A copula approach to test asymmetric information with applications to predictive modeling

joint work with Peng Shi, Northern Illinois University

58th Congress of the International Statistical Institute
Dublin, Ireland, 21-26 August 2011

Emiliano A. Valdez
Department of Mathematics
University of Connecticut
Storrs, Connecticut, USA

Outline

1. **Background**
   - Information asymmetry in insurance
   - Literature
2. **Empirical data**
3. **Model specification**
   - Marginals : ordered logistic regression model
   - Marginals: Negative Binomial regression model
   - Joint distribution: Frank copula
4. **Model calibration results**
   - Estimates
   - Quality of fit
5. **Concluding remarks**

Asymmetric information

- Contract theory: economic transactions between parties
- Asymmetric information (or sometimes called information asymmetry)
  - occurs when one party possesses information not available to the other contractual party
  - information is relevant in the sense that it could affect the economic transaction
- eBay market: transfer of ownership of used goods from one person to another.
  - There are variables, observable only during period of ownership and difficult to trace, that may assist buyer to assess the quality of goods being purchased.
  - Original owners has a sense of history of performance of goods (of which may not be revealed).
- Seminal paper by economist G.A. Akerlof (1970)
  - explained the problems from information assymetry based on the “used car” market
  - defective used cars were referred as “lemons”

Information asymmetry in insurance

- Present in all forms of insurance: life, medical, dental, homeowners, and automobile
  - additional information during the underwriting process may not be revealed
  - prevents the insurer to fairly price-discriminate and create a portfolio of homogeneous risks
- Asymmetric information due to:
  - adverse selection: insurer does not have enough information to assess those “high risks” groups who are more likely to purchase insurance.
  - moral hazard: behavior of policyholder is altered because of the presence of insurance (e.g. careless driving).
- Cohen and Siegelman (2010)
- Our work does not distinguish between these two types of information asymmetry.

Why important in insurance contracts?

- If policyholders are misclassified, this could lead to a deterioration of adequacy of premium.
  - insurer insolvency
  - insurance market collapse
- What can the insurer do - sharing of risks
  - modify insurance policy design within the limits of the law
  - use of policy deductibles, coinsurance, and policy limits
- Rothschild and Stiglitz (1976)
  - showed that if insurer offers a basket of goods with varying levels of coverage, there exists a separating equilibrium
  - optimally, this says that the higher risk group chooses better level of coverage and pays the appropriate higher premium, and vice versa
- This study motivates us to examine whether there is a positive relationship between the risk of policyholders and their choice of the level of coverage.

Literature

Results of empirical investigation of the relationship between risk and coverage in insurance have been mixed:

- Puelz and Snow (1994)
  - automobile insurance - found strong positive correlation
- Cawley and Philipson (1999)
  - life insurance - no evidence of positive correlation
- Chiappori and Salanié (2000)
  - (French) automobile insurance - no evidence of positive correlation
- Dionne, Gouriéroux, and Vanasse (2001)
  - automobile insurance - suggested none with “nonlinearity of the risk classification variables”
- Cohen (2005)
  - (Israel) automobile insurance - found evidence of information asymmetry
- Saito (2006)
  - (Japanese) automobile insurance - no adverse selection or moral hazard

Data

- Data used in our empirical investigation was drawn from a portfolio of automobile insurance policies of a major insurer in Singapore:
  - cross-sectional observations: calendar year 2001
  - total 44,226 policies recorded
  - a sub-sample from studies done in Frees and Valdez (2008) and Frees, Shi, and Valdez (2009)
- policy choices:
  - third party only (1)
  - third party, fire, and theft (2)
  - comprehensive coverage (3)

Summary statistics

Number and percentage of policy choice and reported accidents

Policy Choice Claim Count	1	2	3	Total Number	Percent
0	4721	7324	28411	40456	91.48
1	168	329	2950	3447	7.79
2	6	29	258	293	0.66
3	1	2	26	29	0.07
4	0	0	1	1	0
Total Number	4896	7684	31646	44226	
Percent	11.07	17.37	71.56		100

Explanatory variables

				Third Party		Fire and Theft		Comprehensive	
Variable	Value/Description	Mean	StdDev	Mean	StdDev	Mean	StdDev	Mean	StdDev
Driver characteristics									
ageclass	=1 if less than 25	2.97%		3.98%		6.52%		1.95%	
	=2 if between 26 and 35	31.90%		24.02%		32.96%		32.86%	
	=3 if between 36 and 45	35.33%		31.94%		34.28%		36.1%	
	=4 if between 46 and 55	21.89%		26.02%		20.11%		21.68%	
	=5 if between 56 and 65	6.61%		11.13%		5.39%		6.21%	
	=6 if greater than 65 (reference level)	1.30%		2.91%		0.74%		1.20%	
Sexinsured	=1 if female, 0 if male	17.23%		13.73%		12.16%		19.01%	
Marital	=1 if married, 0 if single	83.90%		84.56%		81.56%		84.37%	
experience	length of driving experience	11.96	8.13	12.45	8.94	10.88	7.66	12.15	8.09
NCD	No claims discount								
	=1 if 0 percent	31.94%		43.22%		49.52%		25.93%	
	=2 if 10 percent	14.89%		14.99%		17.43%		14.25%	
	=3 if 20 percent	11.17%		10.8%		10.09%		11.48%	
	=4 if 30 percent	7.20%		5.23%		5.35%		7.95%	
	=5 if 40 percent	6.23%		4.11%		3.66%		7.18%	
	=6 if 50 percent (reference level)	28.57%		21.65%		13.95%		33.21%	
Vehicle characteristics									
Vage	the age of the insured vehicle	7.55	6.48	18.36	6.20	13.90	3.95	4.33	3.27
vehicleclass	=1 if the vehicle is a private car	86.29%		84.19%		74.97%		89.36%	
	=2 if the vehicle is a goods vehicle	13.13%		13.13%		24.93%		10.27%	
	=3 if others (reference level)	0.58%		2.68%		0.10%		0.37%	
capacityclass	=1 if petty cars	11.27%		19.96%		16.07%		8.77%	
	=2 if small cars	33.25%		35.95%		30.8%		33.43%	
	=3 if medium cars	48.96%		38.30%		45.72%		51.39%	
	=4 if large cars (reference level)	6.52%		5.79%		7.41%		6.41%	
brandclass	=1 if Toyota	18.91%		22.81%		25.43%		16.73%	
	=2 if Honda	13.77%		13.83%		24.62%		11.13%	
	=3 if Nissan	16.88%		16.71%		14.28%		17.54%	
	=4 if Misubishi	10.34%		8.56%		7.26%		11.36%	
	=5 if Mazada	4.71%		4.08%		3.67%		5.06%	
	=6 if other Japanese car	4.10%		4.86%		4.27%		3.94%	
	=7 if Korean car	6.87%		1.61%		1.29%		9.04%	
	=8 If European car	19.92%		24.43%		17.14%		19.9%	
	=9 if others (reference level)	4.50%		3.11%		2.04%		5.30%	

Marginal model for the policy choice

Let y_{i1} indicate the policy choice for policyholder i, with possible values of 1, 2, or 3.

- The value of y_{i1} will be determined according to a corresponding latent variable denoted by y_{i1}^*.
- An ordered multinomial model is used to describe their relationship as

$$y_{i1} = \begin{cases} 1, & \text{if } y_{i1}^* \leq \alpha_1 \\ 2, & \text{if } \alpha_1 < y_{i1}^* \leq \alpha_2 \\ 3, & \text{if } y_{i1}^* > \alpha_2 \end{cases}$$

where α_1 and α_2 are unknown thresholds to be additionally estimated.

Ordered logistic regression model

The distribution function of y_{i1} can thus be expressed as:

$$F_{i1}(y_{i1}) = \text{Prob}(Y_{i1} \leq y_{i1}) = \begin{cases} \pi(\alpha_1 - \boldsymbol{x}_i^{'}\beta), & \text{if} \quad y_{i1} = 1 \\ \pi(\alpha_2 - \boldsymbol{x}_i^{'}\beta), & \text{if} \quad y_{i1} = 2 \\ 1, & \text{if} \quad y_{i1} = 3 \end{cases}$$

$\boldsymbol{x}_i$ denotes the vector of covariates that explain the policy selection.

Possible choices for the π function:

- ordered probit model: $\pi(a) = \Phi(a)$, where $\Phi(\cdot)$ is the std normal cdf
- ordered logit model: $\pi(a) = 1/[1 + \exp(-a)]$

Both methods typically provide consistent results and the selection between the two rests on the user's preference. For our purposes, we considered an ordered logistic regression model.

Marginal model for the number of accidents

The number of accidents y_{i2} is specified using a Negative Binomial regression model with:

$$\text{Prob}(Y_{i2} = y_{i2}) = \frac{\Gamma(y_{i2} + \psi)}{\Gamma(\psi)\Gamma(y_{i2+1})} \left(\frac{\psi}{\lambda_i + \psi}\right)^{\psi} \left(\frac{\lambda_i}{\lambda_i + \psi}\right)^{y_{i2}}$$

with a log link function used for its conditional mean given by

$$\text{E}(y_{i2}|\boldsymbol{z}_i) = \lambda_i = \omega_i \exp(\boldsymbol{z}_i^{'}\gamma)$$

with ω_i the weight (exposure) parameter for policyholder i.

The dispersion parameter ψ in the conditional variance

$$\text{Var}(y_{i2}|\boldsymbol{z}_i) = \lambda_i + \lambda_i^2/\psi$$

provides additional flexibility to accommodate either over or under dispersion.

$\boldsymbol{z}_i$ denotes the vector of covariates that explain the accidents.

See Cameron and Trivedi (1986).

Joint distribution

The joint probability mass function of y_{i1} and y_{i2} could be expressed as:

$$\begin{aligned}\text{Prob}(Y_{i1} =& y_{i1}, Y_{i2} = y_{i2}) \\ &= C(F_{i1}(y_{i1}), F_{i2}(y_{i2})) - C(F_{i1}(y_{i1}-1), F_{i2}(y_{i2})) \\ &\quad - C(F_{i1}(y_{i1}), F_{i2}(y_{i2}-1)) \\ &\quad + C(F_{i1}(y_{i1}-1), F_{i2}(y_{i2}-1)),\end{aligned}$$

where F_{i1} and F_{i2} are the cumulative distribution functions of y_{i1} and y_{i2}, respectively.

Here $C(\cdot,\cdot)$ denotes the copula function that links the marginals to its joint probability function.

Frank copula

To accommodate the fact that the choice of coverage and the frequency of accidents could possibly be either positively or negatively associated, we consider the Frank copula which permits such flexibility:

$$C(u_1, u_2; \theta) = -\frac{1}{\theta} \log \left[1 + \frac{(e^{-\theta u_1} - 1)(e^{-\theta u_2} - 1)}{e^{-\theta} - 1} \right]$$

θ is the dependence parameter.

It is rather straightforward to show that when:

- $\theta \to 0$, the case of independence
- $\theta > 0$ indicates a positive association
- $\theta < 0$ indicates a negative association

Genest (1987)

Estimation results

	Choice - Cumulative Logit				Risk - Negative Binomial		
	Estimate	StdErr	*p*-value		Estimate	StdErr	*p*-value
CHOICE_INT1	-4.1937	0.2293	0.0000				
CHOICE_INT2	-0.8475	0.2269	0.0002	RISK_INT	-3.5866	0.4701	0.0000
CHOICE_AGECLASS1	0.6671	0.1510	0.0000	RISK_AGECLASS1	0.7033	0.1999	0.0004
CHOICE_AGECLASS2	0.9527	0.1309	0.0000	RISK_AGECLASS2	0.4080	0.1803	0.0236
CHOICE_AGECLASS3	0.8321	0.1270	0.0000	RISK_AGECLASS3	0.2059	0.1773	0.2456
CHOICE_AGECLASS4	0.6806	0.1269	0.0000	RISK_AGECLASS4	0.2449	0.1775	0.1677
CHOICE_AGECLASS5	0.4300	0.1342	0.0014	RISK_AGECLASS5	0.2963	0.1850	0.1092
CHOICE_SEXINSUREDF	0.0523	0.0433	0.2268	RISK_SEXINSUREDF	-0.0665	0.0428	0.1204
CHOICE_MARITALM	0.0484	0.0456	0.2895	RISK_MARITALM	-0.0462	0.0466	0.3222
CHOICE_VAGE	-0.4732	0.0041	0.0000	RISK_VAGE	-0.0520	0.0031	0.0000
CHOICE_VEHICLECLASS1	3.6921	0.1730	0.0000	RISK_VEHICLECLASS1	1.3981	0.4254	0.0010
CHOICE_VEHICLECLASS2	3.0326	0.1757	0.0000	RISK_VEHICLECLASS2	1.3491	0.4269	0.0016
CHOICE_CAPACITYCLASS1	0.3666	0.0769	0.0000	RISK_CAPACITYCLASS1	-0.1069	0.0914	0.2423
CHOICE_CAPACITYCLASS2	0.1792	0.0686	0.0089	RISK_CAPACITYCLASS2	0.0771	0.0750	0.3039
CHOICE_CAPACITYCLASS3	0.5342	0.0665	0.0000	RISK_CAPACITYCLASS3	0.1662	0.0717	0.0205
CHOICE_BRANDCLASS1	-0.0980	0.0834	0.2400	RISK_BRANDCLASS1	-0.0284	0.0917	0.7568
CHOICE_BRANDCLASS2	0.0475	0.0850	0.5757	RISK_BRANDCLASS2	0.2164	0.0922	0.0189
CHOICE_BRANDCLASS3	0.1401	0.0867	0.1063	RISK_BRANDCLASS3	0.0210	0.0902	0.8163
CHOICE_BRANDCLASS4	-0.1286	0.0916	0.1604	RISK_BRANDCLASS4	0.0877	0.0942	0.3518
CHOICE_BRANDCLASS5	-0.1348	0.1080	0.2123	RISK_BRANDCLASS5	0.0329	0.1079	0.7605
CHOICE_BRANDCLASS6	-0.3339	0.1021	0.0011	RISK_BRANDCLASS6	-0.1263	0.1251	0.3127
CHOICE_BRANDCLASS7	0.2131	0.1146	0.0629	RISK_BRANDCLASS7	0.0545	0.0995	0.5840
CHOICE_BRANDCLASS8	0.3527	0.0874	0.0000	RISK_BRANDCLASS8	0.1191	0.0912	0.1916
CHOICE_EXPERIENCE	-0.0019	0.0022	0.3852	RISK_EXPERIENCE	-0.0056	0.0025	0.0247
CHOICE_NCD0	-0.6256	0.0452	0.0000	RISK_NCD0	0.3734	0.0484	0.0000
CHOICE_NCD10	-0.5822	0.0522	0.0000	RISK_NCD10	0.2984	0.0554	0.0000
CHOICE_NCD20	-0.3925	0.0570	0.0000	RISK_NCD20	0.1485	0.0609	0.0147
CHOICE_NCD30	-0.0686	0.0688	0.3191	RISK_NCD30	0.1972	0.0675	0.0035
CHOICE_NCD40	0.0568	0.0770	0.4609	RISK_NCD40	0.1889	0.0711	0.0079
				DISPERSION	2.0422	0.3417	0.0000
DEPENDENCE	1.4457	0.1437	0.0000				
Log-likelihood	-29026.14						
Chi-square test	108.30		0.0000				

Quality of fit

Examining the marginals:

Goodness-of-fit tests of the marginals

Choice			Risk		
Value	Observed	Fitted	Value	Observed	Fitted
1	11.07%	10.72%	0	91.48%	91.49%
2	17.37%	16.61%	1	7.79%	7.77%
3	71.56%	72.67%	2	0.66%	0.68%
			3	0.07%	0.06%
			4	0.00%	0.01%

Testing the robustness of the copula:

- We recalibrated the model using two other customarily used Archimedean copulas: the Gumbel-Hougaard and the Clayton copulas.
- The Frank gave a Spearman's rho of 23%; Gumbel-Hougaard 17%; Clayton 21%.

Bias due to underreporting

- Using reported accidents tends to overestimate the risk level of policyholders with high-coverage and thus subsequently distorts the possible coverage-risk relationship.
- To correct for this bias, we re-analyzed the data using only the claims where a third party is involved; such accidents are called bilateral accidents where there is a greater tendency to report.
- Similar in spirit to Chiappori and Salanié (2000) and Kim, et al. (1999).
- We therefore recalibrated our copula model and the positive residual coverage-risk association vanishes.

Our contribution to existing literature

- The copula approach: Unlike linear correlation models used in previous studies, this approach allows to capture both linear and nonlinear coverage-risk relationships.
- Avoiding potential endogeneity: We model the two responses simultaneously to avoid this issue that possibly arises when you examine the effect of a multinomial coverage selection measure on the number of accidents.
- Additional use for predictive modeling: Although mainly motivated to empirically examine asymmetric information, we found use of it for other actuarial applications.

Concluding remarks

- Additional work were done in the paper to demonstrate how the resulting copula model can be used for predictive modeling:
  - predict accident probability, given the choice of coverage
  - calculate pure premium – claim amounts needed to be additionally modeled
- Our article examines the use of copula regression models for investigating the presence of information asymmetry in a portfolio of insurance.
  - Using the reported accidents, we find evidence of the presence of asymmetric information. However, when we corrected the bias from possible underreporting, this evidence vanishes.
- A limitation of our analysis is that we focused on the policyholder's behavior over only a cross-section of a time.
  - examining repeated observations over time will be more informative
  - future work