Industrial Organization II: Markets with Asymmetric Information (SIO13)

Overview

- Will try to get people familiar with recent work on markets with asymmetric information; mostly insurance market, but may talk a little bit about credit markets if time permits.

- The main emphasis of the course is empirical and applied.
    - Will talk about theory, but not much, and will occasionally digress to talk about econometric methods, but not much. Main focus will be on the economics.
    - Some background in econometrics and IO can come handy, but I'm hoping that the class would be penetrable for everyone that has basic graduate-level background. Don't hesitate to stop me if you have no idea what I say ...

- Why this topic?
    - I have something to say.
    - Covers some of the largest and most important markets, and many interesting policy questions, so (good!) research can make a big difference.
    - Amazing data (large, rich, and high quality).

Main topics

- (Quick) Theory background (~0.5 class)
- General framework and "reduced form" tests for asymmetric information (~0.75)
- Empirical models of demand for insurance (~1.5)
- Welfare analysis in the context of asymmetric information (~0.75)
- Pricing and other topics (depends on time left)

Logistics

- Four meetings: Mon 12, Tue 11, Wed 11, Thu 9.
- Each meeting will have three segments of 75 min each (discuss timing preferences!).
- Not enough time for assignments, but try to read as much as you can from one class to another.
- Participation is key: questions would slow me down (which is good!), and discussion is important. I would be totally happy to not finish covering all the material I prepared.
- Grading based on attendance, participation, and a take-home final.
- I'm around this week and generally available before/after class, or during the breaks.

Must-know background: seminal theories of insurance markets

Arrow (1963)

- Pretty interesting to read:
  - Amazing to see how writing style in economics has changed over 50 years.
  - Amazing to realize how the core points about healthcare remained the same.

- Some basic points/assertions:
  - Individuals are risk averse w/ vNM utility function $u(w)$. (Agree?)
  - Insurer is diversified and thus risk neutral. (Agree?)
  - Thus, more insurance or full insurance should always be more efficient.

- Model 1 (also common in many textbooks):
  - Individual has income E, and faces a possible loss X with probability q.
  - Insurance contracts are simple/linear:
    - Coverage costs p for each unit of coverage.
    - Coverage pays 1 for each unit of coverage, in case of a loss.
    - Individual chooses number of units of coverage D.

- Individual's problem:

$$Max_D\ (1-q)u(E-pD)+qu(E-pD-X+D)$$

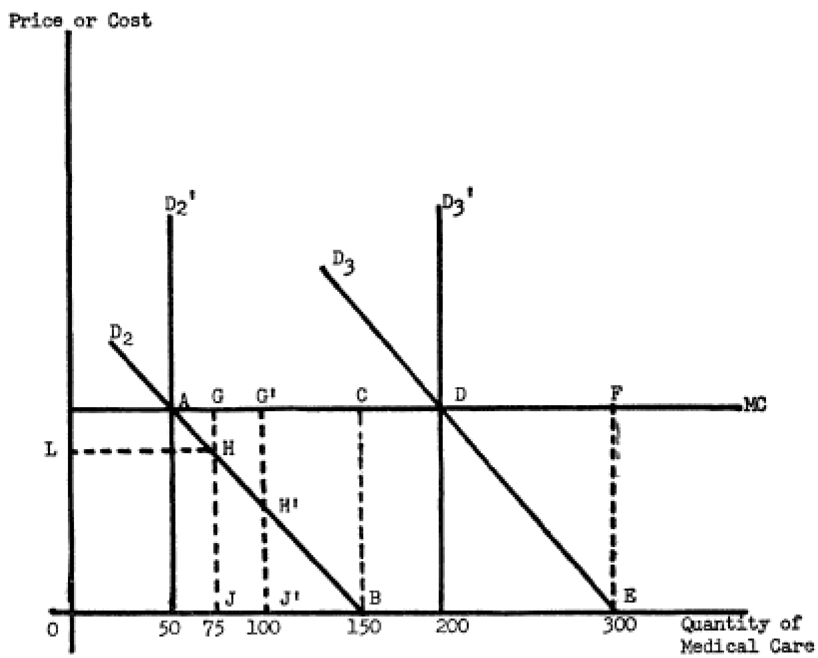
  - This now leads to simple demand derivation $D(p)$.
  - Easy to show that:
    - If $p=q$ (insurance is actuarially fair), individual would choose full coverage, $D=X$.
    - If $p > q$ (insurance is actuarially not fair, e.g. because administrative costs), $D < X$.
    - Furthermore, demand for insurance would obey intuitive comparative statics: all else equal, individuals would choose more coverage if they are more risk averse (in terms of u) or more risky (higher q).
    - The latter case may lead to adverse selection, which we will talk about later.

- Model 2:
  - Individual has income E, and faces a loss distribution Y.
  - Contract costs p and pays $C(Y) \geq 0$.
  - Question: for all contracts with the same expected revenue $E(Y-C(Y))$, which one would the individual prefer the most?
  - Answer: a deductible contract. That is, individual pays the loss up to a deductible d, above which insurer covers everything.
  - Intuition: suppose not, then we can make a risk averse individual better off by smoothing out his risk.

- Other rationales for a deductible contract?

Pauly (1968) (in a response to Arrow)

- Basic point: if moral hazard is present, full insurance would not be optimal.
- Example: Individual is either healthy (prob 0.5), mildly sick (0.25), or very sick (0.25). Consider two situations:



Case 1: Vertical demand curves (no moral hazard):

- Full insurance (priced at 62.5, so zero profit) would be an efficient equilibrium.

Case 2: Downward sloping demand curves ("moral hazard", although bad terminology):

- Full insurance would mean expected costs of 112.5, but if it's so expensive individuals may prefer the risk exposure over the insurance.

General point: efficient coverage will trade off risk exposure with incentives:

- Full insurance would often be suboptimal because it would provide bad incentives.
- No insurance would often be suboptimal because it would expose individuals to too much risk.
- Efficient partial coverage should strike the right balance between the two forces.

Models of adverse selection

Akerlof (1970)

- Presumably familiar to everyone, so we will be brief and adapt to insurance setting as in Einav, Finkelstein, and Cullen (2012) (which we will cover later this week).

- Model illustrates the inefficiency that may arise from asymmetric information.
  - Original model used used-car markets.
  - Many other applications, including labor markets, credit markets, insurance.

- Population of individuals, whose types are given by their willingness to pay for insurance and their expected costs to the insurer (if insured). We also assume perfect competition.

- What would be the equilibrium if types are observed?

- The key is that types are private information, so there is only one price that can be offered to all individuals.
  - Important! No other way to screen people (e.g., by other contract dimensions)

- What would then be the competitive equilibrium? How would it compare to the (efficient) case of observed types?

Graphs:

- Can also graph special cases:
    - Efficiency despite the asymmetric information.
    - Complete unraveling.

- What if individuals vary in other dimensions (e.g. risk aversion)?

Rothschild and Stiglitz (1976)

- Key idea: if insurers can screen people on more than just price, may get some separation.

- Well known R-S graphs:

- Issues:
    - Non-existence when there is a small fraction of high-risk types (see graph). Wilson (1977) and Miyazaki (1977) show that this can be solved if one could change the equilibrium definition (essentially adding implicit dynamics and allowing money-losing contracts to get dropped). Their equilibrium often involves cross subsidization.
    - Very difficult to extend the model for environments with richer heterogeneity.
    - In practice, somewhat surprisingly, many markets converge on standard set of contracts and then clear by price.

Empirical Setting and "Reduced Form" Tests

General framework:

- Consumer with characteristics ζ is offered contract *(ϕ,p)*. Valuation given by:

$$v(\phi, p, \zeta) = \max_{a \in A} \sum_{s \in S} \pi(s \mid a, \zeta) u(s, a, \zeta, \phi, p)$$

  Can also define $a^*(\zeta,\phi,p)$ as the optimal behavior and $\pi^*(\cdot|\zeta,\phi,p)$ the resulting probabilities.

- Insurer's cost:

$$c(\phi, \zeta) = \sum_{s \in S} \pi^*(s \mid \phi, \zeta) \tau(s, \phi),$$

- Consumer choice:

$$v(\phi_j, p_j, \zeta) \geq v(\phi_k, p_k, \zeta) \text{ for all } k \in J.$$

  - Should look familiar for those who took empirical IO

- Adverse selection:

$$\mathbb{E}_\zeta[c(\phi_j, \zeta_i) \mid i \in I(j)] > \mathbb{E}_\zeta[c(\phi_j, \zeta_i) \mid i \in I]$$

  - Important to note that this definition makes it depend on entire sets of contracts
  - Definition depends on actual cost, not on why costs are higher or on whether it could have been higher (will return to it later)

- We will now go over various papers that test for adverse selection. They essentially ask whether

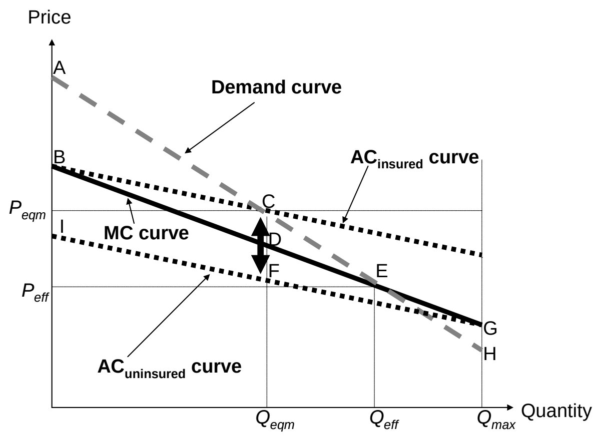
$$\mathbb{E}[y_i \mid i \in I(j), x_i = x] > \mathbb{E}[y_i \mid i \in I(k), x_i = x]$$

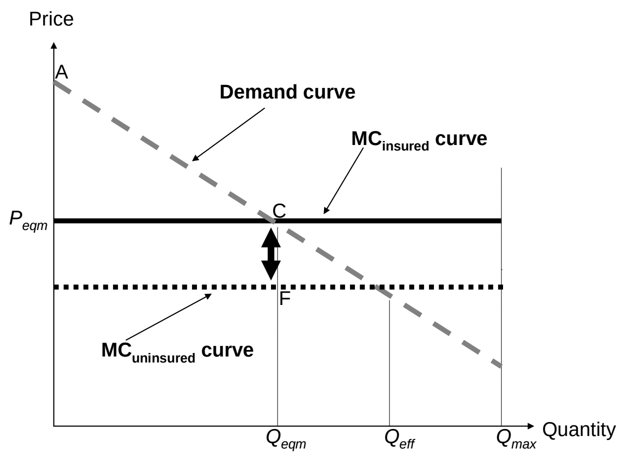
  But it's important to think about the x's:
  - Priced
  - Observed but not priced (examples)
  - Could become observed but are not

- Another important issue that will come up is adverse selection vs. moral hazard:
  - o Adverse selection: high risk types select high coverage contracts
  - o Moral hazard: high coverage causes people to be high risk

  Problem! Both lead to positive correlation between risk and coverage.

- See also graphically:





"Evidence on Adverse Selection: Equilibrium Signaling and Cross-Subsidization in the Insurance Market" by Puelz and Snow (*JPE*, 1994)

Goals:

- Test for adverse selection: are people who chose more coverage (lower deductible) more risky? yes!
- Test for linearity of the pricing schedule: no! concave.
- Test for cross-subsidization: are policies breaking even, type by type? yes (but mushy ...).

Data:

- 3,280 individuals who purchased collision insurance in Georgia in 1986 from a particular company.
- For each individual observe a bunch of stuff, and the price and the deductible level (100, 200, or 250), and whether he/she had an accident during the covered period.
- They don't have: the menu, good proxy for wealth (what do they use?)

What do they do? Estimate a linear hedonic price regression and an ordered logit deductible choice. See tables.

TABLE 1

Summary Statistics for Variables Used in the Model

	Deductible		
	$100	$200	$250
Number of exposures	1,385	703	1,192
Number incurring a loss	42	16	19
Mean collision premium	$122.03	$107.01	$97.17
	(36.26)	(31.57)	(27.00)
Mean wealth (000s)	$106.73	$145.99	$164.05
	(106.20)	(125.57)	(135.02)
Mean age of automobile	3.98	3.55	3.42
	(1.90)	(1.86)	(1.93)
Mean age of individual driver	46.48	42.56	43.17
	(12.48)	(12.01)	(11.79)
Proportion with a multirisk contract	.8779	.8876	.8993

TABLE 2

ESTIMATES OF THE EMPIRICAL MODEL

Dependent Variable: P Estimated Equation (6)			Dependent Variable: D Estimated Equation (7)		
Variable	Coefficient	t-Ratio	Variable	Coefficient	t-Ratio
ONE	259.132	150.01	ONE	.702110	3.082
D_1	−40.1243	−19.034	RT	−.590533	−2.573
D_2	−57.6179	−31.22	$\hat{g}_D$	−.036873	−5.877
A	−10.8933	−61.08	W_1	−1.18118	−8.857
$A \cdot D_1$	2.74736	8.815	W_2	−.503774	−5.082
$A \cdot D_2$	3.3392	13.03	W_3	−.260122	−2.912
MR	−26.3947	−41.43	MALE	.0875617	1.287
SYM_7	−58.5948	−40.965	PERAGE	−.014027	−4.738
SYM_8	−43.8493	−33.084	μ	.950285	28.966
SYM_9	−27.6492	−20.284	Log L	−3,317.525	
SYM_{10}	−12.8008	−8.727	Restricted log L*	−3,483.414	
T_{11}	−19.5718	−9.588			
T_{12}	−25.0184	−14.422			
T_{13}	−15.6466	−11.473			
T_{14}	−20.8188	−28.603			
$SYM_7 \cdot D_1$	5.93706	2.564			
$SYM_8 \cdot D_1$	4.47641	2.246			
$SYM_9 \cdot D_1$	2.4945	1.216			
$SYM_{10} \cdot D_1$	−3.13413	−1.428			
$SYM_7 \cdot D_2$	11.3593	5.698			
$SYM_8 \cdot D_2$	6.99569	3.882			
$SYM_9 \cdot D_2$	4.58894	2.493			
$SYM_{10} \cdot D_2$	−1.52623	−.778			
$T_{11} \cdot D_1$	−.511167	−.169			
$T_{12} \cdot D_1$	.543705	.218			
$T_{13} \cdot D_1$	−1.35770	−.628			
$T_{14} \cdot D_1$	2.57983	2.065			
$T_{11} \cdot D_2$	1.37454	.547			
$T_{12} \cdot D_2$	2.84681	1.302			
$T_{13} \cdot D_2$	3.66697	2.163			
$T_{14} \cdot D_2$	6.14255	5.734			
MALE	.780244	1.909			
PERAGE	−.401531	−24.193			
Adjusted R^2	.892886				

* Slopes = 0.

TABLE 3

EFFECT OF RISK TYPE ON PREDICTED PROBABILITIES

	Prob[D = 0]	Prob[D = 1]	Prob[D = 2]
$RT = 1$	.56592	.20535	.22873
$RT = 0$	.41938	.23197	.34865
Δ in probability	+.14654	−.02662	−.11992
	(.63618)	(−.11556)	(−.52062)

NOTE.—*t*-statistics are in parentheses.

TABLE 4

ESTIMATED LOSS PROBABILITIES

	Risk Type		
	High	Medium	Low
π_τ	.030324	.022756	.015939

TEST FOR DIFFERENCES IN LOSS PROBABILITIES

$\pi_H - \pi_L$	.014334
	(2.09)
$\pi_M - \pi_L$	.007568
	(1.15)
$\pi_H - \pi_M$	.007484
	(1.01)

NOTE.—*t*-statistics are in parentheses.

TABLE 5

Selected Estimated Cross-Subsidization Charges

	k_{3H}	k_{3M}	k_{3L}
		Males	
SYM_7, T_{11}	−6.12	−3.73	9.31
	(−.263)	(−.193)	(.42)
SYM_8, T_{12}	−7.18	−2.89	10.05
	(−.291)	(−.142)	(.425)
SYM_9, T_{13}	−11.97	−4.37	16.49
	(−.412)	(−.185)	(.584)
SYM_{10}, T_{14}	−12.52	−4.17	17.01
	(−.409)	(−.169)	(.57)
		Females	
SYM_7, T_{11}	−5.93	−3.75	9.10
	(−.257)	(−.195)	(.413)
SYM_8, T_{12}	−6.98	−2.91	9.83
	(−.285)	(−.144)	(.419)
SYM_9, T_{13}	−11.78	−4.39	16.28
	(−.407)	(−.187)	(.579)
SYM_{10}, T_{14}	−12.33	−4.19	16.80
	(−.405)	(−.17)	(.566)

NOTE.—*t*-statistics are in parentheses.

"Testing for Asymmetric Information in Insurance Markets," by Chiappori and Salanie (*JPE*, 2000)

Main points:

- Argue that many predictions of the theory are sensitive to various assumptions. For example, the cross-subsidization test of Puelz and Snow required their supply-side model to be correct.
- They look for robust predictions. In particular, the positive correlation between higher coverage and more accidents does not depend on pricing policy, other dimensions of heterogeneity, etc. (as we will see soon, this is only true one-sidedly).
- Can we generate this positive correlation without adverse selection? yes, through moral hazard. Identifying between the two is harder, and requires exogenous change in coverage.
- The distinction between adverse selection and moral hazard may be crucial for counterfactuals policy simulations.

Econometrics:

- To make a convincing case for adverse selection (or moral hazard), we need to worry a lot about:
    - Flexible functional form: coverage choice may be highly non-linear
    - Endogeneity of the menu (omitted variable bias): the insurer may base pricing decisions on stuff unobserved to the econometrician (e.g. past driving record).

Data:

- Sample of about 5% of all auto insurance contracts in France in 1989.
- Rich data of 1,120,000 contracts (and 120,000 claims).
- Focus on young drivers (up to 3 years license years): 20,716 contracts (2% of the data!). Can think about this as a way to control for stuff: this is a more homogeneous group and has no past experience (i.e. no omitted variables probably). Why this may be a bad choice of a sample?
- Throw out one-car accidents to avoid (or reduce) ex-post endogeneity of claims.
- Key dependent variables:
    - y_i - 1 if i bought more coverage (comprehensive insurance; not only the mandatory third-party insurance)
    - z_i - 1 if i was involved in at least one at fault accident

Tests for adverse selection

- Parametric tests:
    - Run separately two probits of y_i and z_i on all observables, and test for correlation in the error terms.
    - Run a bivariate probit of the two. This allows to get a confidence interval for the correlation coefficient ρ.

- Nonparametric test:
    - Split the data into discrete bins (suppose all are dummy variables) and test independence in each bin separately. Loosely speaking, we ask whether $Pr(y_i=1|z_i=0)=N_{01}/N_{00}$ is close to $Pr(y_i=1|z_i=1)=N_{11}/N_{10}$.
    - With 2^m different bins, we have 2^m test statistics. We can aggregate them in different ways to come up with various test statistics (e.g. the number of rejections).
    - Note that the number of bins is restricted by the number of data points.

Results

- Beginners (up to 1 year of license):
  - Test 1: accept conditional independence (low test statistic)
  - Test 2: ρ=*-0.029 (0.049)*.
  - Non-parametric tests: accept the null as well.
- Main issue: perhaps beginners have to learn about their own type before they know it. (see Cohen, 2005)
- More seniors (3 years of license: is this enough to know one's type?): similar results.
- Use some strange anomaly in the French system to test for moral hazard. Do not find evidence for it. (we'll talk later about other ways to test)
- Missing: would we have different results with less controls or more parametric tests?

“Adverse Selection in Insurance Markets: Policyholder Evidence from the UK Annuity Market,” by Finkelstein and Poterba (*JPE*, 2004)

- Test for adverse selection in annuity markets.
- Main point: do not find adverse selection in coverage, but find adverse selection on other dimensions of the contract.
- Annuities: pay annual amounts until one dies. High risk is a longer lived individual.
- Parameters: the NPV of the payments, the level of backloading, and guaranteed payments.
- Data: 42,054 annuity contracts in the UK.
- Use simple hazard model to estimate.
- Find strong adverse selection on other dimensions, little correlation in initial payment. See Table 2.
- Nice features: moral hazard is unlikely to be a problem, so we can attribute results to adverse selection. We also get data about risk-types of non-buyers, so we can check for adverse selection both on the intensive and extensive margins.

TABLE 2
SELECTION EFFECTS AND ANNUITY PRODUCT CHARACTERISTICS

Explanatory Variable	Estimates from Hazard Model of Mortality after Purchasing an Annuity: Compulsory Market (1)	Estimates from Hazard Model of Mortality after Purchasing an Annuity: Voluntary Market (2)	Estimates from Linear Probability Model of Probability of Dying within Five Years: Compulsory Market (3)	Estimates from Linear Probability Model of Probability of Dying within Five Years: Voluntary Market (4)
Index-linked	−.839***	−.894**	−.053***	−.185***
	(.217)	(.358)	(.019)	(.050)
Escalating	−1.085***	−1.497***	−.072***	−.152***
	(.113)	(.253)	(.010)	(.030)
Guaranteed	.019	.216***	.007*	.046***
	(.029)	(.060)	(.004)	(.016)
Capital-protected	...	.056	...	.064***
		(.051)		(.016)
Payment (£100s)	−.003***	.001**	−.0003***	.0003***
	(.0006)	(.0004)	(.0001)	(.0001)
Male Annuitant	.640***	.252***	.044***	.044***
	(.039)	(.051)	(.005)	(.014)
Observations	38,362	3,692	24,481	3,575
Number of deaths in sample	6,311	1,944	2,693	822

NOTE.—Cols. 1 and 2 report estimates from Han-Hausman discrete-time, semiparametric proportional hazard models on the full sample. These are estimated using 17 annual discrete time intervals. Baseline hazard parameters are not reported. Cols. 3 and 4 report estimates from a linear probability model of the probability of dying within five years of purchase; these models are estimates on the sample of individuals who purchased their annuity in 1993 or earlier, so that all observations are uncensored. All regressions include, in addition to the covariates shown above, indicator variables for five-year intervals for age at purchase, indicator variables for year of purchase, and indicator variables for the frequency of payments. Standard errors are in parentheses. The omitted category for the "back-loaded" dummies (index-linked and escalating) is nominal annuities. The omitted category for the guarantee feature dummies (guaranteed and capital-protected) is not guaranteed and not capital-protected.

* Statistically significant at the 10 percent level.

** Statistically significant at the 5 percent level.

*** Statistically significant at the 1 percent level.

"Private Information and its effect on market equilibrium: New evidence from long-term care insurance," by Finkelstein and McGarry (*AER*, 2006)

- Key idea: the lack of positive correlation may be driven by either symmetric information or by two sources of private information which offset each other.
- Why do we care? if this is the former, we should have efficient market. If this is the latter, we will have inefficient pooling in equilibrium, and perhaps we can try to fix it.
- Application: long-term care insurance.
- Note: Little theory. Much effort on data collection, and combining supporting evidence for their story from different (but related) data sources.

Data and results

- First result: use insurer-data to run similar hazard regression to that of Finkelstein-Poterba. Tables 3 and 4. Here they find little evidence of positive correlation (if anything, they find negative correlation) between risk types and policy choices.
- Use a separate source of information (AHEAD survey) that elicits beliefs and preferences. Key point is that insurance companies do not observe this information:
    - Table 1: beliefs help to predict outcome, i.e. private information.

- Table 2: this private information in risk is also translated to coverage choice (this is important to establish this link).
- But the above suggests that there must be some omitted variable that rationalize the difference, i.e. how can else we explain that z (beliefs) affects both x (care) and y (insurance choice) but x and y are uncorrelated? This something else is risk aversion or other preferences - Table 5: proxies for risk aversion (namely, preventive measures taken) can do the trick.

TABLE 3—THE RELATIONSHIP BETWEEN LONG-TERM CARE INSURANCE AND NURSING HOME ENTRY

	No controls (1)	Controls for insurance company prediction (2)	Controls for application information (3)
Correlation coefficient from bivariate probit of LTCINS and CARE	−0.105***	−0.047	−0.028
	(p = 0.006)	(p = 0.25)	(p = 0.51)
Coefficient from probit of CARE on LTCINS	−0.046***	−0.021	−0.014
	(0.015)	(0.016)	(0.016)
N	5,072	5,072	4,780

Notes: Top row reports the correlation of the residual from estimation of a bivariate probit of any nursing home use (1995–2000) and long-term care insurance coverage (1995); p values are given in parentheses. Bottom row reports marginal effect on indicator variable for long-term care insurance in 1995 from probit estimation of equation (3). The dependent variable is an indicator variable for any nursing home use from 1995 through 2000; heteroskedacticity-adjusted robust standard errors are in parentheses. For all rows, control variables are described in column headings; see text for more information. ***, **, * denote statistical significance at the 1-percent, 5-percent, and 10-percent level, respectively. Means of CARE and LTCINS are 0.16 and 0.11, respectively.

TABLE 4—RELATIONSHIP BETWEEN LTCINS AND CARE
(*Sample restricted to individuals with same choice set*)

	No controls (1)	Controls for insurance company prediction (2)	Controls for application information (3)
Correlation coefficient from bivariate probit of LTCINS and CARE	−0.123* (p = 0.08)	−0.122* (p = 0.10)	−0.191** (p = 0.017)
Coefficient from regression of CARE on LTCINS	−0.032* (0.018)	−0.028* (0.015)	−0.033** (0.012)
N	1,504	1,504	1,438

Notes: Sample is limited to individuals in the top quartile of the wealth and income distribution and who have none of the health characteristics that might make them ineligible for private insurance. Top row reports the correlation of the residual from estimation of a bivariate probit of any nursing home use (1995–2000) and long-term care insurance coverage (1995); p values are given in parentheses. Bottom row reports marginal effect on indicator variable for long-term care insurance in 1995 from probit estimation in equation (3). The dependent variable is an indicator variable for any nursing home use from 1995 through 2000; heteroskedacticity-adjusted robust standard errors are in parentheses. For all rows, control variables are described in column headings; see text for more information. ***, **, * denote statistical significance at the 1-percent, 5-percent, and 10-percent level, respectively. Means of CARE and LTCINS are 0.09 and 0.17, respectively.

TABLE 1—RELATIONSHIP BETWEEN INDIVIDUAL BELIEFS AND SUBSEQUENT NURSING HOME USE

	No controls	Control for insurance company prediction		Control for application information
	(1)	(2)	(3)	(4)
Individual prediction	0.091***		0.043**	0.037*
	(0.021)		(0.020)	(0.019)
Insurance company prediction		0.400***	0.395***	
		(0.020)	(0.021)	
pseudo-R^2	0.005	0.097	0.099	0.183
N	5,072	5,072	5,072	4,780

Notes: Reported coefficients are marginal effects from probit estimation of equation (1). Dependent variable is an indicator for any nursing home use from 1995 through 2000 (mean is 0.16). Both individual and insurance company predictions are measured in 1995. Heteroskedacticity-adjusted robust standard errors are in parentheses. ***, **, * denote statistical significance at the 1-percent, 5-percent, and 10-percent level, respectively. Column 4—which includes controls for "application information"—includes controls for age (in single year dummies), sex, marital status, age of spouse, over-35 health indicators, and a complete set of two-way and three-way interactions for all of the variables used in the insurance company prediction (age dummies, sex, limitations to activities of daily living, limitations to instrumental activities of daily living, and cognitive impairment); see text for more details.

TABLE 2—RELATIONSHIP BETWEEN INDIVIDUAL BELIEFS AND INSURANCE COVERAGE

	No controls	Control for insurance company prediction		Control for application information
	(1)	(2)	(3)	(4)
Individual prediction	0.086***		0.099***	0.083***
	(0.017)		(0.017)	(0.016)
Insurance company prediction		−0.125***	−0.140***	
		(0.023)	(0.023)	
pseudo-R^2	0.007	0.010	0.019	0.079
N	5,072	5,072	5,072	4,780

Notes: Reported coefficients are marginal effects from probit estimation of equation (2). Dependent variable is an indicator for whether individual has long-term care insurance coverage in 1995 (mean is 0.11). Both individual and insurance company predictions are measured in 1995. Heteroskedacticity-adjusted robust standard errors are in parentheses. ***, **, * denote statistical significance at the 1-percent, 5-percent, and 10-percent level, respectively. Column 4—which includes controls for "application information"—includes controls for age (in single year dummies), sex, marital status, age of spouse, over-35 health indicators, and a complete set of two-way and three-way interactions for all of the variables used in the insurance company prediction (age dummies, sex, limitations to activities of daily living, limitations to instrumental activities of daily living, and cognitive impairment); see text for more details.

TABLE 5—PREFERENCE-BASED SELECTION

	No controls		Control for insurance company prediction		Control for application information	
	NH Entry (1)	LTC Insurance (2)	NH Entry (3)	LTC Insurance (4)	NH Entry (5)	LTC Insurance (6)
Panel A: Wealth						
Top wealth quartile	−0.095***	0.150***	−0.038**	0.131***	−0.018	0.139***
	(0.013)	(0.020)	(0.014)	(0.020)	(0.015)	(0.022)
Wealth quartile 2	−0.073***	0.104***	−0.025*	0.089***	−0.013	0.092***
	(0.013)	(0.020)	(0.014)	(0.020)	(0.014)	(0.020)
Wealth quartile 3	−0.030**	0.062***	0.0004	0.052***	0.006	0.057***
	(0.015)	(0.020)	(0.016)	(0.019)	(0.015)	(0.020)
Bottom wealth quartile (omitted)	—	—	—	—	—	—
Individual prediction	0.086***	0.089***	0.042**	0.098***	0.035*	0.086***
	(0.021)	(0.017)	(0.020)	(0.017)	(0.019)	(0.017)
Panel B: Preventive health activity						
Preventive activity	−0.106***	0.066***	−0.054***	0.052***	−0.016	0.016
	(0.0118)	(0.017)	(0.018)	(0.017)	(0.019)	(0.017)
Individual prediction	0.095***	0.082***	0.047**	0.095***	0.037*	0.082***
	(0.021)	(0.017)	(0.020)	(0.017)	(0.020)	(0.017)
Panel C: Seat belt use						
Always wear seatbelt	−0.059***	0.053***	−0.031**	0.048***	−0.018	0.029***
	(0.014)	(0.010)	(0.013)	(0.010)	(0.012)	(0.010)
Individual prediction	0.092***	0.084***	0.044**	0.097***	0.038*	0.082***
	(0.021)	(0.017)	(0.020)	(0.017)	(0.019)	(0.016)

Notes: Table reports marginal effects from probit estimation of equations (1) and (2). Additional controls are given in column headings; see text for more information. In panel A, omitted wealth category is quartile 4. For panel A, income controls are omitted from the "application information" controls since they are highly multi-collinear with assets. In panel B, "preventive activity" measures the proportion of gender-appropriate preventive health behaviors undertaken; all estimates in panel B include an additional control for gender. Heteroskedacticity-adjusted robust standard errors are in parentheses. ***, **, * denote statistical significance at the 1-percent, 5-percent, and 10-percent level, respectively.

"Private Sources of Advantageous Selection: Evidence from the Medigap Insurance Market," by Fang, Keane, and Silverman (*JPE*, 2008)

- Show advantageous selection in Medigap coverage, and try to say something about its sources.
- Medigap: private insurance covering risk not covered by Medicare.
- Note: data combinations issues (observe y and x in MCBS and x and z in HRS)
- Surprising? Interesting?

Table 3: OLS Regression Results of Total Medical Expenditure on "Medigap" Coverage in MCBS, with No Health Controls

	Panel A: First "Medigap" Definition					
	(1)	(2)	(3)	(4)	(5)	(6)
Variables	All	Female	Male	All	Female	Male
medigap	-4392.7***	-6037.4***	-1863.4***	-3783.3***	-5687.4***	-1448.2***
	(347.0)	(456.6)	(540.8)	(375.4)	(485.7)	(569.8)
female	270.0			-263.6		
	(356.7)			(389.5)		
(age-65)	387.5***	460.6***	292.9	310.5**	325.1**	281.8
	(138.2)	(176.0)	(229.3)	(136.4)	(169.1)	(229.6)
(age-65)^2	1.94	-1.79	5.58	.94	-.744	.444
	(10.65)	(13.20)	(18.84)	(10.5)	(12.7)	(18.6)
(age-65)^3	.12	.17	.07	.134	.160	.143
	(.22)	(.27)	(.43)	(.220)	(.262)	(.422)
# of Observations	15,945	9,725	6,220	15,784	9,621	6,163
Adjusted R^2	.0702	.0873	.0531	.0869	.1089	.0680
	Panel B: Second "Medigap" Definition					
medigap	-4142.6***	-5883.1***	-1629.3***	-3457.2***	-5520***	-1297***
	(323.4)	(432.2)	(494.9)	(351.4)	(470.2)	(512.0)
female	-35.4			-545.5*		
	(313.1)			(338.0)		
(age-65)	425.2***	470.4***	383.9*	333.6***	336.8**	336.2*
	(122.2)	(156.3)	(198.1)	(120.7)	(150.8)	(197.3)
(age-65)^2	-4.12	-5.85	-3.37	-3.75	-4.42	-5.64
	(9.55)	(11.90)	(16.62)	(9.37)	(11.5)	(16.32)
(age-65)^3	.25	.26	.23	.237	.246	.250
	(.21)	(.25)	(.39)	(.201)	(.240)	(.379)
# of Observations	18,708	11,218	7,490	18,539	11,112	7,427
Adjusted R^2	.0638	.0830	.0442	.0787	.1024	.0588
State Dummy	Yes	Yes	Yes	Yes	Yes	Yes
Year Dummy	Yes	Yes	Yes	Yes	Yes	Yes
Other Demographic Controls	No	No	No	Yes	Yes	Yes

Note: The Dependent variable is "Total Medical Expenditure." All regressions are weighted by the cross section sample weight. See text and Data Appendix for the two definitions of Medigap.

Other demographics included are race, education, marital status, income, working and number of children.
Robust standard errors in parenthesis are clustered at individual level.
*, ** and *** denote significance at 10%, 5% and 1% respectively.

Table 4: OLS Regression Results of Total Medical Expenditure on "Medigap" Coverage in MCBS, with Direct Health Controls

	Panel A: First "Medigap" Definition					
	(1)	(2)	(3)	(4)	(5)	(6)
Variables	All	Female	Male	All	Female	Male
medigap	1937.0***	1677.3***	2420.9***	1732.8***	1426.2***	2210.1***
	(257.6)	(349.0)	(397.4)	(272.4)	(358.4)	(418.9)
female	-751.6***			-754.1***		
	(283.7)			(294.0)		
(age-65)	394.5***	417.5***	355.4*	419.6***	444.2***	392.1**
	(117.4)	(145.0)	(197.6)	(113.3)	(137.4)	(198.9)
(age-65)^2	-27.5***	-32.0***	-22.8	-28.3***	-32.7***	-25.2
	(9.3)	(11.4)	(16.3)	(9.0)	(11.1)	(16.4)
(age-65)^3	.474**	.548**	.466	.491**	.562**	.520
	(.207)	(.254)	(.380)	(.202)	(.247)	(.382)
# of Observations	14,129	8,371	5,758	14,105	8,365	5,740
Adjusted R^2	.2087	.1915	.2462	.2135	.2007	.2484
	Panel B: Second "Medigap" Definition					
medigap	1967.3***	1638.5***	2529.7***	1760.2***	1372.5***	2353.1***
	(238.7)	(311.5)	(377.4)	(255.9)	(330.0)	(398.3)
female	-926.1***			-911.9***		
	(264.0)			(275.7)		
(age-65)	371.6***	404.5***	371.8**	384.3***	417.5***	392.6**
	(104.1)	(129.2)	(171.1)	(101.8)	(124.0)	(172.0)
(age-65)^2	-25.6***	-30.2***	-24.8*	-25.6***	-30.3***	-26.3*
	(8.3)	(10.2)	(14.1)	(8.1)	(9.9)	(14.1)
(age-65)^3	.418***	.504**	.479	.420**	.506**	.520
	(.185)	(.227)	(.330)	(.182)	(.222)	(.331)
# of Observations	16,885	9,860	7,025	16,853	9,852	7,001
Adjusted R^2	.2001	.1906	.2342	.2042	.1991	.2362
State Dummy	Yes	Yes	Yes	Yes	Yes	Yes
Year Dummy	Yes	Yes	Yes	Yes	Yes	Yes
Other Demographic Controls	No	No	No	Yes	Yes	Yes

Note: The Dependent variable is "Total Medical Expenditure." All regressions are weighted by the cross section sample weight. See text and Data Appendix for the two definitions of Medigap.

The variables included as direct health controls are detailed in Data Appendix. The other demographics included are race, education, marital status, income, working and number of children.

Robust standard errors in parenthesis are clustered at individual level.

*, ** and *** denote significance at 10%, 5% and 1% respectively.

Table 8: Sources of Advantageous Selection: Predicting Medical Expenditure Using Only MCBS No Medigap Observations

	Coefficient Estimate of Pred. Exp./10,000			Conditioning Variables								
	A	B	C	risk tol.	pred. var.	risk_tol* pred. variance	educ.	inc.	cogn.	long. expec.	financial planning horizon	# obs.
Panel A: First Definition of Medigap												
(1)	-.0391 (.000)	-.0558 (.001)	-.0574 (.116)	N	N	N	N	N	N	N	N	9973
(2)	…	-.0558 (.001)	-.0570 (.118)	Y	N	N	N	N	N	N	N	3467
(3)	…	-.0301 (.162)	.0393 (.121)	Y	Y	Y	N	N	N	N	N	3467
(4)	…	-.0234 (.281)	.0508 (.063)	Y	Y	Y	Y	N	N	N	N	3467
(5)	…	-.0039 (.843)	.0636 (.060)	Y	Y	Y	Y	Y	N	N	N	3467
(6)	…	…	.0758 (.049)	Y	Y	Y	Y	Y	Y	N	N	1696
(7)	…	…	.0781 (.055)	Y	Y	Y	Y	Y	Y	Y	N	1695
(8)	…	…	.0783 (.061)	Y	Y	Y	Y	Y	Y	Y	Y	1659
Panel B: Second Definition of Medigap												
(9)	-.0534 (.000)	-.0754 (.000)	-.0776 (.022)	N	N	N	N	N	N	N	N	11866
(10)	…	-.0754 (.000)	-.0771 (.022)	Y	N	N	N	N	N	N	N	4295
(11)	…	-.0405 (.060)	.0224 (.398)	Y	Y	Y	N	N	N	N	N	4295
(12)	…	-.0273 (.212)	.0444 (.130)	Y	Y	Y	Y	N	N	N	N	4295
(13)	…	-.0069 (.726)	.0560 (.121)	Y	Y	Y	Y	Y	N	N	N	4295
(14)	…	…	.0683 (.087)	Y	Y	Y	Y	Y	Y	N	N	2146
(15)	…	…	.0694 (.089)	Y	Y	Y	Y	Y	Y	Y	N	2143
(16)	…	…	.0709 (.093)	Y	Y	Y	Y	Y	Y	Y	Y	2103

Note: *p*-value in parenthesis. All regressions include controls for female, a third order polynomial in age-65 and State.