Characteristics of the demand for private long-term care insurance in France : a step by step estimation algorithm in an analytical CRM context [1]

Sébastien Nouet[*], Maximilien Nayaradou[**]
Manuel Plisson[***]

[*]Pôle universitaire Léonard de Vinci, Centre d'Études et Recherche en Management de l'EMLV
Membre de l'Institut des Actuaires
sebastien.nouet@devinci.fr

[**]Université Paris 9 Dauphine, Laboratoire d'Economie de Dauphine
maximilien.nayaradou@dauphine.fr

[***]Université Paris 9 Dauphine, Laboratoire d'Economie de Dauphine
manuel.plisson@dauphine.fr

Abstract. Private long-term care insurance is underdeveloped in European countries. To understand this actual market failure, we study the long term care insurance demand characteristics using banking data and estimate the probability of subscription. Thanks to a logit model and an original step by step estimation algorithm, we show that belonging to upper classes reduces this probability; that women belonging to farmer, worker or employee classes, and having some asset, have 5 times more chances to subscribe (same result for the oldest individuals of this category). Thus insurers are in a favourable position to develop their long-term care portfolio: population ageing effect will make currently defined targets more representatives during the next twenty years. This paper, the sole empirical one using French data, provides insurers with a key decision-making econometric tool for calculating the probability to subscribe, that can be applied to their own portfolio.

1 Introduction

Recent studies show that the incidence of long-term care (LTC) increases strongly with age (*cf.* Gisserot and Grass (2007)). In France, the probability that an individual who has reached the age of 65 will use LTC services before his or her death is 40%(OECD (2005)). The ageing of the population is therefore likely to increase the demand for LTC even though this increase will be tempered by the increase in life expectancy without disability (Gisserot and Grass (2007), Duée and Rebillard (2004)). For individuals cared for at home, this amounts to a cost which fluctuates between €340 a month for light LTC and €5,300 a month in the event of maximum physical and psychological dependence (Ennuyer (2006)). On average it amounts to €1,500 a month (Loones et al. (2005)). Quite apart from the loss of well-being caused by the unexpected onset of this state, LTC represents a highly significant financial risk burden by elderly people. Should there be an increase in the number of people with LTC needs as well as an increase in the average duration of LTC, this would raise a number of concerns, particularly concerning the long-term balance of public finances. In France, the social benefits for LTC are called APA (Personal Independence Allowance) and are paid by the French Departmental General Councils.

[1]The authors gratefully acknowledge, Risk Foundation supports: Chaire Dauphine-ENSAE-Groupama « Les particuliers face au risque » and Chaire « Transition démographique-transition économique », ANR Risk, CAPA Conseil, C. Hess, P. Cazes, H. Haddak of Paris-Dauphine.

The amount depends on the state of dependency measured by the ranking GIR and the level of income. The eligibility rules are quite different from Medicaid system. If we compare the average cost with the average social benefit which amounts to 409€, we note that, on average, the government contribution represents only 30% of the average cost [3]. Other estimates have put the government contribution at 50% of the average cost (Cour des comptes (2005)). On average, there therefore remains a shortfall of €1,100 to be paid by the person receiving long-term care. This average shortfall is higher than the total average pension for women which amounts to €979. LTC therefore represents a highly significant financial risk, especially as in this scenario we can merely contemplate situation of the average person. Yet most of the expenditure risk is uninsured.

Despite this significant financial risk, the LTC insurance market remains small. The coverage rate of this risk in France and in the United States of America, the two biggest markets in the world, is below 10% while the coverage rate for supplementary health insurance in France is 86% (Haut Conseil pour l'Avenir de l'Assurance Maladie (2005)). Several explanations have been put forward to account for this "long-term care insurance puzzle". Due to incomplete markets, insurers only offer an annuity, which may discourage people from buying insurance (Cutler (1993)). Furthermore, while there has been no apparent confirmation of moral hazard in the American market, adverse selection cannot be ruled out (Sloan and Norton (1997)). Indeed, in the American market, high-risk people take out more insurance than low-risk people. This seems to be offset by the fact that the people with the highest risk aversion take out most LTC insurance. These people have also the particularity to invest most heavily in prevention, which reduces the likelihood that they will need long-term care (Finkelstein and McGarry (2006)).

However, as Brown and Finkelstein (2007) note, supply side market failures are unsatisfactory and we should also ask why demand for long-term care insurance is so low. Limited consumer rationality or misconceptions about the extent of public insurance seems less and less relevant in France (CSA (2006)). It has been shown that the crowding out effect is weak when the public insurance do not take into account LTC insurance benefits, which is the case in France(Brown and Finkelstein (2008)) is possible, however, that the demand for LTC insurance has suffered from an intergenerational moral hazard (Zweifel and Struve (1996)). Due to this theory, children would be less careful in the situation their parents bought a LTC contract. To elicit children to care for them in case of LTC, parents would not buy LTC insurance.

Generally speaking, there are a number of theoretical works that attempt to explain why so few individuals buy this product, but very few empirical validations. The main empirical studies have been carried out on the American (Kumar et al. (1995), McCall et al. (1998), Brown and Finkelstein (2007)) and Spanish (Costa-Font and Rivera-Forns (2008))markets. In order to better understand the LTC insurance puzzle, we have first to understand who is buying it (analytical CRM).This is what we set out to do as part of this research. To our knowledge we

3. The AGGIR procedure is a multidimensional tool which measures the level of autonomy. It defines 6 levels of dependency called GIR, which requires a certain level of care for the activities of daily living. The GIR 1 concerns the heaviest dependency.

The French social benefits are based on this ranking.

are the first research to deal with banking data in the French market. The other study that we know was dealing with survey data issued from SHARE basis (Courbage and Roudaut (2007)). Using this banking data, we will therefore try to interpret these results with regards to other estimates carried out in other countries but also with regards to the theoretical literature.

2 Data

2.1 Population

As is often the case when assessing an insurance market, we use company data (Cadoux and Loizau (2004), Hsin Lin and Hern Chang (2008), Trigo-Gamarra (2008)). Company data is often used when assessing insurance markets, particularly that of classical non life insurance (IARD) (Mace (2003), Vasechko et al. (2009)), and when assessing long-term care insurance (Mitchell et al. (2008)).

The data used in our study comes from a large bank in France which proposes to its clients LTC insurance. This bank offering insurance services represents approximately 20% of the LTC insurance market in France (Decoster (2006)). This company has branches evenly distributed across France and caters to all types of customers. With a 20% market share, it is safe to say that it is relatively representative (without being absolutely representative) of the French long-term care insurance market. We note that most of the non life insurance (IARD) studies using company data do not include tests of representativeness, in particular none of the studies mentioned above. Indeed, company data are not often systematically representative of the overall population since insurance company clients are often oriented from a sociological point of view.

From 2002 to 2005, this bank offered all its customers in a position to take out a policy (customers aged between 18 and 75 years) an individual LTC insurance. Therefore all customers were subject to the same commercial policy. It is important to recall that it is crucial and obligatory in this study to reason a customer portfolio from a single insurance company, since the scoring (logistic modeling is a specific case of scoring) is only applicable if all of the customers have similar business procedures (Lebart (1971)). Applying scoring to national data, or which are national in character, would not be useful for this study, except if there was only one insurance company covering the entire long term care national market with the same marketing policy, which is not the case here. For that reason it could be misleading and could introduce serious biais, applying scoring to national data.

We had got the data from “Centre” Region (region with 6 districts: Cher, Eure, Eure-et-Loire, Indre, Indre-et-Loire et Loiret). Of the 275,257 people insurable in the portfolio of this region, 5,027 took out the contract (1.82% of the customer portfolio). We had access to a representative sample of 37.45% of the portfolio of uninsured people (a random sample of 101, 205 from the insurance database of uninsured customers of the bank offering insurance services) and of all the 5,027 insured customers. Descriptive statistics of this data are set out in tables 1 and 2.

The main advantage of the data to which we had access is that it is not data based on reported preferences, but on revealed preferences. Unlike surveys by questionnaire sent to homes, we do not need to calculate the rate of participation or the rate of return since all customers were contacted. There is therefore no bias in our study caused, for example, by the over-participation of certain socio-economic categories (particularly the most educated) to surveys by questionnaire. The trade-off for the exhaustive nature of this data is that we have fewer variables than in declarative surveys. However as it involves a bank offering insurance services, we have access to banking data (income and financial assets), which is rarely the case with insurance companies or mutual benefit societies.

2.2 Insurance's contracts characteristics

The LTC contract offered is not really an insurance contract but an annuity contract. The person can take out this contract up to the age of 75 and for an annuity sum defined upon subscription. Over the considered period the minimum annuity was €600 a month. The monthly premium paid by the insured person depends on the age at which he or she subscribed and the total benefits he or she wishes to receive in the event of long-term care. When his or her level of LTC is certified by the regional medical unit linked to the bank offering insurance services, the insured person ceases to pay his or her premiums and receives a monthly annuity allowing him or her to finance their care. This benefit doesn't depend on the care expenditure. It depends on the premium paid by the policyholder. In the present case, the contract covers heavy LTC needs corresponding to GIR 1 and 2. The rates do not take into account gender, despite the fact that on average women have a higher likelihood of needing long-term care and of remaining in this condition for longer than men.

3 Method

3.1 Econometric Analysis

The logistic model and the bootstrap method. Using a logistic model, we are going to estimate the impact of the explanatory variables (described in paragraph 3.2) chosen on the probability of taking out a long-term care insurance policy (analytical CRM). For an individual i, the likelihood P_i, that he or she insures against LTC depends on an explanatory variable vector, Z_i. We therefore see that: $P_i = proba[Ass_i = 1] = F(Z_i, \beta)$
Ass_i is a binary variable that takes the value 1 if the individual takes out an insurance policy and 0 if not. The vector β reflects the positive or negative marginal effect of changes in Z_i on the likelihood P_i and $F(.)$ designates the link function associated with our logistical modelling. The estimation of coefficients vector β is obtained by the maximum likelihood method.
Put more precisely, our logistic model can be presented as follows: The probability is modeled on the event you are looking for $Y = Ass_i = 1$ that represents the act of taking out dependency insurance. This probability will be estimated for every individual i with a particular profile characterized by our k-tuple$(X_1, X_2, \ldots, X_k)$, with $(X = Z_i)$ for an k-tuple relating to this individual. These individual probabilities will be estimated by proportions. Thus, for example, individuals with the modality x_0 for the variable which could represent income will be divided into policy holders, with the code 1, and non-policy holders, with the code 0. We can therefore

work out the probability of taking out insurance for individuals with incomes x_0 by calculating the ratio between the policy holders with this attribute x_0 and the total number of individuals. Within the framework of our model, insofar as several variables are processed and some are continuous, we apply the model on an k-tuple, not using simple ratios of policy holders with a certain profile in relation to the total population, but using a likelihood function (Thomas (2000)). Thus the model can be presented in the following way with the equation 1 below, using the logit function

$$Logit\left[P\left(Y=1\diagup\mathbf{X}=(x_1,x_2,\ldots,x_k)\right]=\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_kx_k. \qquad (1)$$

We have access to all 5027 policyholders but we only have 37.45% (in other words 101,205 individuals) of uninsured persons. First we will estimate the model by weighting uninsured individuals. In the simple model and in the model with interactions, we have allocated the weighting of 1 to each of our policyholders and the weighting of 2.67 (= 1/ 37.45%) to each of our uninsured individuals. However, essential validation work based on the bootstrap technique remains to be carried out to confirm the previous results. Indeed our segmentation is certainly real and representative of the general population but unbalanced from a statistical point of view. In other words, the very low proportion of policyholders in our portfolio might distort the estimations. We need to show by using the bootstrap method that the factors shown in these two models are genuinely significant. To do this, we will build 20 sub-samples from our portfolio, which will lead us to estimate 20 non - weighted simple models and 20 non - weighted models with interactions. Each of these sub-samples is made up of the representative sample of non - policyholders (37.45% of the portfolio of uninsured people) to which we add a random selection of 37.5% of all available insured individuals. We estimate the bootstrap model with main effects only and the bootstrap model with interactions on these sub-samples. We look at the associated coefficient of variation for each significative variable. The standard deviation of each coefficient (with respect to average) is relatively small or acceptable which confirms our results. The coefficient values are close to those of the weighted model (main effects and interactions), that is, the model in which uninsured individuals have been weighted. Tables 8 and 9 in the annexe contain the model estimations with bootstrap method, and the coefficients of variation. The models were validated by two validation tests: the " rank correlation coefficients" and the " Hosmer - Lemeshow test".

Backward estimation algorithm applied to the weighted model and the bootstrap model. The bootstrap method can be used to check the results obtained with the weighted model. It involves only 20 samples in our case owing to the many manipulations required and time involved to estimate the model, the whole cannot be easily automated.

Estimation algorithm used with the bootstrap model and also with the weighted model:
The logistic econometric model will be estimated step-by-step following the steps for eliminating non-significant coefficients, as well as those for fitting the model to our database. The estimation steps outlined below should be executed in a specific order within a process loop and then stopped at a specific instant. They must then be validated, which is done using two validation tests: the rank correlation coefficient and the Hosmer-Lemeshow test.
It is important to observe each of these criteria and their order of application because removing certain coefficients deemed insignificant compared to certain criteria at the wrong time in the

procedure may deteriorate the final outcome model and the quality of its predictions. Following the step by step estimation algorithm I recommend:

- Step 0) User must look at the overall significance test of model I: with all k variables. The criterion used is equivalent to the Fisher test used in traditional regression. It determines whether the model makes sense, or if the whole of k variables are all statistically null. The general test of the significance of the model with a 5% level confidence consists of testing, with a χ^2_k statistic, the statistical validity, of hypothesis $H_0 : \quad \beta_1 = \beta_2 = \ldots = \beta_k = 0$.

- Step 1) Start by estimating the model with all of its variables using the backward procedure. This procedure of eliminating statistically non-significant coefficients involves eliminating the least significant variable in the model and then automatically repeating the estimation following the same steps. The backward procedure will stop when this type of step-by-step elimination is no longer possible, that is, when each the coefficients estimated as β_i is significant. The significance of coefficient β_i is determined with the corresponding p-value. The user sets the maximum level at which the backward procedure can determine the significance of each β_i : 5%. The p-value must be below this level.
The likelihood ratio test is most widely used, it is applied here to underscore the presence of a variable in a logistic regression model. The test for the presence of an explanatory variable X_i , consists of testing the statistical validity of hypothesis $H_0 : \quad \beta_i = 0$. The statistic G calculated for testing null hypothesis H_0 is presented in the form

$$G = -2\log\left(\frac{\text{likelihood without variable}}{\text{likelihood with variable}}\right) = -2\log\left(\frac{L(\beta_i^*)}{L(\beta)}\right),$$

where β_i^* , indicates the vector β without its i^{th} component. Under null hypothesis H_0, the statistic G follows a chi-square χ^2_1 distribution (with 1 degree of freedom). The variable X_i is significant if the model without β_i is rejected, that is to say null hypothesis H_0 is rejected, which is equivalent to $P(\chi^2_1 < G) < 5\%$.

- Step 2) Once these two steps have been followed, the model's goodness-of-fit test criterion, or the residuals analysis criterion, must be considered. This criterion is equivalent to R-squared for traditional regression. It is one of the means of fitting our model to the data. It consists of measuring the importance of the estimated residuals which characterize the actual difference between the two sides of equality (1) defined by the model. This difference is calculated in relation to all the individuals in the database.
The model is considered in the form :
$y_i = \pi_i(\mathbf{X}) + \varepsilon_i$ with $\pi_i(\mathbf{X}) = P(y_i = 1|\mathbf{X})$.
Thus $E(\varepsilon_i|\mathbf{X}) = 0$ and $Var(\varepsilon_i|\mathbf{X}) = \pi_i(\mathbf{X})(1 - \pi_i(\mathbf{X}))$.

Denote $\widehat{\pi}_i(\mathbf{X})$ the estimated *a posteriori* probability. The normalized Pearson's residuals are given by

$$r_i = \frac{y_i - \widehat{\pi}_i(\mathbf{X})}{\sqrt{\widehat{\pi}_i(\mathbf{X})\,(1 - \widehat{\pi}_i(\mathbf{X})).}}$$

If the model is adjusted correctly, the statistic $\chi^2 = \sum_{i=1}^{N} r_i^2$ approximately follows a chi-square distribution with $N - k - 1$ degrees of freedom, and the corresponding p-value must be higher than 5%.

- Step 3) Once the residuals criterion has been satisfied, we now address the problem of eliminating variables diagnosed by the backward procedure as significant but usually exceptionally not significant with respect to the "odd ratio test" according to the value 1 is one of the extremities of confidence interval (C.I.) relatively to the "odds - ratio criterion" (we look at the output model after the backward procedure). Value 1 cannot <u>strictly</u> belong to C.I. since the variable is significant according backward procedure. In this case with an extremity of C.I. equal to 1, variable involved only slightly affect the probability of belonging to category 0 or 1 for individuals in the database when one alters the values these variables can assume, even though they are significant according to the backward procedure. This problem (<u>1 = one of the C.I. extremities</u>) is usually exceptionally, excepted (*) when the segmentation of the data is unbalanced from a statistical viewpoint, which is the case in our study since around 98% of individuals are coded 0. Thus in our study we need to use elimination criterion referred to as the odds ratio, after the backward procedure, which is not usual. The variable deemed least significant with respect to the "odd - ratio test", (<u>1 = one of the C.I. extremities</u>) is definitely eliminated manually, and the backward procedure is repeated at the step 4).
It should be noted that two variables in the same model can be diagnosed non-significant with respect to the "odd - ratio test" (<u>1 = one of the C.I. extremities</u>). The one whose odds ratio is closest to 1 should be eliminated at this step. (*) It's not exceptional to not obtain the strict equivalence between tests of steps 1) and 3) since: the regression is not linear, these tests are asymptotic, and the data is very unbalanced. Thus experimenter will use both tests.

- Step 4) The backward procedure is repeated using the initial model with all "k" variables (Model I) at the first loop of algorithm stripped of the variable definitely eliminated thanks to the previous odds ratio criteria applied in step 3). In other words backward procedure is repeated using the model II (= Model I stripped of 1 variable) or the model with $k-1$ variables. It's important to underline that the backward procedure is not repeated using the output model provided by the first backward procedure in the algorithm, from which this same variable was removed. The backward procedure is therefore a preliminary step used to "odds-ratio" variable elimination within our model according to the segmentation is unbalanced from a statistical viewpoint. Each loop allows us to definitely eliminate 1 variable.
Then now we obtain a second model output after applying backward procedure to model II, the output obtained is named: model II backward output.
It is now necessary to go back to Step 2) with this "model II backward output", and repeat the same steps than previous, and so forth.
Step 4) stops at the very beginning when it is no longer possible to eliminate variables using the "odd - ratio test" (here 1 = one of the C.I. extremities).
We note that at the second loop in this step 4), we will obtain a new model II or model III stripped of a second new variable eliminated with the same criterions which corresponds to the initial model with all variables stripped of 2 variables (first variable eliminated at the first loop thanks to step 3) and a second eliminated at the second loop thanks to the step 3)). Then the new model II or model III at the second loop has $k-2$ variables, after we apply backward.

- Step 5) It is no longer possible to eliminate new variables. It is now necessary to look at the 2 final validation criteria: the “model goodness-of-fit final test” using the “rank correlation coefficients” and the “back-testing test” or “Hosmer-Lemeshow test”. This involves testing the resulting model on the database. The second criterion is the most important. It involves breaking down the total population into 10 levels of insurance probability.

Applying all of these steps (3 of which are applied in a loop) is extremely important since failure to validate with the model’s goodness-of-fit test criterion step2) implies going back on the local elimination choice of the last variable while still retaining an optimal choice type.

3.2 Explanatory variables

The explanatory variables are defined below.

1. Age :
   Age is a continuous variable, between 18 and 75 years.

2. Gender :
   Gender is a dummy variable, encoded 1 a female and 0 a male . We should note that in the banking portfolio of insurable customers only 28.6% of the main account holders are women and 71.4% men. This is explained by the fact that in most joint couple accounts, it is the man who is the main account holder. Of course, of the 71.4% who are men, not all account holders are men living in a couple; there are also unmarried, widowed and divorced men. Likewise of the 28.6% of accounts whose main holder is a woman, even though it is likely that there are a relatively higher number of unmarried women, widows or divorcees, there may also be women living in a couple having a joint account and being the main account holder (although this situation is less common). Unfortunately our database does not specify the marital status of the customer or whether the account is a joint account. But we can say with certainty that in 28.6% of the accounts where the main holder is a woman there is an over-proportion of single women, whether they are unmarried, widowed or divorced.

3. Socio-economic category :
   Our database provides us with excellent results concerning the account holder’s social background. Furthermore this variable can be considered to be a good indicator of the level of education. Each of seven sub categories is a dummy variable: Farmers, Storekeepers, Senior executive, Middle manager, White collar workers, Blue collar workers, Unemployed. Each variable is encoded 1 or 0: if the client belongs to the category or not.

4. Income :
   Income is a continuous variable. We have used thousands of euros as a unit to make the odds-ratios interpretable.

5. Assets :
   Assets is a continuous variable. It should be noted that it involves the insurable customer’s financial assets and not his or her residential property. We note that financial

assets are a pertinent variable which is not often studied empirically in the literature (only in article of Sloan and Norton (1997) among the main empirical articles on long term care insurance). Crossing this pertinent variable with more traditional variables, such as age, socio-professional category, sex and income, enables obtaining a set of pertinent interaction variables. An example would be to cross financial assets with the socio-professional category of workers. In the database, we have used tens of thousands of euros as a unit to make the odds-ratios interpretable.

	Uninsured population	Policyholders
Number of Customers	101205	5027
Gender		
Men	0.714	0.4582
Women	0.286	0.5464
Socioeconomic categories		
Farmers	0.0571	0.0543
Storekeepers	0.0553	0.0304
Senior Executive	0.0930	0.0191
Middle Management	0.1459	0.0808
White Collar	0.2393	0.3726
Blue Collar	0.2549	0.3203
Unemployed	0.1545	0.1225

TAB. 1 – *Descriptive Results: gender and socio-economic status of population.*

[about *uninsured population*]

Uninsured Population	Mean	Median	Std dev	Min	Max
Age	44.1	44	16.3	18	75
Income (€)	35667	22100	84550	1	8002493
Financial Asset(€)	28386	3000	74500	-1505	4668214

[about *insured population*]

Insured Population	Mean	Median	Std dev	Min	Max
Age	53.8	54	12.6	18	75
Income (€)	35686	24903	45446	24	967242
Financial Asset(€)	52994	9907	107938	-42	1445000

TAB. 2 – *Descriptive Results: Age, Income and Assets.*

4 Main effects model results

The simple model estimated by the weighting method gives us a first series of results (see table 3). Holding financial assets increases the probability of taking out insurance while the impact of income is not significant.

4.1 The working classes insure themselves more than the middle and upper classes

Farmers, blue-collar workers and white-collar workers ensure themselves more than executives. If we take socio-economic category as an indicator of level of education, it is clear that the probability of taking out insurance decreases with the level of education. These results are contrary to the results of McCall et al. (1998), for whom the probability of taking out a long-term care insurance policy is positively linked to the level of education, and those of Kumar et al. (1995), for whom the probability of taking out a long-term care insurance policy is negatively linked to the fact of having a low level of education. The odds-ratios not presented here indicate that blue and white collar workers, respectively are 2.6 and 1.7 times more likely to take out an insurance policy than non-blue and white collar workers. Senior executives are 2.5 times less likely to subscribe.

4.2 Elderly people insure themselves more than young people

Analysis of Table 3 shows that the probability of taking out a policy increases with age. As regards age, we observe a difference between our study and that of Costa-Font and Rivera-Forns (2008). The latter authors note a greater reported preference for LTC insurance by younger people than by older people. We reveal in our study that the probability of taking out an insurance policy against LTC increases significantly with age. Our results are therefore significantly different from those of Courbage and Roudaut (2007), for whom age is negatively linked to subscription. Furthermore, age is often taken as a LTC insurance contract price indicator: the higher the age the higher the contract price. The fact that in our study subscription rates increase with age seems to suggest that price is not a fundamental variable in the decision to subscribe. It seems rather that the approaching risk (as age advances) encourages people to take out insurance. Furthermore, our results tend to agree with theoretical predictions of Meier (1998).

4.3 Women insure themselves more than men

We also note that women insure themselves more than men in accordance with the results obtained by Costa-Font and Rivera-Forns (2008). This type of behaviour can first of all be interpreted rationally. As women have a greater probability of needing long-term care and of needing it for a longer period (Duée and Rebillard (2004)), the insurance is therefore relatively less expensive for them insofar as they are given the same pricing as men. In the American market, Brown and Finkelstein (2007) observed that in most contracts, the prices offered to women were more attractive than the actuarial price. We can suppose that these women are rational. Furthermore, married men (who form part of the 71% of accounts held in the man's name, which account for the majority of the joint accounts) are over-represented among the 45.3% of insured men; and single women (who form part of the 29% of accounts held in the woman's name) are over-represented among the 54.7% of insured women, whether they are divorced, widowed or unmarried. Of course, some of these women have children, but these population categories also include the highest number of women without a partner, without dependent children or without children at all. So the less chance they have of being looked after by a loved one, the more insurance they take out. These results are in line with those of Kumar et al. (1995). In their study the probability of being insured against LTC is negatively linked to the fact of being married. There is therefore a trade-off between LTC insurance and the opportunity for children to provide LTC. Here again we notice a positive relationship between living alone and taking out long-term care insurance.

4.4 Wealthiest people insure themselves more than the whole of people

This property (true in average) is both interesting and more difficult to understand than the three previous properties. It is described precisely in the 2 following paragraphs.

Main effects Model		
ExplanatoryVariables	Coefficients(SE)	p-value
Intercept	-6.6150 (0.0944)	<0.001
Financial Asset	0.0138 (0.00157)	<0.001
Age	0.0338 (0.00151)	<0.001
Gender	1.2942 (0.0514)	<0.001
Senior Executive	-0.9126 (0.1744)	<0.001
White Collar Workers	0.5175 (0.0592)	<0.001
Blue Collar Workers	0.9707 (0.0630)	<0.001
N	103090	
Wald	1557.6130	<0.001
Likelihood Ratio	1669.5836	<0.001

TAB. 3 – *Model with main effects only.*

5 Models with interactions terms results : the specific factors

The model with interactions, estimated by the weighting method allows our results to be refined (see table 4).
The people that take out the most insurance: women, older and better-off blue-collar workers. Within the "farmer" and "blue-collar worker" populations as well as for women, as assets increase, so too does the probability of subscribing. Among "blue-collar workers" and "white-collar workers", increasing age offsets the increasing probability of subscribing, which remains, nevertheless, a positive factor. Among "women-blue-collar workers", "income" slightly favours subscription, while among "men" in the same socio-economic category the opposite is true: there is a slightly negative influence. The "age" variable specific to "women" increases the probability that older members of this category will subscribe. This model shows too that the assets variable specific to "women" increases the probability among women in general, while the same variable doubly specific to "women" and to "blue-collar workers" offsets downwards the influence of the simple variable 'assets', which nevertheless remains a positive

factor. These results allow us to identify the categories of the population that are most likely to take out a long-term care insurance contract: women in the blue-collar worker, white-collar worker and farmer categories. We therefore determine the average probability of taking out a long-term care insurance contract for each of these categories, which allows us to compare this average probability with the average probability for all policyholders. For information purposes we add what assets and average income these categories have at their disposal (see table 5). Being a female blue-collar worker makes an individual 2.7 times more likely to be insured than members of general population. This ratio is 2.9 for female farmers and 2.2 for female white-collar workers. Here we improve our targeting by examining the wealthiest individuals in terms of assets and the oldest individuals in these categories (see table 6). The previous results are confirmed. The wealthiest female blue-collar workers are 5.8 times more likely (4 and 5.2 times more likely for female white-collar workers and farmers respectively) to take out insurance. The effect of age is as strong for the oldest female farmers, blue-collar and white-collar workers: the ratios are, respectively, 5.2, 5.9 and 5 (see table 7).

Interaction terms Model		
ExplanatoryVariables	Coefficients(SE)	p-value
Intercept	-6.9764 (0.2004)	<0.0001
Age	0.0403 (0.0035)	<0.0001
Gender	0.8378 (0.1955)	<0.0001
Senior Executive	-0.7565 (0.1751)	<0.0001
White Collar Workers	1.5468 (0.2078)	<0.0001
Blue Collar Workers	2.2362 (0.2391)	<0.0001
Farmers*Financial Asset	0.0122 (0.0034)	0.0004
Blue Collar Workers*Financial Asset	0.0563 (0.0060)	<0.0001
White Collar Workers*Age	-0.0189 (0.0036)	<0.0001
Blue Collar Workers*Age	-0.0199 (0.0042)	<0.0001
Blue Collar Workers*Income	-0.0103 (0.0024)	<0.0001
Blue Collar Workers*Gender	-0.6067 (0.1497)	<0.0001
Gender*Financial Asset	0.0156 (0.0026)	<0.0001
Age*Gender	0.00888 (0.0034)	0.0097
Blue Collar Workers*Gender*Income	0.0185 (0.0036)	<0.0001
Blue Collar Workers*Gender*Financial Asset	-0.0483 (0.0134)	0.0003
N	103090	
Wald	1670.8765	< 0.0001
Likelihood Ratio	1801.7566	< 0.0001

TAB. 4 – *Model with interaction terms.*

6 Discussion

The fact that female farmers, blue-collar workers or white-collar workers have a higher probability of taking out insurance can be interpreted in different ways. A first interpretation lead us to think that policyholders follow what we can qualify a rational behaviour. Blue-collar and white-collar workers are more likely than executives to need LTC (Mormiche and Boissonat (2003)). This phenomenon is especially pronounced when these blue-collar workers are women. Women have a probability to need LTC higher than men and for a longer average duration. The female blue-collar workers and white-collar workers represent the highest probability of loss. Moreover, they bear the highest financial risk to the extent they receive smaller pensions than men. In the event of LTC, they therefore risk not being able to cope with the costs of being looked after. It is rational that people who bear the highest financial risk insure themselves the most. As long as insurance companies can observe the level of risk, there is no adverse selection. However, it would be appropriate to verify than women don't buy insurance because they have access to hidden information concerning their likelihood of LTC. Second, we notice a significative and positive coefficient associated with the variable Age and Age*Sex. It means that people and particularly women are not sensitive to insurance's price given that age is a very good proxy of price. This is confirmed by all actuarial models used by insurance companies. Moreover, within the category which insure the most (female blue-collar and white-collar), those belonging to the most aged decile buy more insurance (5.9 more time in table 7) than the young. Then, within female blue and white collar workers, those belonging to the last decile of wealth buy more insurance. An interpretation can lead us to hold the altruism hypothesis. Rather than squandering their assets being cared for, women would appear to prefer to take out insurance so as to avoid reducing the inheritance left to their children. It is interesting to note that the middle and upper classes do not display specific behaviour regarding assets. The altruistic behaviour found in the literature on LTC does not seem to apply to well-off categories. Individuals would take out insurance against the fact of bequeathing nothing to their children. The reliability of our results should be considered in the light of several limits. First, we don't have a risk aversion variable. For this reason the fact that women of working classes buy more insurance might be interpreted by the fact that they are more risk averse. However risk aversion is very psychologic. They depend more on the character of individuals than the class to which they belong. No evident reason would explain that on average working classes are more risk averse than the other classes. Second, we don't have access to a variable concerning the sensitivity to advertising campaigns. Working classes or women might be more sensitive to them.

	Average Age	Average Wealth(€)	Average Annual Income(€)	Average Likelihood	Likelihood Ratio [4]
Farmers/Women	62	75 960	32 970	5.30%	2.9
Blue Collars/Women	47	14 700	14 730	4.90%	2.7
White Collars/Women	46	19 980	24 790	4.00%	2.2

TAB. 5 – *Average likelihood depending on gender and socio-economic category.*

4. Likelihood ratio = average likelihood of specific category/(average likelihood of whole pop.=1.82%)..

	Average Age	Average Wealth(€)	Average Annual Income(€)	Average Likelihood	Likelihood Ratio [4]
Farmers/Women					
33% the wealthiest	66	227 090	53 080	9.4%	5.2
Blue Collars/Women					
10% the wealthiest	61	125 970	34 970	10.5%	5.8
White Collars/Women					
10% the wealthiest	60	146 120	57 350	7.2%	4

TAB. 6 – *Average likelihood depending on gender and socio-economic category and wealth.*

	Average Age	Average Wealth(€)	Average Annual Income(€)	Average Likelihood	Likelihood Ratio [4]
Farmers/Women					
33% the oldest	73	106 160	31 080	9.5%	5.2
Blue Collars/Women					
10% the oldest	73	45 800	18 810	10.8%	5.9
White Collars/Women					
10% the oldest	73	49 660	23 560	9.1%	5

TAB. 7 – *Average likelihood depending on gender and socio-economic category and age.*

7 Conclusion

A logit model of long term care insurance subscription probability is developed in this study, thanks to an original estimation algorithm tested in this paper on a bank portfolio. On the one hand it provides insurers with an essential key decision making tool guiding them in implementing of futures marketing strategies. The future marketing strategies will define the operational CRM stage (using Siebel software for instance). On the other hand, it provides us today a better understanding of this long term care market about its possible evolutions and its characteristics. From an analytical CRM viewpoint, the methodology in this paper is interesting: we use an original step by step algorithm using a whole of statistical procedures to characterize the long-term care demand. From an economical viewpoint we obtain a whole of results. An important economical result of the study is that people who are most likely to take out a long-term care insurance contract are the working classes and, within this category, women. At a time when public authorities are considering a tax incentive mechanism for this type of product, it is interesting to note that the working classes display a great appetence for this type of product. A tax incentive may have a significant effect, especially as within these working classes, insurance is taken out by the wealthiest (among: those who pay taxes). LTC insurance is therefore not a product reserved for higher social classes but is better positioned to become a mass product. However the fact that probability increases with age leads us to think that price is not the main determining factor underlying purchase behaviour. The impact of tax incentives may therefore be limited by this latter factor. In a future research we would like to control our results by risk aversion and examine the presence of multidimensional adverse selection. Also, on a professional level, this paper will be able to contribute to increase offer of long term care insurance products in France.

Annexe

Main Effects Model Bootstrap			90%.Conf.Interval		
Variables	Obs.Coef.	Std.dev	Inf	Sup	Coef.of.Var.
Intercept	-6.6022	0.0786	-6.7388	-6.5004	0.0119
Age	0.0338	00010	0.0325	0.0354	0.0335
Financial Asset	0.0137	0.0008	0.0126	0.0148	0.0580
Gender	1.2926	0.0271	1.2662	1.3267	0.0210
Senior Executive	-0.9548	0.1499	-1.1581	0.7390	0.1570
White Collar Workers	0.5079	0.0541	0.4445	0.5765	0.1065
Blue Collar Workers	0.9549	0.0654	0.8616	1.0610	0.0684

TAB. 8 – *Bootstrap statistics: model with main effects only.*

Interaction Terms Model Bootstrap			90%.Conf.Interval		
Variables	Obs.Coef.	Std.dev	Inf	Sup	Coef.of.Var.
Intercept	-7.1017	0.3027	-7.5029	-6.7642	0.0426
Age	0.0425	0.0053	0.0368	0.0495	0.1243
Gender	1.0122	0.3144	0.6703	1.3937	0.3106
Senior Executive	-1.1966	0.8837	-2.8670	-0.6153	0.7386
White Collar Workers	1.5610	0.1622	1.3877	1.8271	0.1039
Blue Collar Workers	2.3080	0.2688	1.8371	2.6607	0.1165
Farmers*Financial Asset	0.0108	0.0072	-0.0000	0.0195	0.6660
Blue Collar Workers*Financial Asset	0.0489	0.0139	0.0326	0.0601	0.2836
White Collar Workers*Age	-0.0192	0.0022	-0.0229	-0.0162	0.1158
Blue Collar Workers*Age	-0.0215	0.0039	-0.0268	-0.0165	0.1831
Blue Collar Workers*Income	-0.0097	0.0040	-0.0139	0.0000	0.4156
Blue Collar Workers*Gender	-0.6240	0.1828	-0.7609	-0.3622	0.2930
Gender*Financial Asset	0.0093	0.0072	0.0000	-0.0153	0.7800
Age*Gender	0.0065	0.0057	0.0000	0.0122	0.8744
Blue Collar Workers*Gender*Income	0.0168	0.0062	0.0000	0.0208	0.3683
Blue Collar Workers*Gender*Financial.Asset	-0.0389	0.0162	-0.0587	0.0000	0.4181

TAB. 9 – *Bootstrap statistics: model with interaction terms.*

The variation coefficient or VC is a dispersion criterion allowing us to highlight the quality of the valuation of our estimated coefficient when this is calculated from a set of possible values taking into account the empirical average. If the variation coefficient is strictly less than 1, we say that the associated set or series has a low variance. Conversely, if the VC is strictly higher than 1, the series has a high variance and, in this case, the valuation obtained by the calculation of the arithmetic mean is necessarily poor in quality. The estimator studied must have a low associated variance in order to give it meaning. Each of the coefficients set out below, which, to a large extent, characterize the 4 main factors drawn from our study (old age, significant financial assets, low level of education, female) is associated with a low variance. However, it is common practice to consider a good VC from a statistical point of view to be strictly less than 33,33%.This is the case here for 9 estimators out of 15 : the constant, "Age", "Gender", "White collar workers", "Blue collar workers", "Blue collar workers*financial asset", “White collar workers*age”, “Blue collar workers*age”, “Blue collar workers*gender”. The 7 estimators, defined by this private set of estimators “White collar workers” and “White collar workers*Age”, infer a good level of predictability in our model in relation to the “Blue collar workers” category and its associated attributes : age, financial assets, gender. We therefore confirm the 4 factors in relation to this category. The “Age” and “Gender” variables have good quality estimators or a VC < 33%. The two criteria : “age” and “gender” are reliable, they are not mixed here with other variables. The “White collar workers” and “White collar workers*Age” estimators each have a very good VC < 12%. Thus, the sub-group formed by the “White collar workers”, “White collar workers*Age”, “Age” and “Gender” form a reliable set of variables in terms of predictability in relation to the “White collar workers” category. What is missing here is the asset factor to confirm the 4 factors in relation to this category. The “Blue collar workers*Gender*Income” and “Blue collar workers*Income” variables characterizing workers’ income, and specifically that of female workers, have a VC > 33%, but are not determinant in this study,which does not generally focus on the “income” factor. The “Blue collar workers*Gender*Financial asset” variable, which corrects the “Blue collar workers*financial asset” variable for female workers has a VC slightly above 33%. The “Age*Gender” variable has the lowest VC: it is almost equal to 0.9, *i.e.* very close to 1. When comparing the 90% maximum value of the coefficient of this variable equaling 0.0122 and that of the estimator of the Age variable equaling 0.04215 with a very good VC (close to 10%), we can put the weight of uncertainty of the “Age*Gender” variable into perspective.

The “Farmer*Financial asset” variable specifically characterizes farmers financial assets. This variable has a variation coefficient > than 33% and a low variance. It plays an important role in relation to the factors drawn out. Long term care insurance is still a new risk and the insurance rate is still low : less than 2% of the overall population. Consequently, it is of no surprise to note that, within this context, reliability problems occur with regard to a small socio-economic group in relation to the overall population, such as ”farmers”, and the estimators specifically associated with them. Given this context, we will accept this variable even with a VC > 33% equal to 66%. Finally: the sub-group made up of “Farmer*Financial Asset”, “Age”, “Gender” forms an acceptable set of variables in terms of predictability in relation to the “Farmer” category and its associated characteristics: age, financial assets, gender. We can therefore confirm the 4 factors in relation to this category.

The “Gender*Financial asset” variable’s VC is far from 33% and < than 0.8. Previously, we demonstrated the significance of the 4 factors produced in relation to the “workers” category and the “farmers” category. The sub-group highlighted for the “employees” category made up of the variables: “White collar workers”, “White collar workers*Age”, “Age”, “Gender”, produced 2 significant factors: “Age” and “Gender”. The “Gender*Financial asset” variable has a poor VC but a low variance classification: its estimators are systematically positive in our “bootstrap”, which points to a positive correlation between high assets and taking out long term care insurance. Furthermore, this variable is the only variable missing for the simultaneous confirmation of the significance of our 4 factors over the 3 categories (employees, workers, farmers), and this is within the context of low numbers of policy holders in relation to a new risk. All of these conditions lead us to consider this variable valid in terms of interpretation. Thus, the sub-group formed by the “White collar workers”, “White collar workers*Age", "Age", "Gender", "Gender*Financial asset" is an acceptable set of variables in terms of predictability in relation to the "employees" category and the attributes attached to it: age, financial assets, gender. We can therefore confirm the 4 factors here.

References

Brown, J. and A. Finkelstein (2007). Why is the market for long-term care insurance so small ? *Journal of Public Economics 91*, 1967–1991.

Brown, J. and A. Finkelstein (2008). The interaction of public and private insurance: Medicaid and the long term care insurance market. *American Economic Review 98*(3), 1083–1102.

Cadoux, D. and J.-M. Loizau (2004). Copules et dépendances : application pratique à la détermination du besoin en fonds propres d'un assureur non vie. *Bulletin Français d'Actuariat 6*(12), 173–208.

Costa-Font, J. and J. Rivera-Forns (2008). Who is willing to pay for long term care insurance in catalonia? *Health Policy 86*, 72–84.

Cour des comptes (2005). Les personnes âgées dependantes. Rapport de la cour des comptes, au président de la république.

Courbage, C. and N. Roudaut (2007). La demande d'assurance dépendance, une analyse empirique pour la france. *29ème Journée des Economistes de la Santé, Lille*.

CSA (2006). Attitudes et comportements face au risque dépendance. *FFSA*.

Cutler, D. (1993). Why doesn't the market fully insure long term care? *NBER Working* (4301).

Decoster, B. (2006). Les contrats d'assurance dépendance en 2005. *FFSA*.

Duée, M. and C. Rebillard (2004). La dépendance des personnes âgees: une projection à long terme. *INSEE, série des documents de travail de la Direction des Etudes et Synthéses Economiques* (G2004/2).

Ennuyer, B. (2006). *Repenser le maintien à domicile: enjeux, acteurs, organisation*. Dunod.

Finkelstein, A. and K. M. McGarry (2006). Multiple dimensions of private information: evidence from the long term care insurance market. *American Economic Review 96*(4), 938–958.

Gisserot, H. and E. Grass (2007). Perspectives financières de la dépendance à l'horizon 2025 : prévisions et marge de choix. *Ministère de la sécurite sociale, des personnes âgées, des personnes handicapées et de la famille*.

Haut Conseil pour l'Avenir de l'Assurance Maladie (2005). Rapport 2005 du haut conseil pour l'avenir de l'assurance maladie.

Hsin Lin, Y. and Y. Hern Chang (2008). Significant factors of aviation insurance and risk management strategy: an empirical study of taiwanese airlines carriers. *Risk Analysis 28*(2), 453–461.

Kumar, N., M. A. Cohen, C. E. Bishop, and S. S. Wallack (1995). Understanding the factors behind the decision to purchase varying coverage amounts of long-term care insurance. *Health Services Research 29*(6), 653–678.

Lebart, L. (1971). Applications des méthodes d'analyse des données à la préparation des enquêtes auprès des ménages. *Journal de la Societe Statistique de Paris*, 108–120.

Loones, A., E. David-Alberola, and S. Eugène (2005). Approche du coût de la dépendance des personnes âgées à domicile. *CREDOC* (C221).

Mace, Y. (2003). Effets des mesures de prévention : un paradoxe lié à l'antisélection. *Bulletin Français d'Actuariat 6*(10), 11–22.

McCall, N., S. Mangle, E. Bauer, and J. Knickman (1998). Factors important in the purchase of partnership long-term care insurance. *Health Services Research 33*(2), 187–203.

Meier, V. (1998). Long-term care insurance and life insurance demand. *The Geneva Papers on Risk and Insurance Theory 23*, 49–61.

Mitchell, O., J. Piggott, and S. S. (2008). An empirical analysis of patterns in the japanese long-term care insurance system. *The Geneva Papers 33*, 694–709.

Mormiche, P. and V. Boissonat (2003). Handicaps et inegalités sociales : premiers apports de l'enquête « handicaps, incapacités, dépendance ». *Revue française des affaires sociales* (2), 267–285.

OECD (2005). The oecd health project: long term care for older people. *OECD publishing*.

Sloan, F. A. and E. C. Norton (1997). Adverse selection, bequests, crowding out, and private demand for insurance : Evidence from the long-term care insurance market. *Journal of Risk and Uncertainty 15*, 201–219.

Thomas, A. (2000). *Econométrie des variables qualitatives*. Dunod.

Trigo-Gamarra, L. (2008). Reasons for the coexistence of different distribution channels : an empirical test for the german insurance market. *The Geneva Papers 33*, 389–407.

Vasechko, O., M. Grun-Réhomme, and N. Benlagha (2009). Modélisation de la fréquence des sinistres en assurance automobile. *Bulletin Français d'Actuariat 9*(18), 41–63.

Zweifel, P. and W. Struve (1996). *Long term insurance and trust saving in a two-generation model*. Kluwer Academic Publisher.

Résumé

Afin de comprendre la défaillance du marché européen de l'assurance dépendance privée nous étudions les caractéristiques de sa demande en modélisant la probabilité de souscription sur données bancaires, grâce à un modèle logistique et à un algorithme original d'estimation pas à pas. L'appartenance aux classes aisées réduit cette probabilité tandis que les femmes des classes populaires possédant un patrimoine financier conséquent (même résultat pour les plus âgées) ont 5 fois plus de chances de souscrire. Ainsi l'assureur se trouve dans une position favorable pour développer son portefeuille dépendance : le vieillissement rendra ces cibles plus prépondérantes dans le futur. Cette étude, la seule étude empirique sur données françaises, fournit à l'assureur un outil d'aide à la décision économétrique d'estimation de la probabilité de souscription, qu'il peut appliquer sur son propre portefeuille.