

## RESEARCH PAPER

# Big Data and Insurance: Advantageous Selection in European Markets

Francesco Corea

LUISS Guido Carli, IT  
fcorea@luiss.it

Rothschild and Stiglitz (1976) argued that people signal their risk profile through their insurance demand, i.e. individuals with a high risk profile would buy insurance as much as they can, while people who are not going to buy any insurance are the ones with a lower risk profile. This issue is commonly known as adverse selection. Even if their prediction seems to work quite well in a lot of different markets, Cutler et al. (2008) proved that there exist some insurance markets in United States in which the expected result is completely different. In the wake of this study, we provide empirical evidences that there are some European insurance markets in which the low risk profile agents are the ones who buy more insurance.

**Keywords:** Adverse selection; Asymmetric information; Fixed-effects; Annuity; Long-term care; Medigap

## 1 Introduction

The insurance market is usually the most common example used in textbooks trying to explain the impact of the information on any economic activity. Indeed, the model proposed (Rothschild and Stiglitz, 1976) is usually quite straightforward: an insurance company should be suspicious concerning people who want to buy some coverage because only individuals with a high expected claims are willing to pay a premium for being compensated in case an accident occurs. Therefore, asking for an insurance is thus a signal that a person will need to be reimbursed at some point in future. Since the insurance company makes profit on the probability that not every client will need to be paid more than the premium deposited, it is also not going to sell any insurance if it is certain that every client will need to be paid in the contract lifetime. On the other hand, people that are not expected to have a high claim in future are not willing to pay any premium for being insured. This is an asymmetric informational issue called in literature *adverse selection* (Akerlof, 1970).

Hence, the insight behind this concept is that the correlation between the individual's demand for insurance and the risk of losses has to be positive. In the health sector, many works tested this positive correlation idea, such as Mitchell et al. (1999) for the American annuities market, while McCarthy and Mitchell (2003) focused on the Japanese annuities market and Finkelstein and Poterba on the English one in different works (2002, 2004, 2006). A more extensive review of the verification of the positive correlation between insurance coverage and risk occurrence can be found in Cutler and Zeckhauser (2000).

The framework has also been extended in several different ways, but the prediction is again confirmed, as for instance proved in Chiappori and Salanie (2000) and Chiappori et al. (2006).

On the other hand, even if the classic and intuitive adverse selection hypothesis has been validated and proved to be robust in many circumstances, some influential exceptions exist. Indeed, Einav et al. (2010), Einav et al. (2011), Cardon and Hendel (2001), as well as Cutler et al. (2008), and Finkelstein and McGarry (2006a) among all, showed that the prediction of positive correlation fails in some countries and markets, even in sectors other than health (Dionne et al., 2001; Cohen and Einav, 2007). In particular, Medigap insurance demand seems to be negative correlated with the risk occurrence (Ettner, 1997; Fang et al., 2006; Hurd and McGarry, 1997), as well as life insurance (Cawley and Philipson, 1999) and long-term care (Finkelstein and McGarry, 2006b). This seems to be due to a wider spectrum of private information owned by the individuals

that would entail a preference heterogeneity and an unexpected irrational coverage. The result of these analysis has been named *advantageous* or *propitious* selection (De Meza and Webb, 2001; Hemenway, 1990).

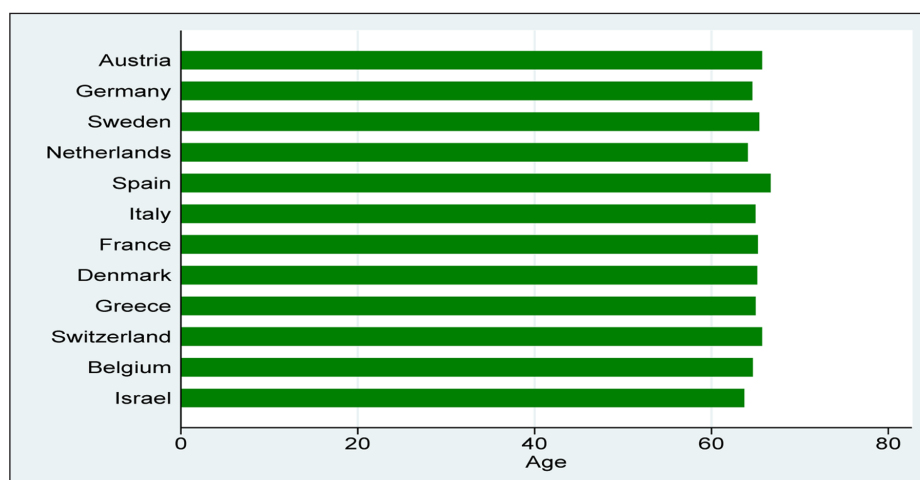
Some explanations are identified in the variable individual's risk tolerance. In fact, preference heterogeneity for both risk tolerance and risk type may let the sign between insurance demand and accident occurrence to be anyone (Einav et al. 2007), since I) individuals with lower (higher) risk tolerance can either buy more (less) insurance or invest in instruments or activities that lower (higher) the expected claims; and II) individual's behaviour may vary across different markets, i.e. the correlation may be positive for some markets and negative for others.

On the wave of the works above mentioned, the purpose of our analysis was to verify (or disprove) whether the correlation between insurance coverage and risk occurrence was indeed positive, or on the other hand negative or absent within European countries. Five different health insurance markets have been considered: term life, annuity, long-term care, acute health and eventually Medicare supplemental insurance (Medigap), as already proposed in Cutler et al. (2008). The demand of each one of this insurance type has been studied with respect to both risky behaviours (i.e., behaviours suitable to proxy risk tolerance) and risk occurrence (i.e., the event that should trigger the payment from the insurance company).

The work is then structured as follows: the next section will explain the data used, how the main database and variables have been built, and the kind of approach used for the analysis. Section 3 will present the results from the different regressions run, and it will compare and comment on the outcomes obtained. Finally, section 4 will sum up and conclude.

## 2 Data and empirical framework

As already mentioned in Section 1, the purpose of the analysis is to see what kind of relationship exists between five insurance market demands, risk tolerance and risk occurrence within different European countries. The analysis implemented used micro panel data on health from the Survey of Health, Ageing and Retirement in Europe (SHARE) project. We used a sample of people aged more than 51 in 2004–2005, for eleven European countries (Austria, Germany, Sweden, Netherlands, Spain, Italy, France, Denmark, Greece, Switzerland, Belgium) plus Israel.<sup>1</sup> The Appendix presents key summary statistics for each country. **Figure 1–2** show the average age to the population and the average medigap expenses during the period considered. As it can be seen, the average age is pretty stable across countries, with the highest pick corresponding to Spain, followed then by Austria, Sweden and Switzerland. On the other hand, Sweden is the country in which people spend the most in additional medicines and/or cure, i.e. where the people buy a supplementary insurance more likely. Another Scandinavian country, the Denmark, is ranked second, followed directly by Israel and Italy. If we instead have a look to the **Figure 3**, we can observe that for almost each country the population is on average slightly overweighted. There are more obese than underweighted, and these two measures seem to be at a glance inversely correlated. The **Figure 4** claims instead that, on the total population considered, only a small amount of persons undertake preventive health actions, and this happens in particular in Germany, Greece and Spain (Italy just following). Within the group of persons who take actions of the kind described above, it is very common to do the minimum possible, i.e. undertaking only one preventive measure (this is particularly true in Greece and in Switzerland, Germany and Austria).



**Figure 1:** Key summary statistics for average age per country.

<sup>1</sup> The panel nature of the dataset was essential, for instance, to track mortality and nursing home.

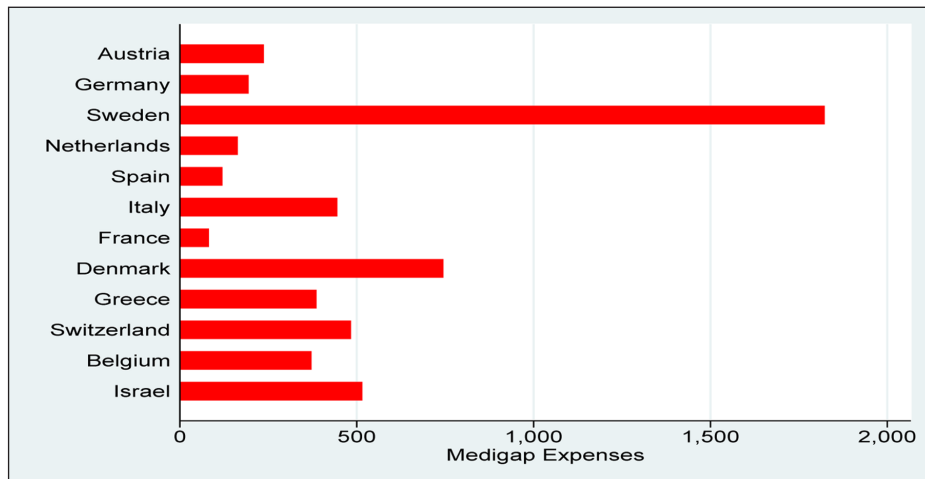


Figure 2: Key summary statistics for medigap expenses per country (%).

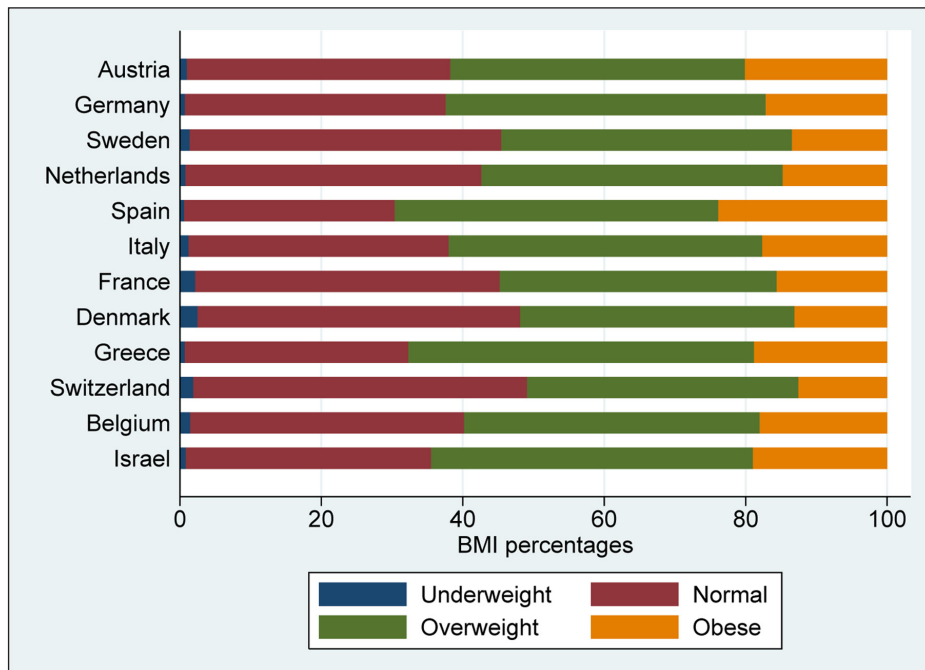


Figure 3: Key summary statistics for bmi index per country (%).

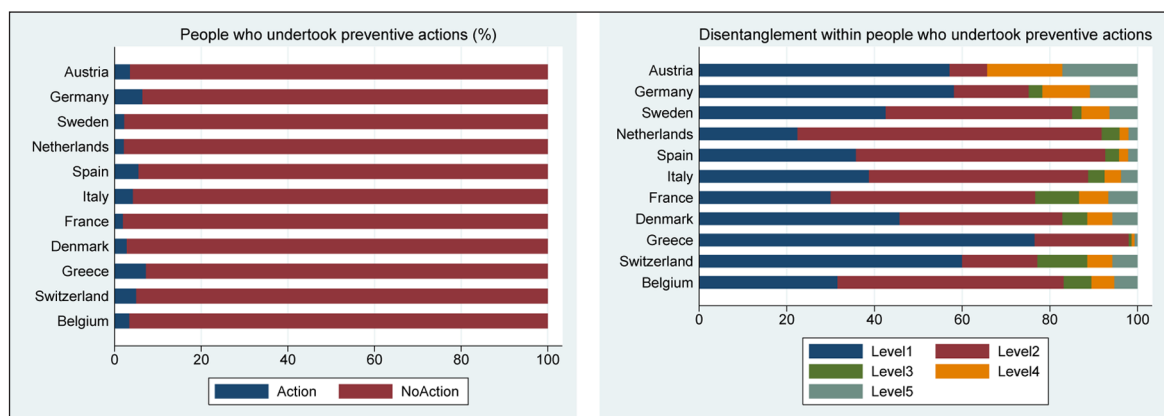
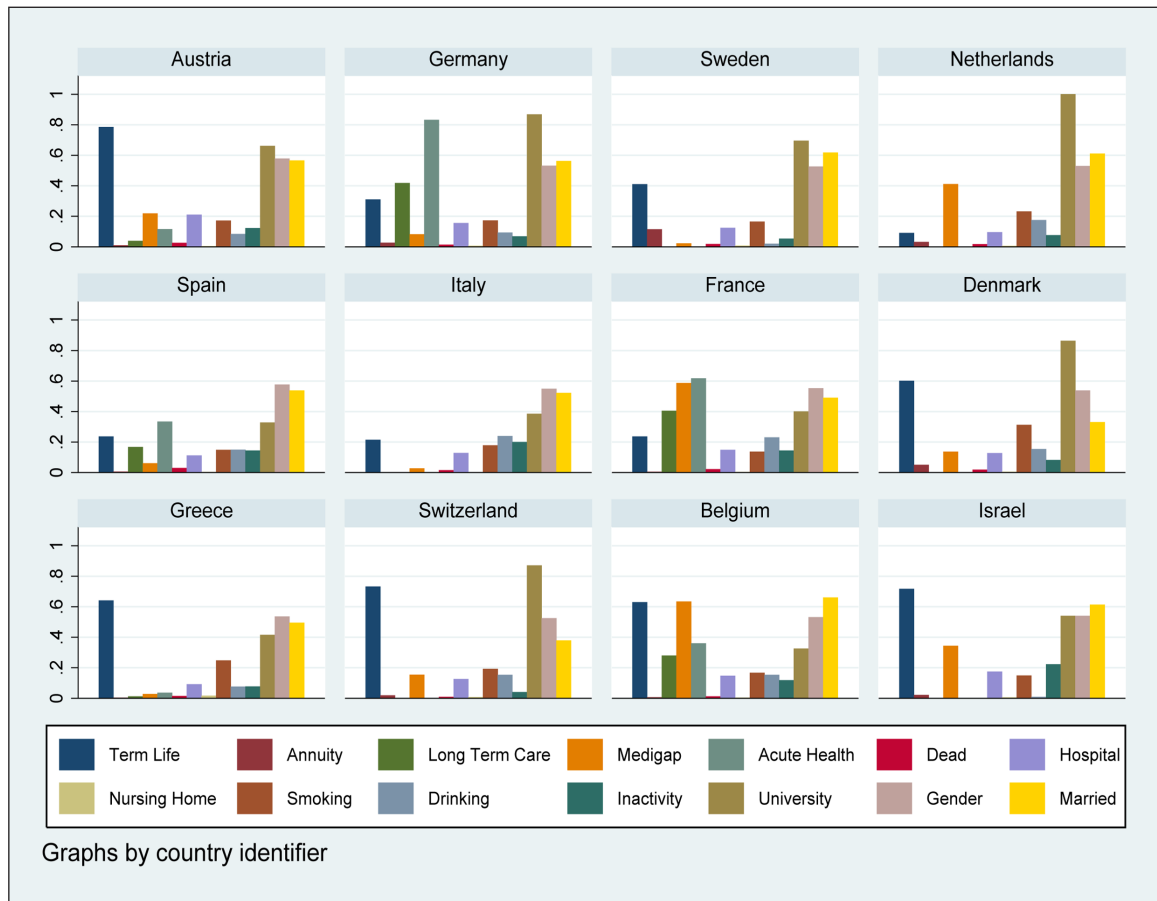


Figure 4: Key summary statistics for different level of prevention (number of preventive actions) per country (%).



**Figure 5:** Key summary statistics for other variables (%).

There is a consistent amount of people who go further and implement a second preventive action as well, but above that threshold the number shrinks toward very low levels (Greece is emblematic from this point of view, since it has the highest percentage of people undertaking one single action and the lowest of who undertakes more than two preventive actions). The best examples here are The Netherlands, Spain and Belgium. Finally, **Figure 5** exhibits a wider spectrum of variable summary statistics, expressed in percentage terms, for the groups of insurance coverages, the remaining risky behaviour and risk occurrence variables, and finally for the controls as well. Instead of focusing on a single variable, what we infer from this last figure is the high heterogeneity within the population. Already since this figure, we observe how this sparsity may be reflected in heterogeneous preferences, a fundamental concept which may help us in enlighten the advantageous selection phenomenon.

From the SHARE survey, we indeed extracted several answers to construct the variables used in our regressions. In particular, as insurance and risk occurrence, we measured:

- Life insurance as whether the individual has a term life insurance at the time of the survey (or both term and whole life policies), and the correspondent occurrence is whether the individual dies between 2004 and 2006/7. According to Cutler et al. (2008), we use the term life insurance since it represents a pure investment compared to a whole life insurance, where we should take care also about the saving component;
- Acute health as whether the individual has a hospital care with unrestricted choice of hospitals/clinics and/or hospital care with limited choice of hospitals and clinics. The risk occurrence is whether the individual has been in a hospital in the last twelve months;
- Annuity as whether the individual has a personal and private annuity insurance, with the corresponding risk occurrence of whether the interviewed is alive at the time of the second survey (2006/7);
- Medigap as whether the individual has a supplementary insurance.<sup>2</sup> The risk occurrence here is

<sup>2</sup> In particular, a person has a supplementary insurance if he has at least one of the following: Medical care with direct access to

the amount the individual incurred as extra medical expenses;<sup>3</sup>

- Long-term care as whether the individual has at the time of the survey a long term care in nursing home insurance and/or a nursing care at home in case of chronic disease or disability. The corresponding risk occurrence is whether the individual has been into a nursing home between 2004 and 2006/7.

Instead, as proxy for risk tolerance, we decided to use the following measures able to capture the risk preferences:

- Smoking, i.e. whether the individual currently smokes;
- Drinking problems, that is whether the individual drinks two or more glasses of alcohol each day or 5/6 days a week;
- Body mass index (BMI), considered as an indicator of incorrect actions about individual's diet, is computed as individual's weight divided the square of the height, times 10,000. In this way, it has been possible to classify the individual under the following four categories: Underweight (BMI below 18.5), Normal (18.5–24.9), Overweight (25–29.9) and Obese (30 or higher). Finally we assigned 0 to the variable if the weight was in the normality range, 1 otherwise;
- Level of physical inactivity, defined as never or almost never engaging in neither moderate nor vigorous physical activity;
- A variable reflecting preventive health actions followed out by the interviewed.<sup>4</sup>

Therefore, we run the following two different regressions:

$$Pr(Y_i | PRT_i) = \alpha_0 + \alpha_1 * PRT_i + X_i\Gamma + \epsilon_i \quad (1)$$

$$Z_i = \beta_0 + \beta_1 * PRT_i + X_i\Pi + \eta_j \quad (2)$$

where  $Y_i$  represents the fact that an individual has or not the particular kind of insurance under analysis,  $PRT$  stands for *Proxy of Risk Tolerance*, that is the behavioural variables discussed above, while  $Z_i$  is the risk occurrence for the insurance studied, and  $X_i$  are the covariates (gender, age, education and marital status).<sup>5</sup> We then run both the unconditional regression and the one controlled for the covariates. The control variables are used according to the usual insurance practices and are applied differently with respect to the insurance markets. Indeed, about the term life/long term insurance we will control for education, age and gender; then we will check the Medigap for education and age, the annuity for age, gender, education and marital status and the acute health only for education.<sup>6</sup> We decided after careful consideration to use the probit in the model 1 because, although does not differ almost at all from a standard least squares regression model, it provides a better probabilistic interpretation. The model 2 is instead a classic least square estimation.

Since we should also embed somehow the differences due to being analysing different countries, we decide to follow the Bryan and Jenkins' approach (Bryan and Jenkins, 2013) on hierarchical (multilevel) datasets. According to them, to prove the robustness of our analysis we are going to run a simple pooled

---

specialists; Medical care with an extended choice of doctors; Dental care; A larger choice of drugs and/or full drugs expenses (no participation); An extended choice of hospitals and clinics for hospital care; (Extended) Long term care in a nursing home; (Extended) Nursing care at home in case of chronic disease or disability; (Extended) Home help for activities of daily living (household, etc.); Full coverage of costs for doctor visits (no participation); Full coverage of costs for hospital care (no participation).

<sup>3</sup> It has been computed as the total sum in euros of paid out-of-pocket expenses for inpatient care, paid out-of-pocket expenses for outpatient care, paid out-of-pocket expenses for prescribed drugs and paid out-of-pocket expenses for day care, nursing home and home-based care.

<sup>4</sup> This variable has been constructed as an indicator of whether the individual has consulted a specialist for regular controls, whether he had a flu vaccination in the last year, whether he had a sigmoidoscopy or colonoscopy less than 10 years ago, whether he had a mammogram (x-ray of the breast) and if he had another test to detect hidden blood in his stool in the last 10 years. From each action undertaken, he got one point and the final indicator is expressed as the sum of all the point obtained, i.e. if an individual has the preventive variable equal to two it means that he did only two preventive actions out of five.

<sup>5</sup> The education variable has been set as a binary variable on whether the individual has pursued or not a higher level of studies, such as university, college, nursery school, etc. In addition, the marital status variable has been created as well as a binary variable, on whether the individual is married/in a registered partnership or not married/divorced/widowed.

<sup>6</sup> For a more detailed definition of risk classification controls, see Cutler et al. (2008).

regression, a separate regression for each country and a country fixed effect model. This multiple choice could prove the results to be not related to the technique used and will improve the understanding of the phenomenon we are trying to capture providing different interpretations of the data.

First of all, a pooled regression with clustered-robust errors is going to be run. This would ignore that different countries have different unobserved features and will underestimate the standard errors of  $\beta$ , but it could be easily corrected using countries-robust standard errors that allow for a more general correlation within countries.

The second analysis implemented concerns instead a separate regression for each country. The country effect is in this way internalised and it is merged with the intercept of each regression model. It is a bit computationally more demanding, but it allows to put no restrictions on the variances of country-specific errors and to let  $\beta$  to vary across countries.

The final approach used is the fixed effect estimation, and it is set as a middle way between the two models explained above. It indeed pooled all the data but allows the intercept to differ across countries to be able to capture individual-specific effects. The other greatest difference with the single-country regression is that the residuals are here constrained to be the same across countries. Besides, it is useless to include further country-level variables, since the intercept embeds country differences. Every regression will then be corrected for cluster-robust errors and cross-sectionally weighted by the weights system provided by SHARE.<sup>7</sup>

### 3 Results

The first two regressions presented in the Appendix are the pooled regressions. At a first glance, it seems that at an aggregate level the effects are not so weak, although very sparse. Indeed, as shown in **Table 1**, even if some of the results are generally either not significant or confirming the classic adverse selection theory, some relationships between insurance coverage and risky behaviours proved to be robust, meaningful and able to confirm our initial hypothesis of advantageous selection in European markets. Furthermore, the control variables seem to not affect considerably the estimation results. For instance, according to the classic theory individuals who currently smoke or drink should buy more insurance, but in reality they are more likely to buy less insurance. This is particularly true for long term care and term life/acute health respectively for smoking and drinking, and the same it is also verified for annuity markets and long term care for people physically inactive and for who implements more preventive health care actions. In addition, people not in the normality weight range are actually going to buy few insurance in three different markets, i.e. annuities, medigap and acute health.

In addition, the **Table 2** shows that both smoking and physical inactivity increase the likely to die (and to not live long). While drinking seems to not be statistically significant in any circumstances, physical inactivity will also involve a higher level of medigap expenses as well as a higher likely to be hospitalised, as expected. On the other hand, preventive health actions reduce this risk and the smoking does not increase the chance to get hospitalised. This may seem counterintuitive, but since we considered a short time hospitalisation period and since the smoking effect are quite long term, it may be reasonable that the two variables are not positively correlated. Surprisingly, some anomalies characterise the BMI variable, meaning that the BMI seems to not reduce the life expectation. Further studies may be necessary in order to understand the reason why these kind of anomalies happen, but in general we may think of some psychological disease, misperception of the illness or simply the stress as possible causes of those strange phenomena, since it seems reasonable that people who, for example, are hypochondriac (or that somatizing a lot) are the ones who implement more prevention, who then spend more in extra medicines and cure and the ones who go to the hospital more likely as well. One general interpretation of the deviations presented is that maybe more risk averse individuals have less risky behaviours, and are the ones who value the insurance the most.

As above mentioned, the results are not verified for all the insurance markets and with respect to each dependent variable, but already in the comprehensive overall regression they provide robust insights about the advantageous selection issue.

After that, we run instead the Linear Probability Model analysis at a country-level. A regression for each country has been run and the results are visualised in Appendix as **Figures 6–9**. There are five subgraphs corresponding to each insurance market and each coefficient for every independent variable is drawn by a smaller circle and a line that represents the confidence interval for that coefficient estimates at a level of 95%. For the sake of completeness, even if the results are not extremely different, the following figures have also

<sup>7</sup> For a more detailed explanation of the weights system, look at SHARE release guide for wave 1, pag. 39–46.

**Table 1:** Relation between Insurance and Risky behaviours (Pooled Probit regression).

	Term life		Annuity		Lt care		Medigap		Acute health	
	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Main Smoking	0.110* (2.37)	0.130** (3.25)	0.0389 (0.37)	0.0276 (0.29)	-0.296*** (-5.43)	-0.277*** (-4.25)	-0.103 (-1.08)	-0.138 (-1.94)	-0.0322 (-0.49)	-0.0308 (-0.65)
Drinking	-0.285*** (-4.25)	-0.310*** (-5.28)	-0.0729 (-0.68)	-0.0678 (-0.66)	-0.175 (-1.15)	-0.286 (-1.88)	0.241 (1.43)	0.243 (1.87)	-0.387*** (-3.44)	-0.389** (-3.26)
BMI	0.0283 (0.59)	0.0122 (0.23)	-0.115** (-2.62)	-0.116 (-1.91)	-0.0138 (-0.16)	-0.114 (-1.70)	-0.151*** (-4.28)	-0.154*** (-3.33)	-0.311*** (-4.40)	-0.311*** (-4.70)
Preventive	-0.0431 (-1.17)	-0.0363 (-1.06)	-0.100** (-3.25)	-0.110*** (-4.47)	0.139 (1.52)	0.183 (1.53)	-0.0598 (-1.37)	-0.0525 (-1.29)	0.189** (3.03)	0.189** (2.75)
Inactivity	-0.179 (-0.97)	-0.215 (-1.50)	-0.408*** (-7.19)	-0.468*** (-6.72)	-0.716* (-2.13)	-0.819** (-3.07)	-0.146 (-1.49)	-0.0985 (-1.73)	-0.0624 (-0.62)	-0.0683 (-0.36)
N	2657	2657	22221	22221	1269	1269	22233	22233	1269	1269

There are two different regressions for each variable: on the left the unconstrained one, while on the right the one controlled for covariates.  
 \* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$ .

**Table 2:** Relation between Risk occurrence and Risky behaviours (Pooled LPM regression).

	Dead		Alive		Nursing Home		Medigap Exp		Hospital	
	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Smoking	-0.00399 (-1.86)	0.00529* (2.47)	-0.0217 (-1.33)	-0.0285 (-1.84)	-0.00198 (-1.42)	-0.000663 (-0.42)	-56.18 (-1.20)	-8.348 (-0.31)	-0.0223** (-3.92)	-0.0238*** (-4.64)
Drinking	0.00745 (1.87)	0.0000605 (0.01)	0.0507 (1.87)	0.0483 (1.93)	-0.00166 (-1.05)	-0.00272 (-1.23)	-45.09 (-1.66)	-42.96 (-1.28)	-0.00409 (-0.60)	-0.00232 (-0.29)
BMI	-0.00285 (-0.88)	-0.00398 (-1.00)	0.0202** (3.88)	0.0161** (3.17)	-0.000794 (-0.42)	-0.000787 (-0.44)	-77.96 (-0.96)	-71.38 (-0.88)	0.00773 (0.84)	0.00825 (0.89)
Preventive	0.00267 (0.72)	0.00301 (1.20)	-0.000823 (-0.11)	0.00228 (0.29)	-0.000145 (-0.10)	-0.000224 (-0.16)	-6.056 (-0.39)	-16.64 (-1.08)	0.0262*** (9.14)	0.0260*** (8.91)
Inactivity	0.0777*** (12.06)	0.0635*** (10.11)	-0.103* (-3.08)	-0.0841* (-2.68)	0.0164 (1.81)	0.0151 (1.90)	253.3** (4.05)	190.1** (3.93)	0.169*** (17.17)	0.173*** (13.96)
N	22233	22233	22233	22233	15040	15040	22233	22233	22226	22226

There are two different regressions for each variable: on the left the unconstrained one, while on the right the one controlled for covariates.

\* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$ .

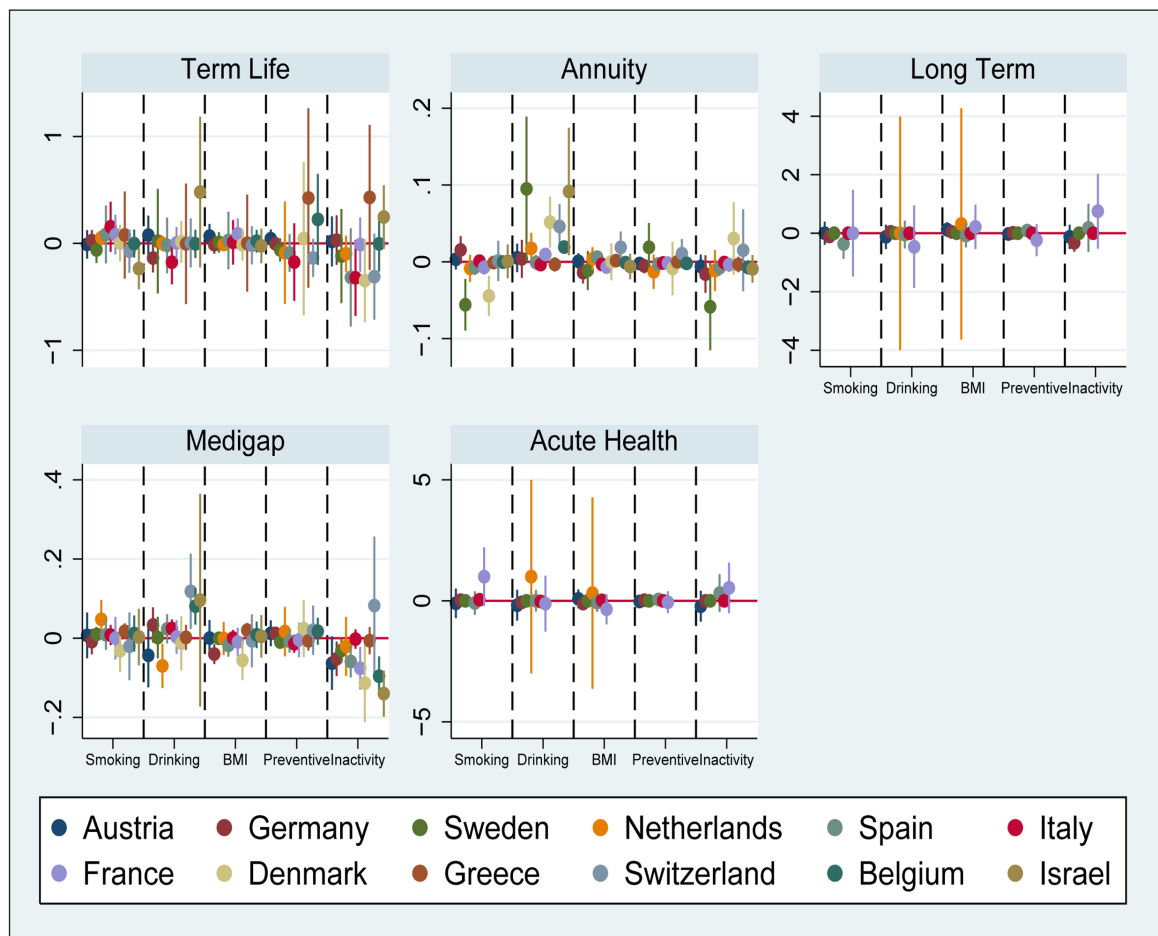


the coefficient estimates taking into account the control variables. The results are clearly not so distant from the ones observed at an aggregate level, but they are again really mixed within each country and insurance market. What it should be noticed from these graphs are the numbers of point under/above the zero line, since as before we are more interested in the sign of the relations more than in the magnitude. In particular for the term life, the annuity and the medigap insurance markets, having riskier behaviours or taking less care about own health does not directly entail a higher demand of insurance. Again, the relation between the risk occurrence and the risky behaviours is instead generally confirmed, in particular regarding physical inactivity or the smoking addiction.

The final regressions showed in the Appendix regards the country fixed effect model (with cluster-robust errors), that is usually used in this situation because, with respect to for instance a random effect model, it underlines the unique features of each country. In the regressions run here, the control variables looked still to not have a crucial role.

The **Table 3** points out again that, as expected, people who smoke or drink/with weight problems, are more likely to buy a term life or a long term care insurance, respectively. The opposite is instead verified still for smoking, drinking and BMI with respect to long term care, term life and acute health markets. The prevention is still ambiguous, since if from one hand shows an expected result such as the negative correlation with the annuity insurance purchase, on the other hand involves a positive relation with the acute health market, that is to some extent counterintuitive. Finally, physical inactivity proved again to provide the most robust results, i.e. it is negatively correlated with annuities, long term care and medigap as well. All our consideration may still make sense, behaviourally speaking, if we think again about people affected by apprehension or hypochondria, or physical inactivity reflected also in disregarding for personal care.

On the risk occurrence side instead (see **Table 4**), smoking is as expected associated to a higher chance to die (and to not live long), as well as physical inactivity, that proved also to be positively correlated with



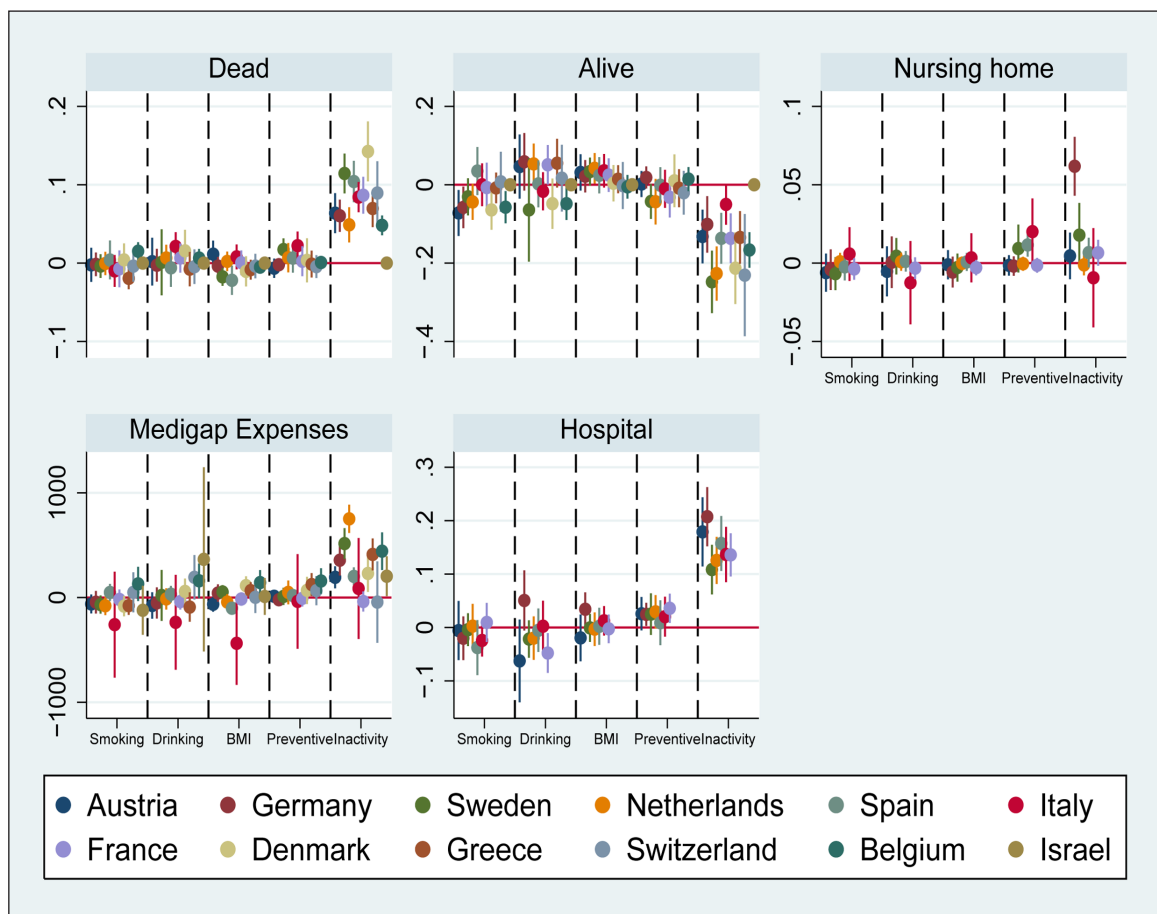
**Figure 6:** Relation between Insurance and Risky behaviours (LPM regression) per country.

medigap expenses and hospitalisation. Prevention may require, as above mentioned, a higher possibility to get hospitalised, while counterintuitively the BMI is positively correlated with a higher life expectation and the smokers are less likely to go to the hospital (in one year time).

Even in the country-fixed effect framework, although the results are less strong than in the pooled regression case, some anomaly seems to persist, and we believe the reasons behind this deviation could be interesting to be investigated in future works. We cannot conclude univocally in favour of our initial hypothesis neither in the fixed-effect scenario, but we can claim that the standard adverse selection theory seems to not hold strongly as the theory stated.

### 4 Conclusions

Our analysis aimed to investigate whether an advantageous selection phenomenon was proved to be robust in different insurance markets, as in Cutler et al. (2008). We focused on five insurance markets for eleven European countries plus Israel, specifically on term life, annuity, long term care, Medigap and acute health insurances. Our main finding has been that it looks like that riskier behaviours are not always associated with higher mortality, but above all they are not unconditionally associated with higher insurance demand as the classic theory would predict. This result does not hold for each country and each market with respect to each risky behaviour, but the outcomes are mixed, suggesting that further analysis may shine a light on this puzzle. In particular, in the most robust analysis, no systematic relation between risky behaviours and any of the insurance market, although some risky behaviours are not coherent (while others are) with Rothschild and Stiglitz (1976). In any case, it is interesting to notice that the adverse selection proposed in the '70s does not hold anymore so strongly and extensively, but also to consider that maybe preferences heterogeneity for insurance could explain the different behaviours of the participants. A different risk tolerance may indeed explain the insurance puzzle, but of course further investigations will be required in order to test this hypothesis.



**Figure 7:** Relation between Risk occurrence and Risky behaviours (LPM regression) per country.

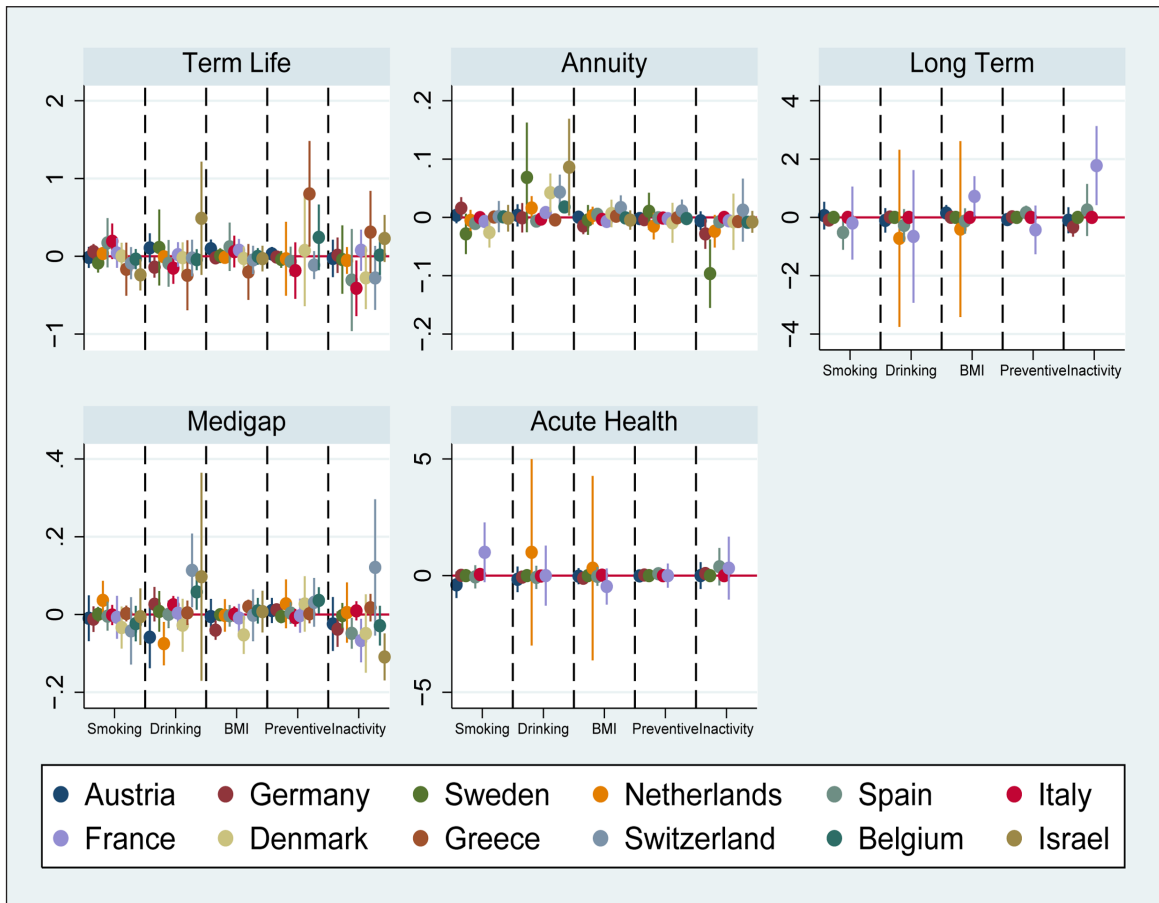


Figure 8: Relation between Insurance and Risky behaviours (LPM regression) per country with control variables.

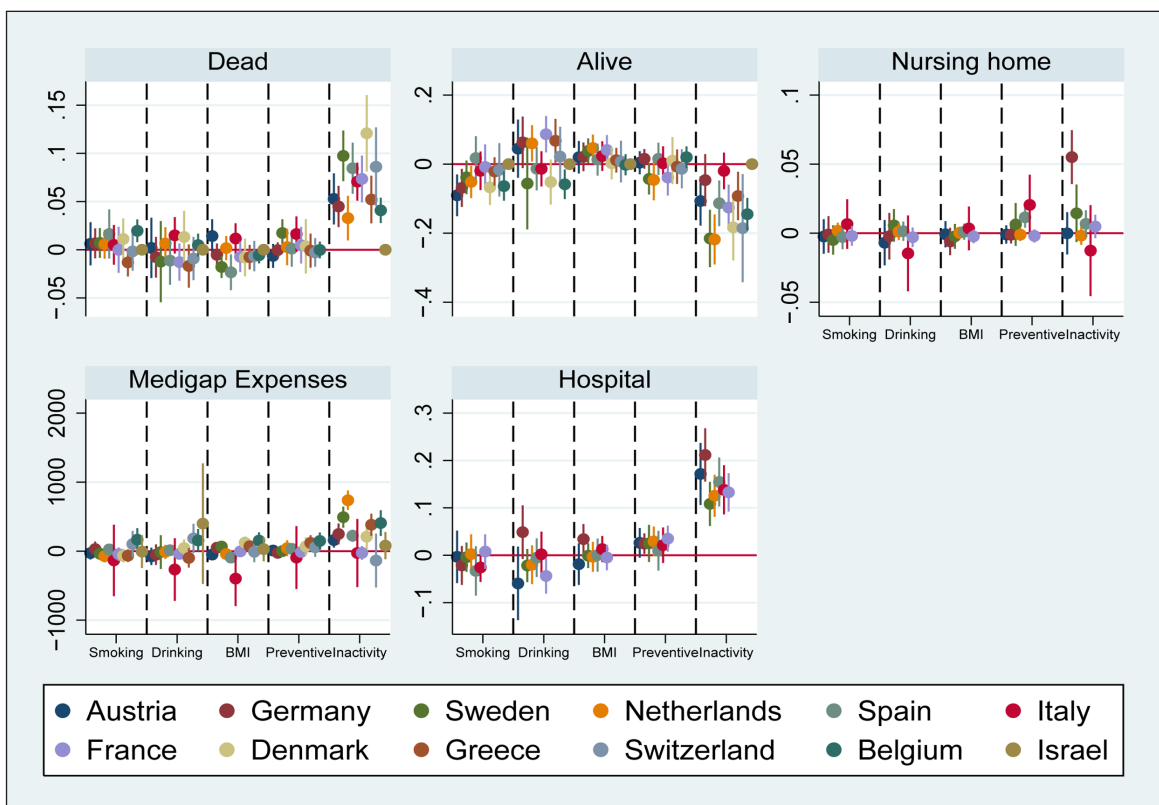


Figure 9: Relation between Risk occurrence and Risky behaviours (LPM regression) per country with control variables.

**Table 3:** Relation between Insurance and Risky behaviours with fixed-effect (Probit regression).

	Term life		Annuity		Lt care		Medigap		Acute health	
	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Main Smoking	0.0947* (2.40)	0.120** (2.70)	-0.00152 (-0.01)	0.0190 (0.18)	-0.347*** (-5.78)	-0.315*** (-3.99)	0.0218 (0.86)	-0.0340 (-1.68)	0.0961 (1.29)	0.0445 (0.70)
Drinking	-0.194 (-1.76)	-0.227* (-2.22)	0.161 (1.68)	0.0964 (0.98)	0.0955* (2.23)	-0.0229 (-0.33)	0.0988* (2.00)	0.0764 (1.74)	-0.0981 (-0.68)	-0.0613 (-0.37)
BMI	0.0435 (0.83)	0.0270 (0.50)	-0.123 (-1.90)	-0.128 (-1.87)	0.0999* (2.48)	-0.00520 (-0.12)	-0.0804 (-1.81)	-0.0709 (-1.49)	-0.459*** (-13.05)	-0.464*** (-14.13)
Preventive	-0.0428 (-1.18)	-0.0376 (-1.31)	-0.111** (-3.10)	-0.104*** (-3.59)	0.0757 (1.33)	0.102 (1.38)	0.0229 (0.87)	0.0337 (1.75)	0.150*** (3.38)	0.138** (2.96)
Inactivity	-0.193 (-1.03)	-0.241 (-1.58)	-0.290*** (-4.98)	-0.426*** (-5.02)	-0.835** (-2.65)	-0.890** (-3.12)	-0.234*** (-4.49)	-0.140* (-2.38)	0.113 (0.79)	0.333*** (3.53)
N	2657	2657	22221	22221	332	332	22233	22233	390	390

There are two different regressions for each variable: on the left the unconstrained one, while on the right the one controlled for covariates.

\* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$ .

**Table 4:** Relation between Risk occurrence and Risky behaviours with fixed-effect (LPM regression).

	Dead		Alive		Nursing Home		Medigap Exp		Hospital	
	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Smoking	-0.00379 (-1.75)	0.00536* (2.65)	-0.0232 (-1.60)	-0.0344* (-2.33)	-0.00285* (-2.76)	-0.00141 (-1.21)	-67.52 (-1.66)	-20.02 (-0.72)	-0.0205** (-4.08)	-0.0200** (-4.18)
Drinking	0.00629 (1.29)	-0.00132 (-0.25)	0.0212 (1.31)	0.0279 (1.46)	0.000407 (0.31)	-0.000893 (-0.48)	-70.66 (-1.28)	-76.88 (-1.32)	0.00319 (0.31)	0.00362 (0.36)
BMI	-0.00290 (-0.84)	-0.00383 (-0.95)	0.0248*** (8.55)	0.0235*** (6.04)	-0.00106 (-0.57)	-0.00121 (-0.70)	-82.38 (-1.01)	-71.16 (-0.91)	0.00858 (1.05)	0.00831 (0.97)
Preventive	0.00268 (0.74)	0.00300 (1.21)	0.00378 (0.42)	0.00498 (0.65)	-0.000324 (-0.21)	-0.000322 (-0.23)	-4.803 (-0.38)	-15.51 (-1.11)	0.0248*** (9.41)	0.0247*** (9.56)
Inactivity	0.0780*** (11.91)	0.0639*** (9.88)	-0.107*** (-5.18)	-0.0757** (-3.37)	0.0183 (1.97)	0.0164 (2.06)	201.2* (3.03)	141.7 (2.12)	0.173*** (13.38)	0.172*** (12.98)
N	22233	22233	22233	22233	15040	15040	22233	22233	22233	22226

There are two different regressions for each variable: on the left the unconstrained one, while on the right the one controlled for covariates.  
 \* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$ .

## Acknowledgements

The author thank Giuseppe Ragusa and Alberto Cybo Ottone for the help received in framing the work.

This paper uses data from SHARE Waves 1 and 2 release 2.6.0, as of November 29th 2013 (DOI: <https://doi.org/10.6103/SHARE.w1.260> and DOI: <https://doi.org/10.6103/SHARE.w2.260>). The SHARE data collection has been primarily funded by the European Commission through the 5th Framework Programme (project QLK6-CT-2001-00360 in the thematic programme Quality of Life), through the 6th Framework Programme (projects SHARE-I3, RII-CT-2006-062193, COM-PARE, CIT5-CT-2005-028857, and SHARELIFE, CIT4-CT-2006-028812) and through the 7th Framework Programme (SHARE-PREP, N. 211909, SHARE-LEAP, N. 227822 and SHARE M4, N. 261982). Additional funding from the U.S. National Institute on Aging (U01 AG09740-13S2, P01 AG005842, P01 AG08291, P30 AG12815, R21 AG025169, Y1-AG-4553-01, IAG BSR06-11 and OGHA 04-064) and the German Ministry of Education and Research as well as from various national sources is gratefully acknowledged. For more information about the survey, data collection or preliminary analysis check Borsch-Supan et al. (2005; 2008; 2013) and Borsch-Supan and Jürges (2005).

## Competing Interests

The author has no competing interests to declare.

## References


- Akerlof, G A** 1970 The Market for Lemons: Quality Uncertainty and the Market Mechanism. *The Quarterly Journal of Economics*, 84(3): 488–500. DOI: <https://doi.org/10.2307/1879431>
- Borsch-Supan, A, Brandt, M, Hunkler, C, Kneip, T, Korbmacher, J, Malter, F, Schaan, B, Stuck, S and Zuber, S** 2013 “Data Resource Profile: The Survey of Health, Ageing and Retirement in Europe (SHARE)”. *International Journal of Epidemiology*. DOI: <https://doi.org/10.1093/ije/dyt088>
- Borsch-Supan, A, Brugiavini, A, Jürges, H, Kapteyn, A, Mackenbach, J, Siegrist, J and Weber, G** 2008 “First results from the Survey of Health, Ageing and Retirement in Europe (2004–2007). Starting the longitudinal dimension”. *Mannheim: Mannheim Research Institute for the Economics of Aging (MEA)*.
- Borsch-Supan, A, Brugiavini, A, Jürges, H, Mackenbach, J, Siegrist, J and Weber, G** 2005 “Health, ageing and retirement in Europe – First results from the Survey of Health, Ageing and Retirement in Europe”. *Mannheim: Mannheim Research Institute for the Economics of Aging (MEA)*.
- Borsch-Supan, A and Jürges, H** (Eds.). 2005 “The Survey of Health, Ageing and Retirement in Europe-Methodology”. *Mannheim: Mannheim Research Institute for the Economics of Aging (MEA)*.
- Bryan, M L and Jenkins, S P** 2013 Regression Analysis of Country Effects Using Multilevel Data: A Cautionary Tale. IZA Discussion Paper Series No. 7583.
- Cardon, J and Hendel, I** 2001 Asymmetric Information in Health Insurance: Evidence from the National Medical Expenditure Survey. *RAND Journal of Economics*, 32: 408–427. DOI: <https://doi.org/10.2307/2696362>
- Cawley, J and Philipson, T** 1999 An empirical examination of information barriers to trade in insurance. *American Economic Review*, 89(4): 827–846. DOI: <https://doi.org/10.1257/aer.89.4.827>
- Chiappori, P A, Jullien, B, Salanie, B and Salanie, F** 2006 Asymmetric Information in Insurance: General Testable Implications. *Rand Journal of Economics*, 37(4): 783–98. DOI: <https://doi.org/10.1111/j.1756-2171.2006.tb00057.x>
- Chiappori, P A and Salanie, B** 2000 Testing for Asymmetric Information in Insurance Markets. *Journal of Political Economy*, 108(1): 56–78. DOI: <https://doi.org/10.1086/262111>
- Cohen, A and Einav, L** 2007 Estimating risk preferences from deductible choice. *American Economic Review*, 97(3): 745–788. DOI: <https://doi.org/10.1257/aer.97.3.745>
- Cutler, D M, Finkelstein, A and McGarry, K** 2008 Preference Heterogeneity and Insurance Markets: Explaining a Puzzle of Insurance. *American Economic Review: Papers & Proceedings*, 98(2): 157–162. DOI: <https://doi.org/10.1257/aer.98.2.157>
- Cutler, D M and Zeckhauser, R** 2000 The Anatomy of Health Insurance. Culyer, A and Newhouse, J (ed.). In: *Handbook of Health Economics*, 1A: 563–643. Amsterdam: Elsevier. DOI: [https://doi.org/10.1016/S1574-0064\(00\)80170-5](https://doi.org/10.1016/S1574-0064(00)80170-5)
- De Meza, D and Webb, D C** 2001 Advantageous selection in insurance markets. *RAND Journal of Economics*, 32(2): 249–262. DOI: <https://doi.org/10.2307/2696408>

- Dionne, G, Gouriéroux, C and Vanasse, C** 2001 Testing for Evidence of Adverse Selection in the Automobile Insurance Market: A Comment. *Journal of Political Economy*, 109(2): 444–453. DOI: <https://doi.org/10.1086/319557>
- Einav, L, Finkelstein, A and Levin, J** 2010 Beyond Testing: Empirical Models of Insurance Markets. *Annual Review of Economics*, 2: 311–336. DOI: <https://doi.org/10.1146/annurev.economics.050708.143254>
- Einav, L, Finkelstein, A and Levin, J** 2011 Selection in Insurance Markets: Theory and Empirics in Pictures. *Journal of Economic Perspectives*, 25(1): 115–138. DOI: <https://doi.org/10.1257/jep.25.1.115>
- Einav, L, Finkelstein, A and Schrimpf, P** 2007 The Welfare Cost of Asymmetric Information: Evidence from the U.K. Annuity Market. NBER Working Paper No. 13228.
- Ettner, S** 1997 Adverse Selection and the Purchase of Medigap Insurance by the Elderly. *Journal of Health Economics*, 16(5): 499–624. DOI: [https://doi.org/10.1016/S0167-6296\(97\)00011-8](https://doi.org/10.1016/S0167-6296(97)00011-8)
- Fang, H, Keane, M and Silverman, D** 2006 Sources of Advantageous Selection: Evidence from the Medigap Insurance Market. NBER Working Paper No. 12289.
- Finkelstein, A and McGarry, K** 2006a Multiple dimensions of private information: evidence from the long-term care insurance market. *American Economic Review*, 96(4): 938–958. DOI: <https://doi.org/10.1257/aer.96.4.938>
- Finkelstein, A and McGarry, K** 2006b Private Information and its Effect on Market Equilibrium: New Evidence from Long-Term Care Insurance. *American Economic Review*, 96(4): 938–58. DOI: <https://doi.org/10.1257/aer.96.4.938>
- Finkelstein, A and Poterba, J** 2002 Selection Effects in the Market for Individual Annuities: New Evidence from the United Kingdom. *Economic Journal*, 112(476): 28–50. DOI: <https://doi.org/10.1111/1468-0297.0j672>
- Finkelstein, A and Poterba, J** 2004 Adverse Selection in Insurance Markets: Policyholder Evidence from the U.K. Annuity Market. *Journal of Political Economy*, 112(1): 183–208. DOI: <https://doi.org/10.1086/379936>
- Finkelstein, A and Poterba, J** 2006 Testing for Asymmetric Information Using ‘Unused Observables’ in Insurance Markets: Evidence from the U.K. Annuity Market. NBER Working Paper No. 12112.
- Hemenway, D** 1990 Propitious selection. *Quarterly Journal of Economics*, 105(4): 1063–1069. DOI: <https://doi.org/10.2307/2937886>
- Hurd, M and McGarry, K** 1997 Medical Insurance and the Use of Health Care Services by the Elderly. *Journal of Health Economics*, 16(2): 129–154. DOI: [https://doi.org/10.1016/S0167-6296\(96\)00515-2](https://doi.org/10.1016/S0167-6296(96)00515-2)
- McCarthy, D and Mitchell, O S** 2003 International Adverse Selection in Life Insurance and Annuities. NBER Working Paper No. 9975.
- Mitchell, O S, Poterba, J M, Warshawsky, M J and Brown, J R** 1999 New Evidence on the Money’s Worth of Individual Annuities. *American Economic Review*, 89: 1299–1318. DOI: <https://doi.org/10.1257/aer.89.5.1299>
- Rothschild, M and Stiglitz, J** 1976 Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information. *Quarterly Journal of Economics*, 90: 630–649. DOI: <https://doi.org/10.2307/1885326>

**How to cite this article:** Corea, F 2017 Big Data and Insurance: Advantageous Selection in European Markets. *Data Science Journal*, 16: 33, pp. 1–15, DOI: <https://doi.org/10.5334/dsj-2017-033>

**Submitted:** 07 November 2016    **Accepted:** 09 June 2017    **Published:** 23 June 2017

**Copyright:** © 2017 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

 *Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

**OPEN ACCESS** 