Journal of Business Administration Online
HTTP://JBAO.ATU.EDU

Fall 2004, Vol. 3 No. 2

DO REGULATORS GET AUTO PREMIUMS RIGHT? A REGRESSION APPROACH

Marshall J. Horton
Puterbaugh Professor of Free Enterprise
Southern Arkansas University

Arthur Parry
Adjunct Professor of Business
LeTourneau University

James Rech
Property and Casualty Actuary
Towers Perrin Tillinghast

Abstract

This study uses linear regression analysis to provide independent analysis of actuarially determined auto liability premiums in Texas. The authors use generalized least squares with cross-sectional, publicly available data to obtain a high degree of correlation between benchmark premiums and several demographic variables which serve as proxies for insurer costs. The high correlation between premiums and proxy variables, which are outside the control of insurers or regulators, is taken as independent evidence that regulators in Texas have established a reasonable and not unfairly discriminatory ratemaking process for private passenger automobile liability insurance.

Do Regulators Get Auto Premiums Right? A Regression Approach

Introduction

Automobile liability insurers are under state mandates to justify premiums as reasonable and adequate, but not unfairly discriminatory. State insurance regulators are subject to criticism from consumers and consumer groups, which do not understand the actuarial methods by which allowable rates are determined. The authors seek to address this issue by trying to answer the question: "Can a simple model which uses publicly available data explain benchmark automobile liability premiums which were determined using actuarial methods, thereby

providing independent support for current regulatory practice?" A carefully chosen, simple regression model seems to provide an answer of "Yes."

Methodology

The auto liability line is unique among insurance products in that it is effectively required, in fulfillment of state financial responsibility laws, of virtually every adult in the United States. As a result, automobile premiums are not market determined, but are instead allowed on a jurisdictional level on a cost-plus basis. Because of this pricing process, consumer groups frequently monitor the premiums which regulatory agencies allow. Of the major states, Texas has perhaps the simplest structure of rates, determining a benchmark rate for each rating territory independently of risk class or age of primary policyholder.

Whitehead (1997), writing from the viewpoint of social responsibility, cautioned against the use of proprietary models, such as catastrophe models, in rate regulation, characterizing their potential use as "unverifiable 'black boxes.'" While Whitehead's warning was well founded, the authors submit that those outside the insurance industry, such as consumers and academicians, could view current actuarial practice as a black box. To the extent that outside constituencies hold this view, an independent methodology which uses data from outside the industry to evaluate premiums could be of great use to regulators in justifying their ratemaking approach.

The primary question at hand is whether a regression approach can be of potential use for regulators in justifying automobile liability insurance rates to outside constituencies. This study focuses on the situation in Texas, using the second-largest auto insurance market as a relatively straightforward example of state standards. The data were gathered independently of the insurance industry, making the results an outside, independent assessment of current industry practices. Thus the results are potentially useful for consumer advocacy groups which may wish to evaluate company rate filings. In addition, regulators may wish to incorporate some of the approach developed in this paper for evaluating company standards and practices. Practitioners themselves, such as actuaries, may find it useful to have outside, independent confirmation (or refutation) of their extremely thorough, yet complicated, methodology. To these ends, a simple econometric model of underwriting criteria is used in an attempt to explain the black box in which Texas rates are determined as an approximation to the actuarial approach.

Regression models have been used in insurance for many years. Standard references on linear regression models include Maddala (1977) and Judge, et. al. (1988). Van Slyke (1980) presented a critique of such approaches in insurance to property-liability insurance. In 1990, the State of California, in response to Proposition 104, implemented a multiple linear regression based automobile rate filing system using information supplied by the insurance industry. Bouzouita and Bajtelsmit (1997) used data from the National Association of Insurance Commissioners in a linear regression model to show relationships between regulated rates and underwriting standards, population density and residual coverage, and industry concentration and the size of the residual market. Bajtelsmit and Bouzouita (1998) used a regression approach to relate profitability of private passenger automobile insurance to market concentration. While each of these studies has made a major contribution to a better understanding of the regulatory process,

none has concentrated on providing a means of relating premiums to variables which are beyond the insurance industry's control.

The Actuarial Approach to Price Determination

The markets for private passenger automobile liability insurance are highly fragmented, consisting of the fifty states and the dozens of counties or territories within each state. Premium, the analog to market price, may be set on a county or territorial basis, based on variables which company experience has proven to be of value in assessing the probability of losses and claims. Actual dollar amounts of losses are difficult to obtain or altogether unavailable, even on a statewide basis, to either the regulatory agencies or insurance companies. Even if they were, attributing relevant loss dollars to accidents would be problematic, since the losses associated with bodily injury typically stretch out many years. Therefore, ratemaking authorities and companies use the number of accidents and, at the state level, projections of loss dollars, to approximate losses.

In most states, premiums and policy details are regulated by a centralized state agency. In Texas, companies must justify rates by filing detailed statements with the Department of Insurance. Actuaries working within the industry and the state commission have accepted the specific filing requirements. Therefore, solely market forces do not set the price for a personal automobile insurance contract. Rather, the price is regulated by a state agency. The agency generally allows a price equal to the full-expected costs of the contract plus an allowance for profit and contingencies. In general, an automobile rate is a function of the following variables: losses, loss adjustment expense, coverage, limits, class (based on risk characteristics), investment income, operating expenses, profit and contingencies, taxes, and governmental regulation).

To simplify the approach, a linear model may be used, in which the premium is determined on a cost-plus pricing formula similar to that of other regulated industries. Unlike most other enterprises, the cost of contract benefits (losses) is unknown at the inception of the contract. Automobile contract benefits represent a series of stochastic events. As a result, the aggregate losses for a specific group of insureds may vary significantly from year-to-year. This variation may result from a number of statistical reasons, an important one being the result of the composition of the company's portfolio of individual risks.

Over the years, automobile risk underwriting has provided a series of common risk factors gleaned from the observations of losses and associated with subgroups of insureds. The insurance carriers have received approval by the regulatory agencies that these risk factors have statistical validity in the underwriting of individual insureds. These subgroups are designated as "classes."

The more refined the underwriting system, i.e. the more classes, the greater is the probability that the insured's individual risk characteristics will more closely approximate their cohorts within the subgroup. At the most accurate extreme, each insured would be assumed unique, resulting in each insured's being placed in his or her own class. Unfortunately, such a classification system would tell us nothing about the expected losses for the class group. The more narrowly defined the classes, the less likely that the class will be sufficiently large to accurately approximate the

characteristics of the population. As a result, the number of classes is limited. In automobile insurance, underwriting is synonymous with class definition and the assigned of individual risks to their most appropriate risk class (or subgroup).

Because of the inclusion of "risk characteristics" in the pricing equation, the pricing function for an insurance company operation quickly becomes very complex. Risk characteristics that have been identified in the automobile insurance industry include: geographical location (urban or rural), mileage driven, age of drivers, sex, marital status, type of vehicle, number and severity of prior driving violations, number of prior accidents, number of years accident free, and driver's school certification. The price for a specific insured's policy varies directly with the magnitude of that individual's risk characteristics. Basic to the determination of the premium is the initial assignment of the insured to a specific underwriting class.

In the actual establishment of class premiums, the class rate is based on a modification of the statewide base rate. These modifications are known as relativity factors. First, however, the statewide base rate must be developed.

The statewide base rate is the average rate for all insureds which, when multiplied by each exposure, accumulates to the expected aggregate premium level needed to provide for all losses, cover all expenses, plus add a profit. The adequacy of the statewide base rate is established by reviewing the company's historical losses and pro forma expenses. The company's historical losses are adjusted for inflation to the expected midpoint of the future period in which the rates will be effective. These adjusted losses are then combined with a pro forma loss adjustment and other operating expenses, investment income, and profit and contingencies load to determine the aggregate premium requirement. The aggregate premium is simply divided by the exposures for the same period to determine the statewide base rate. To evaluate whether the suggested rates will be adequate to cover the future expected losses and expenses, the historical exposures are rerated using the revised estimates of losses and expenses. The base rates are refined to reflect characteristics of state-determined rating territories, which may span as few as one or as many as several dozen counties.

The premium for a specific individual insured is developed from this base rate through a series of adjustments. These adjustments are achieved through the application of relativity factors, as the overall state base rate is adjusted up and down on a scale of individual risk characteristics as determined by the weight of the appropriate factors. These factors may be either additive or multiplicative. By summing over all the insureds, the aggregate premium for the company's current in-force business is developed. Using an iterative process, the base rate is adjusted until the aggregate premium of the current in-force business equals the projected aggregate losses, expenses, and profit levels. According to Bickerstaff (1995), this methodology assumes that losses are correlated from each period to the next. Therefore, an actuarial approach to premium determination would contain an autoregressive component.

A Simple Regression Model

The extensive data and computation requirements for the actuarial approach described in the previous section impose a substantial burden on companies, ratemaking analysts, and consumer groups. Many of the variables, which cannot be directly observed or accurately recorded are proxied by other variables such as number of losses, mentioned above. Can a simple linear regression model, without incorporating actual loss data but rather using proxy variables of risk characteristics determined by actuaries to be of value, explain the dispersion of the state base rate among the different rating territories?

The State of California experimented with a regression approach in its 1990 implementation of Proposition 104. This implementation involved a large-scale econometric model, which incorporated many variables considered important by demographers and actuaries. The model to be considered in this paper is different from the Proposition 104 approach in that the model is a multiple linear regression using publicly available data.

To estimate the base line premium across territories, the authors specified a linear regression with actual base rates as established by the Texas Department of Insurance for January, 1997, from forty-nine such territories across Texas as the dependent variable baseline premium. These rates were established for the minimum liability requirements of $20,000 bodily injury per person, $40,000 bodily injury per accident, and $15,000 property damage per accident. Average annual rainfall was used as a proxy for weather conditions and weighted into the independent variable for precipitation using land area in square miles as the weighting variable. Also, two additive dummies were added to the regression, suburban dwellers and rural dwellers. Suburban dwellers were represented by a "1" if the territory adjoined one of the major metropolitan counties: Bexar (San Antonio), Cameron (Brownsville), Dallas, El Paso, Harris (Houston), Lubbock, Nueces (Corpus Christi), Potter (Amarillo), Tarrant (Fort Worth), Tom Green (San Angelo), Travis (Austin), and Webb (Laredo), and a "0" otherwise. Rural dwellers were represented by a "1" if the territory did not touch one of the major metropolitan counties and a "0" otherwise. Of course, the major metropolitan counties acted as the implicit dummy variable, receiving a "0" across both the explicit dummies. The demographic data and designations of major metropolitan counties were taken from the 1998 edition of the *Texas Almanac*.

Yet more demographic information was available from census data on the county level. Average weekly income, number of automobiles in the rating territory, and six age variables which assigned ages of the population of each rating territory according to the following groups: 5 - 17 years, 21-24 years, 25 - 34 years, 45 - 54 years, 55 - 64 years, and 75 and over were used to formulate additional independent variables. These variables are likely to relate to the baseline premium in that they are potentially related to losses. The youngest drivers, for example, are likelier to have much higher accident rates than are older drivers, at least those under retirement age, when driver reaction capabilities decline. Young married drivers and middle-aged drivers, as well as the oldest drivers, were also included to address a broad array of ages. Areas whose residents have higher incomes are likely to host more litigation over auto accidents and losses than are those areas with low-income residents, therefore the weekly income variable was included. The number of registered automobiles is likely a good proxy for population and driver density in a particular rating territory.

Results

The results of the multiple linear regression model are reported in Table 1.[1] Both the precipitation variable and overall F ratio are statistically significant at the 1 percent level. The rural dummy and age group 75 years and over are statistically significant at the 5 percent level. A plot of the residuals by independent variable and subsequent Park tests indicates that the residuals are heteroscedastic with respect to the first age group variable. Since heteroscedasticity typically results in inefficient estimates, in such cases, the model should be transformed if the nature of the heteroscedasticity is known. Transforming the model by dividing each variable through by the first age group variable, the authors obtained a generalized least squares (GLS) specification. The estimates from the GLS specification are also reported in Table 1. These results indicate that the slope parameters for the precipitation and automobiles registered variables are statistically significant at the 1 percent level, with the rural dummy and oldest age group again statistically significant at the 5 percent level. The young adult age group parameter estimate is statistically significantly different from zero at the 10 percent level.

The signs on the parameter estimates make sense, except that of the number of automobiles registered in the rating territory. Increased precipitation results in more accidents, therefore more losses and higher premiums. Rural drivers likely log less road miles than other drivers, leading to fewer accidents and lower premiums. Young drivers have high premiums. Old drivers have lower premiums. One would expect that the more automobiles in an area, the higher the risk of accidents and losses. Perhaps more congested areas also have warier drivers, better roads and lighting, or slower traffic. Possibly, those accidents that occur in congested areas can be handled more efficiently by insurance companies, resulting in economies of scale.

[1] The specification is the result of a specification search with variables representing all available age groups. The first model, not reported here, was a "kitchen-sink" model with a high R^2 and low t-values, a classic sign of high multicollinearity. Since the age groups were likely highly correlated with one another, the authors calculated cross-correlations between each of the quantitative dependent variables, eliminating those, which had little historical risk significance. The variables whose economic significance is noted in the previous section were kept in the OLS regression reported in Table 1. Strictly speaking, such a specification search should result in higher standard errors and, therefore, lower significance levels, but the authors' a priori determination of economically significant variables made adjustment of the standard errors unnecessary (see Leamer (1978)).

Table 1

Regression Results: A Simple Linear Model of Benchmark Premium Across Rating Territories

	Ordinary Least Squares	Weighted Least Squares (normalized by Percentage of Population Aged 5 to 17 years)
Intercept	530.341 (12.930)	11.600 (6.112)
Average Annual Precipitation	2.907* (-3.594)	2.646* (3.922)
Average Weekly Income	-0.037 (-0.341)	0.010 (-0.571)
Number of Registered Automobiles	3.50 E-5 (1.710)	-5.51 E-5* (2.968)
Additive Dummy representing Suburban Drivers	-33.499 (-1.352)	-19.972 (-1.239)
Additive Dummy representing Rural Drivers	-57.151** (-2.526)	-24.284** (-2.626)
Percentage of Population Aged 5 to 17 years	0.026 (0.006)	
Percentage of Population Aged 21 to 24 years	-7.960 (-0.838)	6.722*** (1.951)
Percentage of Population Aged 25 to 34 years	-2.063 (-0.190)	4.554 (0.910)
Percentage of Population Aged 45 to 54 years	-12.551 (-0.957)	-4.484 (-0.723)
Percentage of Population Aged 55 to 64 years	.657 (0.047)	1.175 (0.316)
Percentage of Population Aged 75 years and higher	-20.207** (-2.300)	-6.555** (-2.048)
R^2	.6743	.7507
Adjusted R^2	.5657	.6766
Overall F ratio	6.383*	10.402*

(t-values in parentheses)
* Statistically significant at one percent level (two-tail test)
** Statistically significant at five percent level (two-tail test)
*** Statistically significant at ten percent level (two-tail test)

What is probably most notable from Table 1 is the high coefficient of determination which the WLS models provides: over three-quarters of the variation in premium across rating territory can be explained using publicly available data independent of insurance companies. That a cross-sectional model using proxy variables and publicly available data can explain so much of premium variation indicates a potentially useful approach for public interest groups, regulators, and even actuaries, who wish to independently check up on insurance industry information or actuarial methods.

Given the stability of losses and highly iterative nature of the ratemaking process over time, access to Texas Department of Insurance historical data for benchmark rates would allow researchers to estimate an autoregressive component in a time series or panel data model, and would likely yield an extremely good fit with the actuarially determined benchmark rates.

Recommendations and Extensions

Treating the auto pricing decision (ratemaking) as a black box, the authors used a simple, linear, cross-sectional model to explain more than three-quarters of the variation in baseline premium across rating territories. The analysis was completed using publicly available data, substituting proxies for those variables whose values are either unavailable or closely held by the industry and regulatory authorities. Notably, none of the variables included in the final specification was a proxy for race, gender, or other unfairly discriminatory considerations.

As noted above, the ratemaking process is highly dependent upon ongoing business from year-to-year. This would imply that an autoregressive model, using time-series data, might explain well over 90 percent of premium variation across rating territories.

To a great degree, the results of this model independently confirm actuarial methodology. Regulators, however, and consumer advocacy groups who are sensitive to charges of unfairly discriminating against demographic groups would do well to develop linear regression approaches in analyzing just how rates are determined. Practitioners, such as actuaries, may find that factors previously not considered, such as precipitation, can be of use in explaining potential loss behavior.

The approach can be extended to the county level for the State of Texas, since the Texas Department of Insurance maintains baseline premium information for each county. In addition, other states, such as Massachusetts, can be analyzed. Some data are also available on the companies, which sell auto. A study which evaluates company premium using economic and demographic variables in a regression approach could be useful to an industry which frequently has to justify the price it charges its customers.

Conclusion

In this paper, it was explained that the pricing process for automobile liability insurance is administratively driven at the state level. Using actuarial practice, regulatory authorities set base rates for each rating territory.

The situation in Texas, a large automobile liability insurance state, was examined using a simple regression model. Specifically, a simple, single-equation, cross-sectional model explained more than 75 percent of the variation in base rates using publicly available data, without recourse to actual loss data. This result demonstrates that economic theory is likely to make a significant contribution to the automobile ratemaking process as a less complicated alternative to the actuarial approach. This alternative should be of use to consumer advocates, regulators, economists, and insurance industry practitioners. Perhaps most important, there is independent evidence that, at least in Texas, insurance regulators have determined auto liability premiums correctly.

References

Bajtelsmit, Vickie L. and Raja Bouzouita, 1998. "Market Structure and Performance in Private Passenger Automobile Insurance." *Journal of Risk and Insurance*, 65 (3): 503-514.

Bickerstaff, David R., 1995. Testimony Before Texas Auto Benchmark Rate Hearing (Public Record), (November 28).

Bouzouita, Raja and Vickie L. Bajtelsmit, 1997. "The Impact of Rate Regulation on the Residual Market for Automobile Insurance." *Journal of Insurance Regulation*, 16(1): 61-72.

Judge, George G., and R. Carter Hill, William E. Griffiths, Helmut Lutkepohl, and Tsoung-Chao Lee, 1988. *Introduction to the Theory and Practice of Econometrics, Second Edition*. New York: John Wiley and Sons.

Leamer, Edward E., 1978. *Specification Searches: Ad Hoc Inference with Nonexperimental Data*, New York: John Wiley and Sons.

Maddala, G.S., 1977. *Econometrics*. New York: McGraw-Hill, Inc.

Van Slyke, Oakley E., 1980. Is Econometric Modeling Obsolete?" with Review by Fusco, Michael. *CAS Discussion Paper Program*: 650-687.

Whitehead, Selwyn, 1997. "Risky Business: Proprietary Modeling and Insurance Ratemaking." Journal of Insurance Regulation, 15(3): 372-381.