한국경제학보 제19권 제1호
The Korean Journal of Economics Vol. 19, No. 1(Spring 2012)

Spurious Correlation in Testing for Asymmetric Information: The Case of Automobile Insurance Data

Yong-Woo Lee*

Abstracts

We examine the possibility of spurious conditional correlation in testing for asymmetric information in the context of the automobile insurance contracts. An important characteristic of insurance data is that the insurer can observe only claims, not accidents. As pointed in Chiappori(2000), this may cause a spurious correlation in the conditional correlation approach. In particular, it may be that certain losses are only covered under more comprehensive contracts and are only reported by a policyholder who has indeed purchased this more comprehensive insurance. Using a rich data set obtained from a major automobile insurance firm in Korea, we found out the possibility of spurious conditional correlation. Considering claims involving only bodily injuries enables us to overcome this problem. Using claims involving 2 vehicles is still susceptible to the problem.

KRF Classification : B030104
Keywords : Insurance, Asymmetric Information, Econometrics, Spurious Correlation

* The author is grateful to Jaap Abbring and Vassilis Hajivassiliou for providing him with intellectual stimulus to launch this research project. Also, constructive comments from two anonymous referees are deeply appreciated. Correspondence: Yong-Woo Lee, Korea Capital Market Institute, 18F, KOFIA Building, 143 Uisadang-daero, Yeongdeungpo-gu, Seoul, Korea. E-mail: leastsquares@gmail.com

Ⅰ. Introduction

Since the seminal paper by Chiappori and Salanie(1997), there have been numerous empirical research to test for asymmetric information using conditional correlation approach (for instance, Chiappori and Salanie(2000) and Dionne et al.(2001)). Further, Cohen(2005) has presented the evidence of learning process in the automobile insurance market using conditional correlation approach. More recently, Kim et al.(2009) has improved the method using the detailed coverage structure in the Korean automobile insurance contracts.

In this paper, we contribute to the method, focusing on the accident side. As emphasized by Chiappori(2000), both the insurer and econometricians observe only claims, not actual accidents. Under this circumstance, it is highly likely that this may cause a spurious correlation in the conditional correlation approach due to the endogenous feature of claim filing decision. In particular, even if losses are not affected by behavior - that is, if there is no moral hazard - it may be that certain losses are only covered under more comprehensive contracts and are only reported by a policyholder who has indeed purchased this more comprehensive insurance. That is, partial coverage cannot cover certain losses and, as a result, a policyholder cannot claim the reimbursement for this kind of losses. Besides, due to the existence of bonus-malus system, a policyholder always has some incentives not to report accidents after taking the costs of filing a claim and the expected benefits into consideration. Therefore, a regression using claims as the dependent variable is likely to generate misleading results. Overall, although coverage and accident occurrence have no direct causal connection, yet it may be wrongly inferred that they do, due to either coincidence or the presence of a certain third unobserved factor. This phenomenon precisely corresponds to what is known as spurious correlation. To

overcome this possible biases, we exclusively consider accidents involving bodily injuries since reporting is mandatory in that case.[1] Under this circumstance, claim records are likely to approximate true accident occurrence and, as a result, spurious correlation may be overcome.

In the next section, we explain the key features of the Korean automobile insurance coverage. Then, in section 3, data description, the results with full claim records, the results with claim records involving only bodily injuries, and the results with claim records involving 2 vehicles (suggested by Chiappori and Salanie(2000)) are presented turn by turn. We show the striking contrast between the results. In the last section, we conclude.

Ⅱ. The Korean Automobile Insurance Coverage and Accident Types

One of the main features of the Korean automobile insurance contract is the complexities of the insurance coverage structure. In the Korean market, there are six types of insurance coverage. Details are given in table 1 below.

【Table 1】 Types of Coverage

1. against the injuries or deaths inflicted on third party
2. against the remaining losses beyond the compensated amount made by the coverage 1
3. against the damages caused on the third party's property or vehicle
4. against the damages or theft on policyholder's own vehicle
5. against the injuries or death on policyholder or family members
6. against the injuries or deaths on policyholder or family members
7. caused by the third party's vehicle that is not sufficiently insured or kick and run vehicle

1) In Korea, this is specified in The Road Traffic Act article 54.

Policyholders can freely choose insurance cover amongst the coverage available. However there are some points to mention:

1. Coverage [1] is compulsory by law so that every policyholder has to buy this coverage in order to drive a vehicle.
2. Coverage [1] has a maximum possible reimbursement. When a policyholder is responsible for injury or death to a third party and this exceeds coverage [1]'s maximum possible reimbursement - having not purchased coverage [2], a policyholder must pay the exceeding money by himself. Coverage [2] was introduced to compensate for this kind of loss.
2. Purchase of coverage [1, 2 and 3] altogether corresponds to 'third party' automobile insurance in some developed countries.
3. Purchase of coverage [1, 2, 3, 4, 5, and 6] altogether is 'comprehensive' insurance in some developed countries.
4. Purchase of coverage 6 is possible only if the policyholder also purchases coverage [1, 2, 3 and 5].

【Table 2】 Accident Types

1. bodily injuries
2. automobile to automobile (frontal collision)
3. automobile to automobile (broadside collision)
4. automobile to automobile (fender bender)
5. automobile to automobile (rear end collision)
6. automobile to automobile (other)
7. one automobile (to motor bicycle)
8. one automobile (rollover)
9. one automobile (lane departure)
10. one automobile (collision to objects)
11. one automobile (other)
12. one automobile (theft)
13. one automobile (fire)
14. one automobile (flooding)
15. one automobile (typhoon)
16. other

Normally, the insurer classifies the accident types according to table 2 when there is a claim filing. Broadly speaking, there are 3 categories

of accident types, which are bodily injuries, accidents involving 2 vehicles, and accidents involving only one vehicle.

Ⅲ. Data and Empirical Results

1. Data

I have obtained data set on automobile insurance contracts from a major Korean non-life insurance firm that have been operating since 1983[2]. The data set from the firm covers 262,499 policyholders contracted between 01/01/2000 and 31/12/2000; samples are limited to those who are younger than 32 years old in 2000.

The data has four broad categories:

The first component is 'individual characteristics', which includes each policyholder's gender, age, place of residence (by post code), 'contract experiences coefficient' and 'bonus-malus' coefficient. The contract experience coefficient reflects how long a policyholder has contracted the automobile insurance. When the policyholder begins an insurance contract for the first time, the coefficient is 1.4, in subsequent years it decreases to 1.15 and 1.05 respectively. After the policyholder has contracted the insurance for 3 or longer years, this coefficient becomes 1.

The second part is 'contract information', which consists of dummy variables for family-limited policy and age-limited policy, contract starting and terminating date, coverage types, coverage amount for

2) Before 1983, there had been a monopoly within the automobile insurance market in Korea. As the number of cars increased and the market size got bigger this system was no longer compatible with the changing economic environment, both domestically and internationally. Thus, to enhance the competitiveness of the market, the government allowed 10 domestic and 2 foreign not-life insurance companies to operate in the car insurance market from 1983.

each coverage type, and applied insurance premiums. If a policyholder buys a family-limited contract, those who can drive are limited to family members - prescribed in the clauses - as the name suggests. There is a discount in insurance premiums for this policy because this choice is supposed to reduce accident probability. With regards to age-limited policy there are three types: all ages, over 21 year's old, and over 26 year's old. There is also a discount whereby the 'over 26 year old contract' has the highest discount. Coverage types are recorded separately (from 1 to 6 in table 1). Upon choosing coverage type, a policyholder can choose the monetary coverage amount for coverage [2, 3, 4, and 5]. This specifies the maximum possible reimbursement. Finally, insurance premiums for each policyholder are recorded.

The third part concerns information on the vehicle. This contains vehicle age (measured by the production year), vehicle size (measured by CC), gear types (automatic, semi-automatic or manual), dummy variable for ABS equipment and valuation of the vehicle.

The final component is information on accident occurrence. This includes: accident place (by post code), accident date, fault rate, loss amount, accident type (16 types) and vehicle loss type. Loss amount shows the actual amount of monetary compensation awarded to the policyholder when there was an accident. Vehicle loss type informs us whether vehicle was a total 'write off' or only partially damaged. Further to this it also separately shows whether a vehicle was stolen.

2. Method

Here, we closely follow the conditional correlation approach proposed by Chiappori and Salanie(2000). There are two probit models, one for 'contract choice' and the other for the 'occurrence of an accident'. Denote two independent errors following normal distributions with zero mean and unit variance by ϵ_i and η_i. Then we

have

$$d_i = 1(X_i\beta + \epsilon_i > 0)$$
$$n_i = 1(X_i\gamma + \eta_i > 0).$$

d_i is a dummy dependent variable for contract choice. As explained in section 2, the Korean coverage system is quite sophisticated. We have formulated a dependent variable, d_i, according to whether a policyholder purchased below coverage [3] listed in table 1. Thus, in our estimation, if a policyholder bought above coverage [3] in addition to coverage [1, 2 and 3], then $d_i = 1$ and $d_i = 0$ otherwise.

Also n_i is a dummy variable for an accident occurrence. If a policyholder had at least one accident in which s/he were judged to be at fault, then, $n_i = 1$ and $n_i = 0$ otherwise.

As for the exogenous variables, the most relevant ones in terms of premium pricing are included. Thus, we have dummy variables for gender (1), family-limited policy (1), age-limited policy (2), gear type (2), ABS equipment (1), car age (3), and car size (3). Overall, 13 exogenous variables are available. Insurer uses those variables in the determination of insurance premiums for each policyholder.[3)]

Finally, we also concentrate on young drivers.[4)] According to Chiappori and Salanie(2000), this has several advantages: One benefit being that the 'heteroscedasticity problem' is probably much less severe on a sample of young drivers since their experience is much more homogeneous than in a population in which different seniority

3) In Korean automobile insurance market, the base premium is mainly determined by vehicle characteristics. Individual characteristics such as place of residence, marital status, and so on do not play a role, apart from through family limited and age limited term indirectly. The premium structure is as follows: premium = base premium×limitation policy×individual characteristics coefficient×bonus-malus coefficient

4) This usually means the most recent drivers in insurance industry.

groups are mixed up: And, more importantly, concentrating on young drivers avoids the problems linked to the experience rating (bonus-malus coefficient). If we include it, the test may be biased since this variable is likely to be correlated with η_i in the second equation. To prevent this problem, we choose young drivers using a 'contract experience coefficient'; in particular we selected drivers whose number of years of contract experience is '1' so that they do not have an extended driving record history.[5] Given the model and variables, we first estimate two probits independently. Then we compute the generalized residuals $\hat{\epsilon_i}$ and $\hat{\eta_i}$. For instance, $\hat{\epsilon_i}$ is given by

$$\hat{\epsilon_i} = E(\epsilon_i | d_i, X_i) = \frac{\phi(X_i\beta)}{\Phi(X_i\beta)} d_i - (1-d_i)\frac{\phi(X_i\beta)}{\Phi(-X_i\beta)},$$

where ϕ and Φ denote the probability density and cumulative distribution function of $N(0,1)$. Finally, we compute a test statistic by

$W = \frac{(\sum_{i=1}^{n} \omega_i \hat{\epsilon_i}\hat{\eta_i})^2}{\sum_{i=1}^{n} \omega_i^2 \hat{\epsilon_i}^2 \hat{\eta_i}^2}$, where ω_i denotes the number of days with insurance cover.[6]

It is proposed that, under the null of conditional independence $cov(\epsilon_i, \eta_i) = 0$, W is distributed asymptotically as a $\chi^2(1)$. This provides a test of symmetric information. Overall, the idea is that when there is asymmetric information this should result in a positive correlation between d_i and n_i, which is equivalent to a positive

5) Since the data set contains the policyholders younger than at most 32 years old, this procedure adequately captures young drivers.

6) Chiappori and Salanie(2000) weigh each individual by ω_i due to the fact that they have data set in a calendar year. We have data sets also in contract years and almost all policyholders had completed one year of the contract (89.77%).

correlation between ϵ_i and η_i.

After testing two 'independent probits', we also estimate a 'bivariate probit' in which ϵ_i and η_i are jointly distributed. It has been argued that estimating the two probits independently is appropriate under conditional independence, but it is inefficient under the alternative. Thus, the 'bivariate probit estimation' is a reasonably complementary piece of work.

3. Results using Full Claim Records

When we use full claim records, we have 44,889 young drivers sample, among whom 7,914 policyholders reported claims.

【Table 3】 Claim and Coverage for Full Claim Records

Claims						Frequency	Percent
no claim						36,975	82.37
claim						7,914	17.63
Coverage						Frequency	Percent
1	2	3	4	5	6		
✓						6	0.01
✓	✓					1	0.00
✓	✓	✓				1,574	3.51
✓	✓	✓		✓		2,175	4.85
✓	✓	✓		✓	✓	14,529	32.37
✓	✓	✓	✓			25	0.06
✓	✓	✓	✓	✓		27	0.06
✓	✓	✓	✓	✓	✓	26,551	59.15

This claim records contain all kinds of accident types described in table 2. With this sample, we obtain test statistic, 89.58 which far exceeds $\chi^2(1)$ test statistic. Thus, we reject the 'conditional independence'. However, we suspect the possibility of spurious correlation. Particularly, it may result from the fact that certain losses are only covered under more comprehensive contracts and are only reported by a policyholder who has indeed purchased this more

comprehensive insurance. For instance, when we look at the theft data, this becomes very obvious. In the whole original data set, we have 61 theft claims. In this case, all of the coverage choices are 123456, which is full coverage. For theft accident, a policyholder who has not purchased coverage 4 does not need to report since she cannot be reimbursed. Likewise, some losses will not be reported by partially insured policyholders and, as a result, it would create biases in conditional correlation approach. As a complement work, we implement the bivariate probit model. Correlation coefficient is 0.15 and the likelihood ratio test rejects the null hypothesis of zero correlation ($\chi^2 = 68.891,\ p < .0000$).

4. Results using Claims involving Bodily Injuries only

If we use the claims involving bodily injuries only, we have 37,575 young drivers sample, among whom 600 policyholders reported claims.

【Table 4】 Claim and Coverage for Bodily Injuries

Claims						Frequency	Percent
no claim						36,975	98.40
claim						600	1.60
Coverage						Frequency	Percent
1	2	3	4	5	6		
✓						4	0.01
✓	✓					1	0.00
✓	✓	✓				1,447	3.85
✓	✓	✓		✓		1,983	5.28
✓	✓	✓		✓	✓	13,198	35.12
✓	✓	✓	✓			18	0.05
✓	✓	✓	✓	✓		19	0.05
✓	✓	✓	✓	✓	✓	20,905	55.64

With this sample, we have test statistic, 0.08 which is smaller than $\chi^2(1)$ test statistic. Thus, in contrast to the previous results,

conditional independence vanishes. For this case, we again implement the bivariate probit model. Although we obtain the correlation coefficient, 0.01, this is not statistically significant ($\chi^2 = 0.07$, $p < .7878$).

With regard to accident vs. claim problem, Chiappori and Salanie(2000) attempted to overcome the spurious correlation using the claim files involving two automobiles so that a claim is much more likely to be filed. We also have implemented this approach using accident type 2-7, totalling to 5,813 claims. However, the test statistic is still 59.22. Correlation coefficient from the bivariate probit model is 0.13 and statistically significant ($\chi^2 = 45.37$, $p < .0000$). Thus, we suspect the remaining endogeneity of claim filings. Further, in this case, we cannot rule out the possibility of bilateral transfers among policyholders involved in the automobile accidents.[7)]

【Table 5】 Claim and Coverage for Claims involving 2 Vehicles

Claims						Frequency	Percent
no claim						36,975	86.41
claim						5,813	13.59
Coverage						Frequency	Percent
1	2	3	4	5	6		
✓						5	0.01
✓	✓					1	0.00
✓	✓	✓				1,541	3.60
✓	✓	✓		✓		2,117	4.95
✓	✓	✓		✓	✓	14,197	33.18
✓	✓	✓	✓			21	0.05
✓	✓	✓	✓	✓		25	0.06
✓	✓	✓	✓	✓	✓	24,880	58.15

In both subsections, for a robustness check, we implement the same

7) Although Chiappori(2000) argues that this kind of 'street-settled' deal is difficult to implement, we cannot be sure about this argument, since we have been told from the profession that such bilateral agreements cannot be neglected in Korea.

test using only 123000(those who purchased coverage 1, 2, and 3) and 123456 coverage(those who purchased all 6 coverages) to compare full third-party coverage with full coverage. Although the quantities of the test statistic change, the qualitative feature remains the same (test statistic 93.32 for full claim records and 0.22 for claim records involving bodily injuries).

Ⅳ. Conclusion

As we have presented in the last section, in the conventional correlation approach, there is a possibility to obtain spurious correlation. This phenomenon is originated from the fact that we observe the claim records rather than directly observing accident occurrences. Therefore, if we use claim history without careful consideration, it is always likely to have biases in the regression. From the results given in this paper, we suggest to use more restrictive definition of claim, bodily injuries, to avoid the possible biases in the conditional correlation approach. The possible biases may also entail the distortion in policy design. According to the results in this research, there is no such a phenomenon as adverse selection in the Korean automobile insurance market. Therefore, the current risk sorting system used in the insurance firms may be properly working although the sorting is not as sophisticated as the one in developed countries.[8] Therefore, if we wrongly infer the existence of asymmetric information with wrong identification, then the policy based on this results are likely to distort market outcome.

Received: December 12, 2011. Revised: January 18, 2012. Accepted: January 30, 2012.

8) See footnote 3.

◈ *References* ◈

Chiappori, P-A (2000), "Econometric Models of Insurance under Asymmetric Information," In Handbook of Insurance, Springer.

Chiappori, P-A. and B. Salanie (1997), "Empirical Contract Theory: The Case of Insurance Data," *European Economic Review*, Vol. 41, 943-950.

Chiappori, P-A. and B. Salanie (2000), "Testing for Asymmetric Information in Insurance Markets," *Journal of Political Economy*, Vol. 108, 56-78.

Cohen, A(2005), "Asymmetric Information and Learning: Evidence from the Automobile Insurance Market," *The Review of Economics and Statistics*, Vol. 87, 197-207.

Dionne, G., Gourieroux, C. and C. Vanasse (2001), "Testing for Evidence of Adverse Selection in the Automobile Insurance Market: A Comment," *Journal of Political Economy*, Vol. 109, 444-453.

Kim, H., Kim, D., Im, S. and J. Hardin (2009), "Evidence of Asymmetric Information in the Automobile Insurance Market: Dichotomous Versus Multinomial Measurement of Insurance Coverage," *Journal of Risk and Insurance*, Vol. 76, 343-366.

비대칭적 정보의 검증에서 발생 가능한 가성적 상관관계: 자동차보험데이터의 경우

이 용 우*

논문초록

본 연구는 자동차보험계약에 존재하는 비대칭적 정보의 검증과정에서 발생할 수 있는 가성적 조건부 상관관계의 가능성을 고찰한다. 보험데이터의 중요한 특징 중의 하나는 보험회사가 실제 교통사고가 아닌 보험청구가 요청된 사고만을 관찰한다는 사실이다. Chiappori(2000)가 지적하듯 이러한 현상은 조건부 상관관계 접근법에서 가성적 상관관계를 초래할 수 있다. 특정 손실의 경우 포괄적인 보험에서만 커버가 가능하며 따라서 이러한 손실은 포괄적 보험에 가입한 보험소비자들만 청구할 것이기 때문이다. 본 연구는 우리나라 대형손보사의 데이터를 이용하여 분석한 결과 전체 보험청구를 이용하는 경우 가성적 상관관계가 나타나는 반면, 인적 상해만을 수반하는 보험청구를 이용하는 경우 이러한 가성적 상관관계가 사라짐을 보인다.

주제분류 : B030104
핵심 주제어 : 보험, 비대칭적 정보, 계량경제, 가성적 상관관계

* 자본시장연구원 연구위원, e-mail: leastsquares@gmail.com