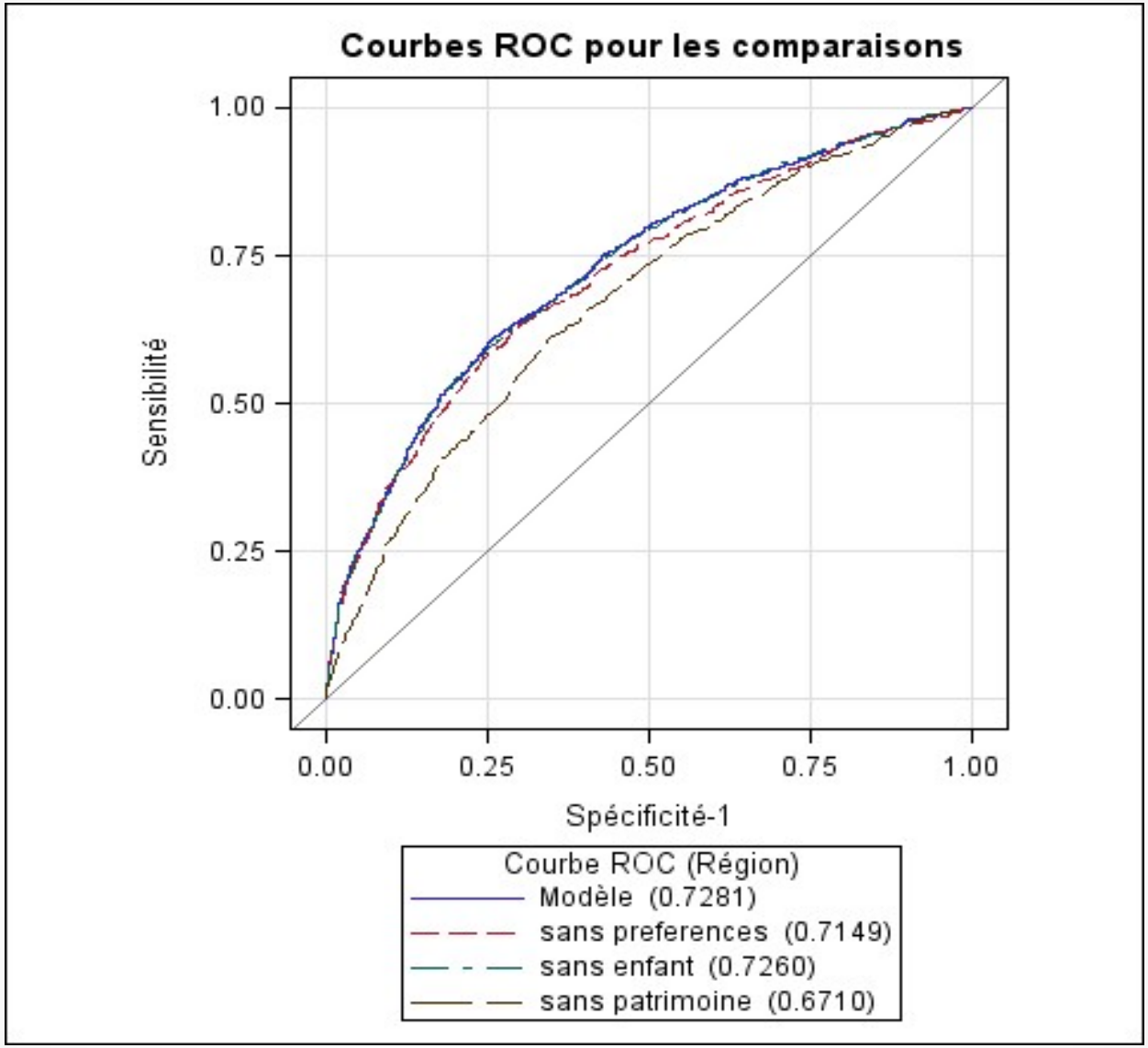
- The 45-degree line represents the expected ROC curve for a model with an intercept only, that is, one with no predictive power. The more the curve departs from the 45-degree line, the greater the predictive power. The standard statistic for summarizing that departure is the area under the curve, which here is reported as 0.7281.
- this is the same as the c statistic that is reported in the "Association of Predicted Probabilities and Observed Responses" table

ROC Curves (II)



ROC Curves (III)

- Another attractive feature is the ability to compare the ROC curves and c-statistics for different models. In the Sas Program, I request ROC curves and c-statistics for the basic model and for three different submodels, each omitting one of the model predictors.
- The four ROC statements produce separate ROC graphs for each model (not shown) and a final graph that displays all the curves in one graph.
- Besides the graphs, we also get the table which reports various ordinal measures of association between observed and predicted values for the three models specified in ROC statements, as well as for the original model.
- The ROCCONTRAST statement enables us to test for differences in the area under the curve (c-statistic) for the different models. We first get an overall test of the null hypothesis that the c-statistic is the same across all four models (the main model with all three covariates and the three submodels that omit one covariate each).
- Including the ESTIMATE=ALLPAIRS option, we get the difference in the c-statistics for each pair of models, along with its standard error, 95 percent confidence interval, and p-value for testing the null hypothesis that the true difference is 0.


Courbes ROC pour les comparaisons
Sensibilité
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Spécificité-1
Courbe ROC (Région)
Modèle (0.7281)
sans preferences (0.7149)
sans enfant (0.7260)
sans patrimoine (0.6710)

Predicted Values and Residuals

- L'instruction OUTPUT permet d'engendrer un fichier SAS contenant un certain nombre de statistiques importantes sur les valeurs prédites et les résidus:
- The OUTPUT statement can produce a large number of case-wise statistics, for each individual observation Here are some of the statistics that can be selected:
- **Linear prediction** : Predicted log-odds for each case. In matrix notation, this is $x\beta$, so it's commonly referred to as XBETA.
- **Standard error of linear predictor**, Used in generating confidence intervals, STDXBETA.
- **Predicted values** : Predicted probability of the event, based on the estimated model and values of the explanatory variables, PREDICTED.

- **Deviance residuals**, contribution of each observation to the deviance chisquare, RESDEV
- **Pearson residuals**, contribution of each observation to the Pearson chisquare, RESCHI

- since the dependent variable can only take on values of 0 or 1, the utility of residuals is somewhat limited. A high residual would mean that the individual had the event, even though the predicted probability of the event was low. But we know that low probability events do happen sometimes, so this doesn't necessarily mean a failure of the model, or that something is amiss with that individual.

Obs.	xbeta	stdxbeta	predicted	reschi	resdev	ass_vie
1	1.16613	0.13857	0.76244	0.55819	0.73651	1
2	-0.73729	0.10862	0.32360	-0.69167	-0.88427	0
3	-0.65123	0.11517	0.34271	-0.72208	-0.91611	0
4	1.14085	0.13126	0.75784	0.56528	0.74470	1
5	**-0.94061**	**0.09085**	**0.28078**	**1.60048**	**1.59386**	**1**
6	-0.38079	0.11793	0.40594	-0.82663	-1.02056	0
7	-1.16951	0.14818	0.23694	-0.55724	-0.73542	0
8	-0.94123	0.10488	0.28065	-0.62462	-0.81168	0
9	-0.07243	0.09747	0.48190	-0.96443	-1.14681	0
10	-0.24632	0.07368	0.43873	-0.88412	-1.07476	0
11	-0.96771	0.13262	0.27534	-0.61640	-0.80256	0
12	**-0.76594**	**0.11828**	**0.31736**	**1.46664**	**1.51508**	**1**
13	-1.07915	0.11822	0.25367	-0.58300	-0.76496	0
14	**-0.79927**	**0.14491**	**0.31018**	**1.49128**	**1.53010**	**1**
15	0.15862	0.16829	0.53957	0.92375	1.11084	1
16	-1.50836	0.13247	0.18118	-0.47040	-0.63229	0
17	-0.74241	0.09415	0.32248	-0.68990	-0.88240	0
18	**0.78246**	**0.14010**	**0.68621**	**-1.47880**	**-1.52252**	**0**
19	0.03515	0.09672	0.50879	-1.01773	-1.19237	0
20	-1.07095	0.08883	0.25522	-0.58539	-0.76768	0
21	-1.52312	0.10153	0.17900	-0.46694	-0.62807	0
22	-0.08636	0.11490	0.47842	-0.95774	-1.14096	0
23	**1.37116**	**0.11252**	**0.79757**	**-1.98492**	**-1.78737**	**0**
24	-0.83228	0.10704	0.30316	-0.65959	-0.84995	0
25	-1.60493	0.13357	0.16729	-0.44822	-0.60510	0

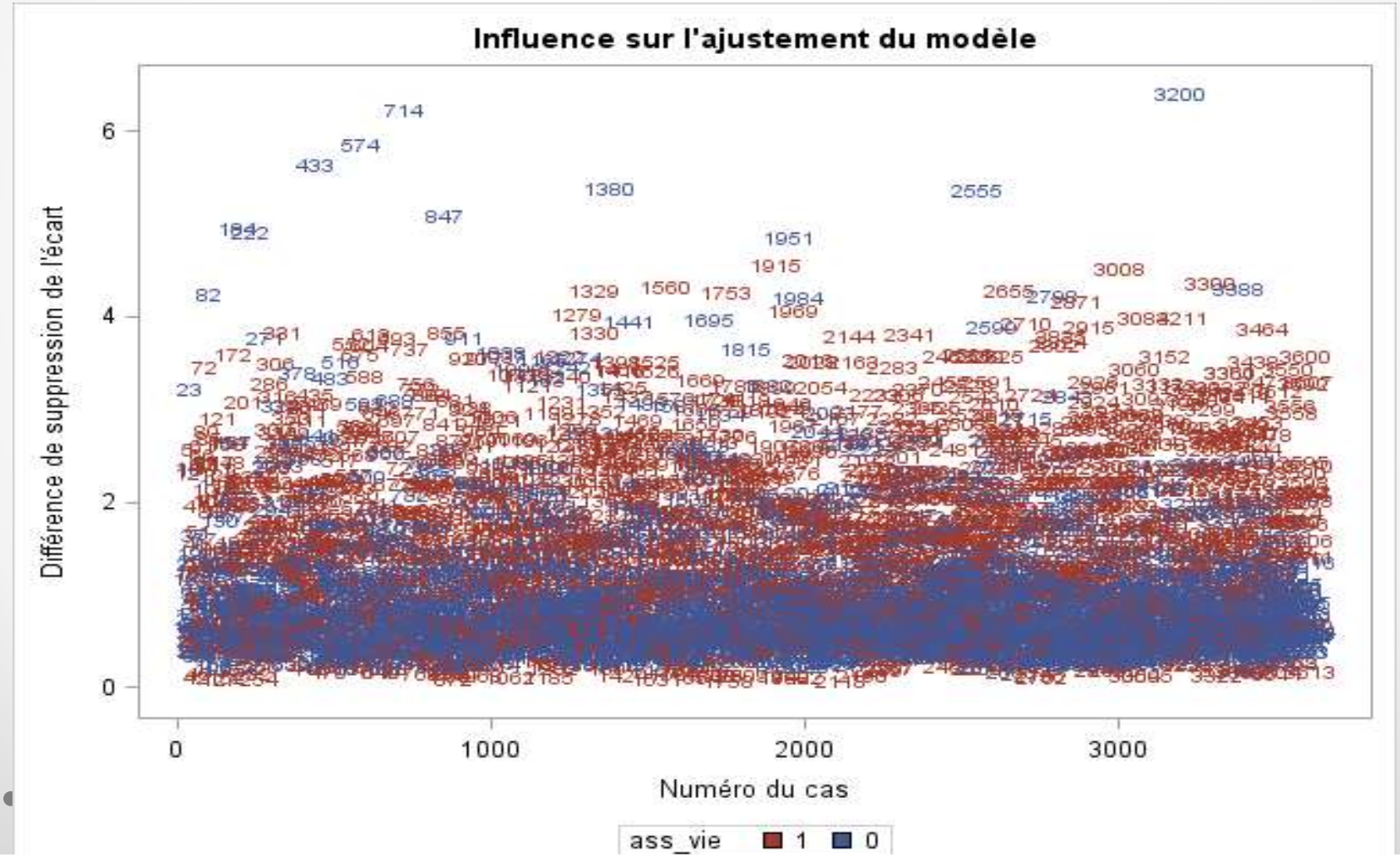
Influence of individual observation

- The OUTPUT statement can also produce several statistics that are designed to measure the influence of each observation. Basically, influence statistics tell you how much some feature of the model changes when a particular observation is deleted from the data set.
- **DFBETAS_**: These statistics tell you how much each regression coefficient changes when a particular observation is deleted. The actual change is divided by the standard error of the coefficient.
- **DIFDEV**: Change in deviance with deletion of the observation.
- **DIFCHISQ**: Change in Pearson chi-square with deletion of the observation.
- **C et CBAR**: Measures of overall change in regression coefficients, analogous to Cook's distance in linear regression.
- **LEVERAGE**: Measures how extreme the observation is in the space of the explanatory variables. The leverage is the diagonal of the "hat" matrix. This matrix is defined as the projection matrix of the explained variable y in the plan defined by $X\beta$, that is giving the predicted value of y :

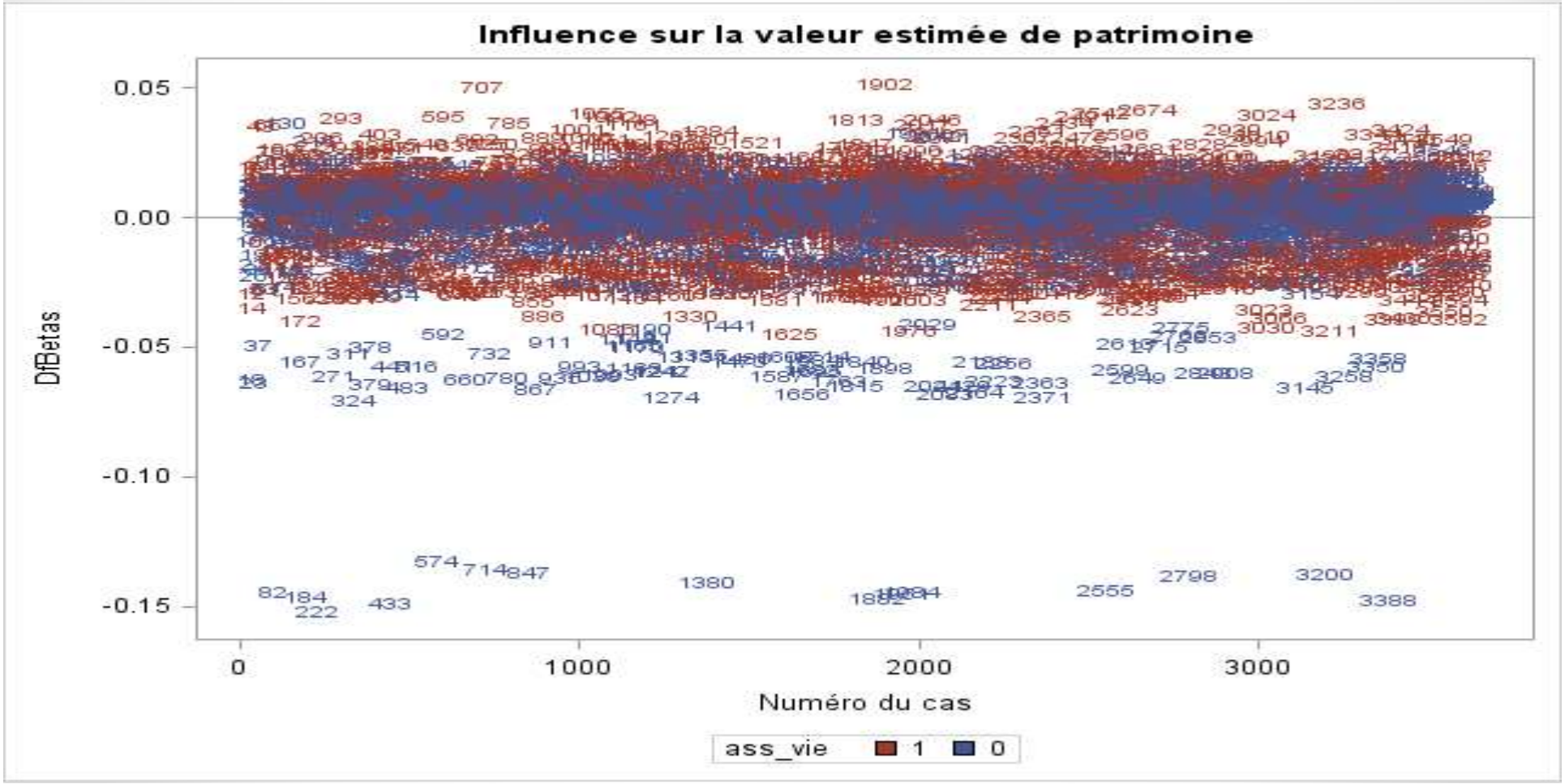
$$\hat{y} = Hy$$

- The diagonal term of this « hat » matrix evaluated the effect of y_i on the predicted value. The higher that value, the more the observation i takes high values in the space of the explained variables.

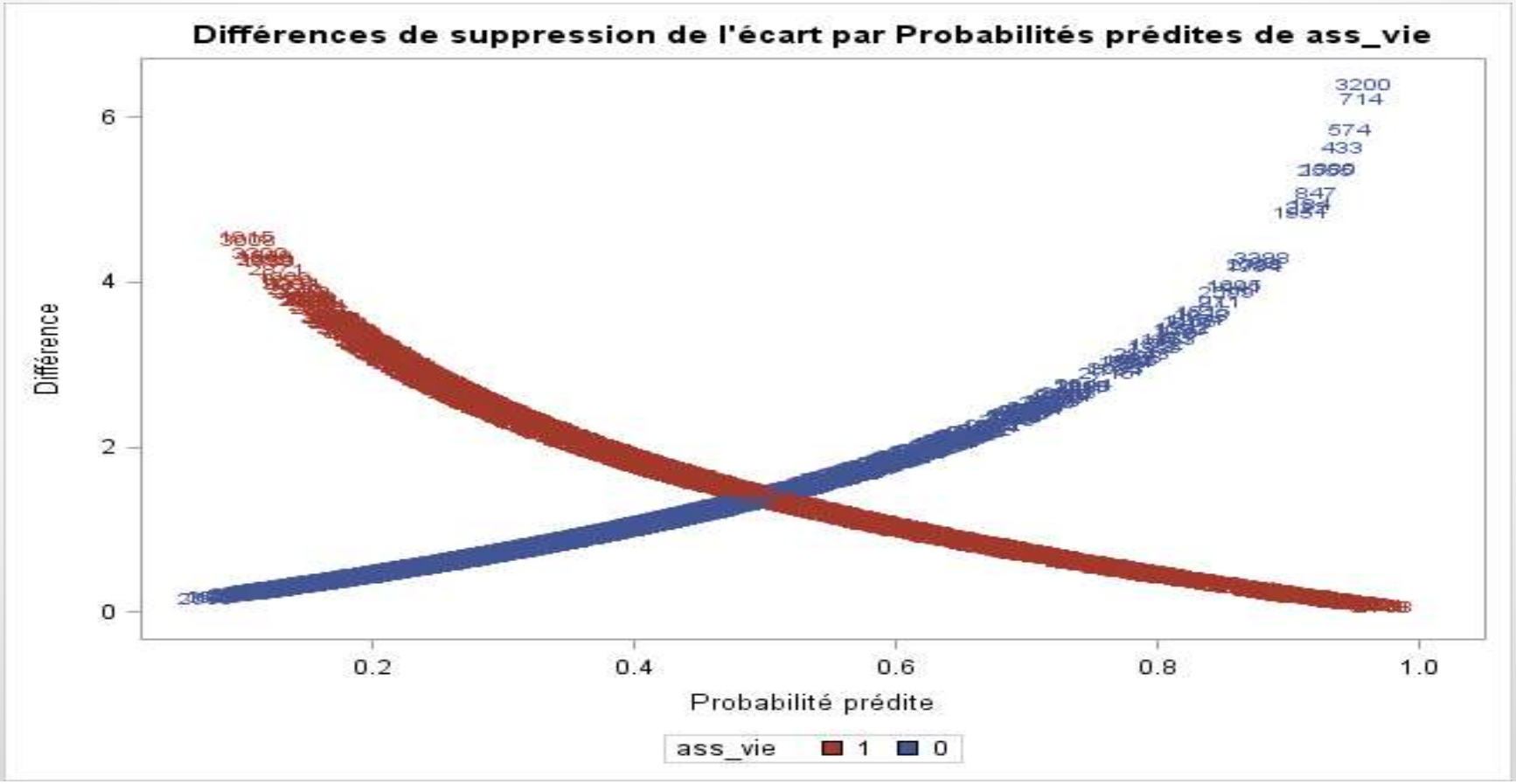
The graph below is one of the seven generated by the INFLUENCE option and measures the change in deviance following the removal of an observation. We observe that observations 433, 714, 847, 3200 generate a large change in deviance, because they are poorly predicted by the model, which predicts that they have a life insurance policy when they do not.



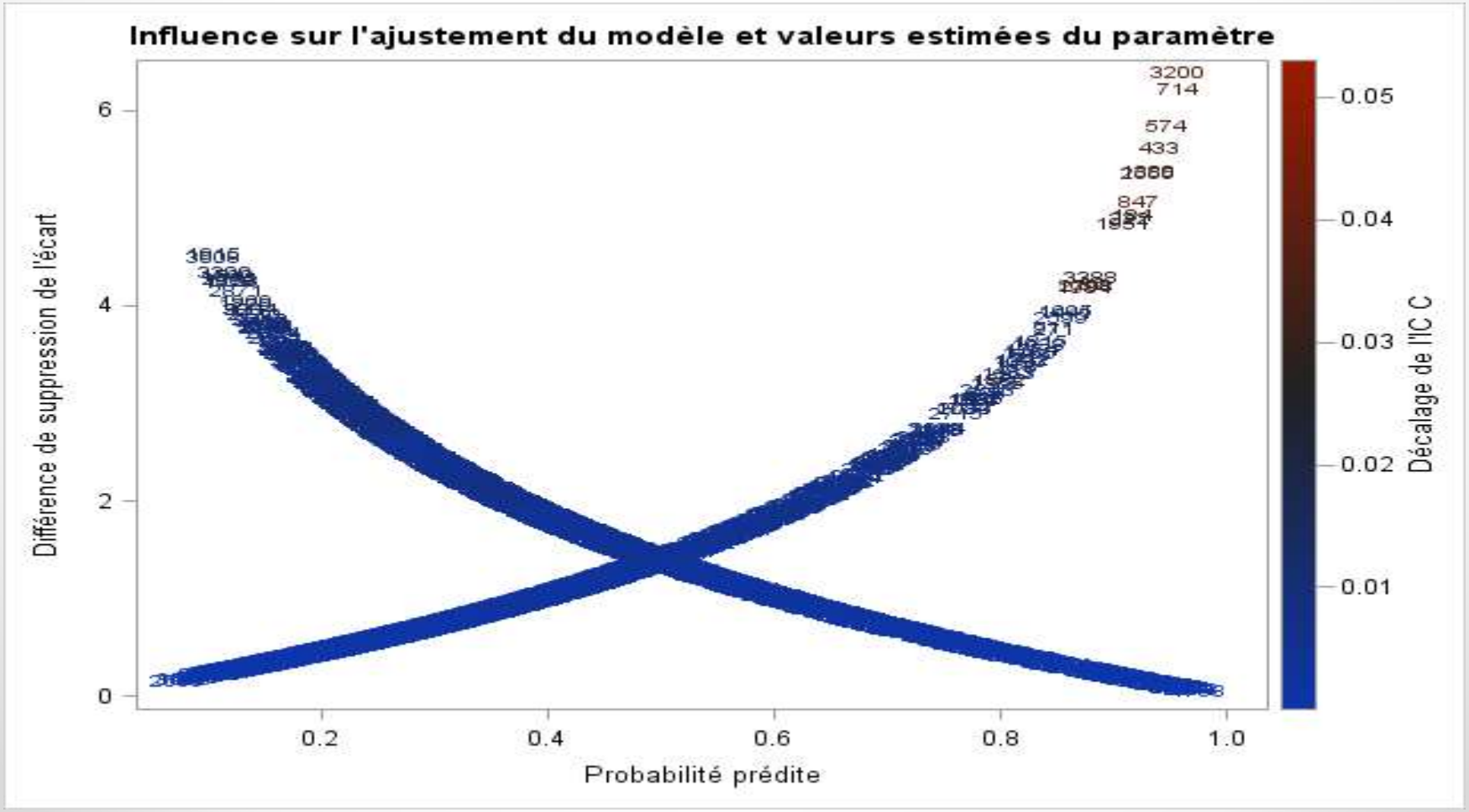
The graph below is produced by the DFBETAS option and measures on the vertical axis the change in the parameter associated with the wealth variable. In the case of our model, observations 714 and 222 when excluded lead to a decrease in the coefficient because the observations correspond to high wealth values and an absence of life insurance contracts.



The graph below is one of the PHAT graphs representing the change in deviance in relation to the probability predicted by the model. The curve increasing towards the right corresponds to observations with no event, the curve decreasing towards the left to observations with event. The principle is to focus on observations with high values of change in deviance and which are well separated from the others: 3200, 714 etc...Translated with www.DeepL.com/Translator (free version)



The graph below is a version of the graph above, produced by the DPC option. It has the originality that the observations are colored according to the value of another statistic, the change statistic in the confidence interval C.



This graph relates the observations leading to a significant change in deviance on the ordinate and extreme values of the explanatory variables on the abscissa.