Selection procedures of regression analysis applied to automobile insurance [continued]

Autor(en): **Lemaire, Jean**

Objekttyp: **Article**

Zeitschrift: **Mitteilungen / Vereinigung Schweizerischer Versicherungsmathematiker = Bulletin / Association des Actuaires Suisses = Bulletin / Association of Swiss Actuaries**

Band (Jahr): **79 (1979)**

PDF erstellt am: **26.05.2024**

Persistenter Link: https://doi.org/10.5169/seals-967122

Selection Procedures of Regression Analysis Applied to Automobile Insurance

Part II: Sample Inquiry and Underwriting Applications

By Jean Lemaire, Bruxelles

1. Objectives of the Sample Inquiry

In [1], we observed during one year the entire automobile third-party liability portfolio of a Belgian company (106,974 policies) and applied different selection procedures of regression analysis in order to sort out the criteria that significantly influence the risk. These were
- the power and the age of the car,
- the age, the territory, the language of the main driver,
- the merit-rating class,
- the number of claims without responsability, and
- the type of coverage

when the dependent variable was the number of claims, and
- the power of the car,
- the merit-rating class,
- the language of the driver, and
- the number of claims without responsability

when the amount of the claims was to be explained.

Those results raised many questions, concerning principally the causes of the significance of unexpected variables like the type of coverage, the driver's language and the number of claims without responsability. On the other hand, the values of several variables that were felt to be important, like the annual mileage or the marital status, were not known to the company. We therefore conducted a sample inquiry among the policy-holders, asking
- the profession, the nationality, the marital status and the number of children of the main driver,
- if the car was regularly driven by somebody else,
- the number of cars of the household,
- the annual mileage, the mileage driven at work, during the holidays, and the distance between the place of residence and the office.

3,995 valid answers were collected.

2. Claim frequencies

The following tables summarize the main results for the new variables. The average claim frequency of the sample (.0999) nearly equals the overall portfolio frequency (.1011). We only computed the claim frequencies for the classes containing more than 50 policy-holders.

1. By profession

Profession	Frequency
Worker	.0997
Clerk	.1146
Senior staff	.1111
Teacher	.1099
Civil servant	.1037
Tradesman	.1364
Craftsman, farmer	.0435
Professional job	.0676
Retired	.0886
Housewife	.0411

2. By nationality

Nationality	Frequency
A	.0968
B	.1176
C	.1212
D	.1373
E	.2500
F	.4167

3. By marital status

Marital status	Frequency
Married	.0909
Widower	.1171
Single	.1451
Separated	.1846
Divorced	.2152

4. *By number of children*

Children	Frequency
0	.1063
1	.1018
2	.0967
3	.0849
4	.0822
> 4	.1034

5. *By mileage*

Number of Kilometers							
Annual	Fre-quency	During Holidays	Fre-quency	At Work	Fre-quency	Between House and Office	Fre-quency
0– 4,999	.0584	0– 999	.0921	0– 999	.0952	0– 5	.0984
5,000– 9,999	.0681	1,000–2,999	.0995	1,000– 4,999	.1192	5–10	.0934
10,000–14,999	.0949	3,000–4,999	.1014	5,000– 9,999	.0838	10–15	.1099
15,000–19,999	.1042	5,000–7,999	.1362	10,000–19,999	.1109	15–20	.0979
20,000–24,999	.1313	8,000+	.1566	20,000+	.1134	20–25	.0990
25,000–29,999	.1418					25–35	.1183
30,000+	.1044					35+	.1024
Averages	15,344		1,697		4,104		9,758

3. Selection of the Significant Criteria

In order to apply the selection techniques (described in [1]) of regression analysis, it was necessary to group some sub-classes of policy-holders, in order to avoid too weak absolute frequencies. Considering the results of the preceding section, we created the following new variables (see [1] for the definition of the variables x_1 to x_{25});

x_{26}: number of cars of the household;
x_{27}: number of children;
x_{28}: annual mileage;
x_{29}: mileage during the holidays;
x_{30}: mileage at work;
x_{31}: distance between place of residence and office;

x_{32}: dichotomous variable characterizing the most dangerous profession (tradesman);

x_{33}: dichotomous variable characterizing the least dangerous professions (craftsman, farmer, housewife and professional jobs);

x_{34}: dichotomous variable characterizing the dangerous nationalities B to F;

x_{35} to x_{39}: dichotomous variables characterizing the marital status.

We then applied the stepwise selection procedure in order to determine the variables that significantly influence the number of claims. 14 were selected, representing 8 criteria.

Criterion	Variable		Regression Coefficient
Age of the driver		x_9	−.001492
Merit-rating class		x_{10}	.002328
Power of the car		x_{12}	.000585
Territory	Towns	x_{23}	.016388
	Suburbs	x_{24}	.0
	Villages	x_{25}	−.016802
Annual mileage		x_{28}	.000480
Tradesman		x_{32}	.030861
Nationality		x_{34}	.036997
Marital status	Married	x_{35}	−.053492
	Widower	x_{36}	−.026126
	Single	x_{37}	.0
	Separated	x_{38}	.027551
	Divorced	x_{39}	.056692
Constant			−.005127

The significant criteria are thus

1. The *age of the driver:* A 40-year old driver causes 22.38% less accidents than a 25-year old policy-holder.
2. The *merit-rating class:* The frequency rises by 2.328% per bonus point, or 11.64% for each bonus class. This confirms the fact that the actual merit-rating system is not sufficiently severe for the bad drivers.
3. The *power of the car:* We notice an increase of the claim frequency of 5.85% per class of 10 H. P.
4. The *territory:* The people living in the suburbs stand about half-way between the citizens of large towns (+16.8%) and the countrymen (−16.38%).
5. The *tradesmen:* They form a very bad risk: +30.86%. Notice that x_{34}, the variable characterizing the least dangerous professions, is not selected. This can be explained by the fact that the policy-holders of this group present an

annual mileage markedly inferior than the average. Thus the introduction of x_{28} suffices to explain the discrepancy between the claim frequencies.

6. The *marital status:* The effects of this criterion are very spectacular indeed: comparing to the singles, the married drivers (–53.49%) and the widowers (–26.13%) should be entitled to a discount, while the separated (+27.55%) and the divorced policy-holders (+56.69%) present very high average claim frequencies. Note that x_{27}, the number of children, is not significant when the marital status is introduced in the regression.
7. The *nationality:* The policy-holders of nationalities B to F cause on the average 37% more accidents than the others! It is interesting to note that, comparing to the results of the first study, the variable "nationality" replaces the variable "language of the driver": the difference between the Dutch group and the French group seems spurious and can be entirely explained by the fact that most foreign workers fill their forms in French: the partial correlation coefficient between the number of claims and the language, controlling for the effects of nationality, does not differ significantly from zero.
8. The *annual mileage:* As was suspected, this is an important variable: one notices a significant correlation between the actual premium and the annual mileage. The introduction of this single variable accounts for the disappearance of three awkward variables of the former study: the number of claims without responsability, the age of the car and the type of coverage. For instance, it was demonstrated that the dangerous drivers tend to be involved in more accidents in which they have no responsability. This can be explained by the fact that they spend more time on the road than the average, and are more exposed to the risk: there exists a positive correlation between the annual mileage and the number of claims without responsability.

   Notice that the other mileage variables are less significant than x_{28}; consequently they do not appear in the final regression equation.

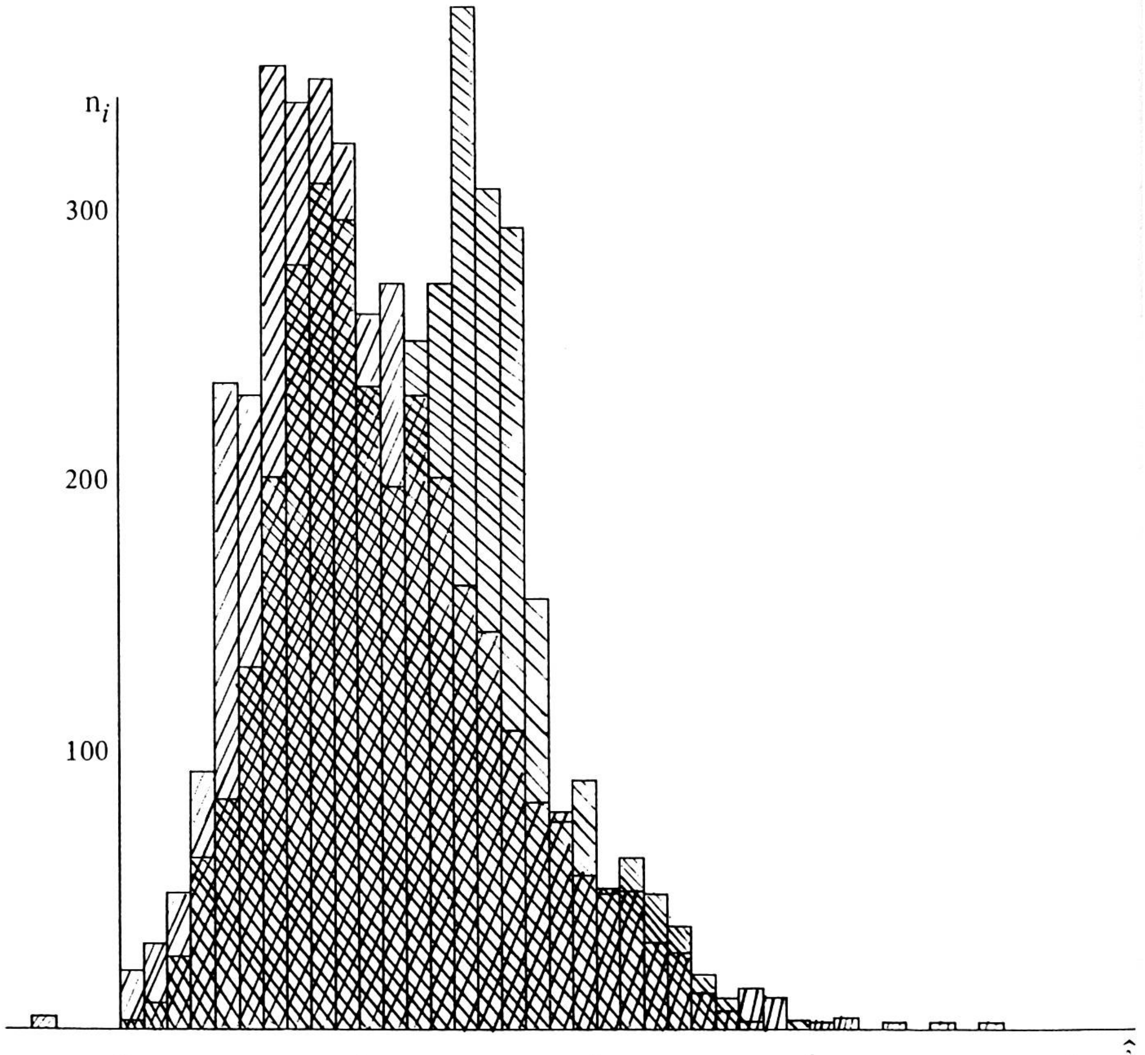
   The multiple correlation coefficient between the number of claims and the set of independent variables is .1519. This represents an increase of the efficiency of the regression (measured by the explained percentage of the variance of the number of claims) of 32% comparing to the first study, and of 105.5% comparing to the actual tariff[1].

[1] The sample has now been under observation for two and a half years. The sample size reduced to 3,892 (since most of the companies apply the same rates and all the contracts are signed on a ten-year basis, very few policy-holders leave the company). Only very small variations of the regression coefficients were obtained. The multiple correlation between x_1 and the selected variables rose from .1519 to .2056, a further increase in efficiency of 77.8%.

The number of claims in the sample (399) is too low to perform an analysis of the claim amounts. We therefore did not attempt to select the criteria influencing x_3, the total amount of the claims.

4. Practical Applications

In Belgium the automobile premiums are fixed by law, and it is therefore not allowed to use the preceding findings in order to modify the rates. Consequently, those results can only be used for underwriting purposes. In fact, the final regression equation provides an index which is the «best» linear a priori estimation of the claim frequency $\hat{\lambda}$. Fig. 1 presents the distribution of this index for the policies of the sample inquiry (hachures with positive slope), which by construc-



tion have been under observation for at least a full year, and for the first 3,995 policy-holders that entered the company since january 1978 (hachures with negative slope). The discrepancy between those distributions demonstrates the necessity of some filtering, for instance by rejecting the policy-holders presenting an index higher than some given quantile λ_R of the distribution of λ, and/or by imposing a surcharge to the drivers whose index is higher than another quantile λ_S (with $\lambda_S < \lambda_R$).

Reference

[1] *Lemaire, Jean:* Selection procedures of regression analysis applied to automobile insurance. *Mitteilungen der Vereinigung Schweizerischer Versicherungsmathematiker, Heft 2,* 1977, pp. 143 to 160.

Jean Lemaire
Institut de Statistique
Université Libre de Bruxelles
Campus de la Plaine, C.P. 210
50, bd du Triomphe
B-1050 Bruxelles

Zusammenfassung

Wir beschreiben die Ergebnisse einer Untersuchung von ungefähr 4000 Autoversicherten bei einer wichtigen belgischen Gesellschaft, um die Befunde einer früheren Arbeit über die Faktoren, die das Autohaftpflichtrisiko beeinflussen, genauer analysieren zu können.

Résumé

Nous décrivons les résultats d'un sondage, d'effectif proche de 4000, effectué auprès des assurés automobiles d'une grande compagnie belge, dans le but de préciser les résultats d'une étude antérieure ([1]) concernant les facteurs influençant le risque automobile.

Riassunto

Descriviamo i resultati di un campionamento di dimensione di circa 4000, effettuato presso gli assicurati automobile di un importante compagnia belga d'assicurazioni per spiegare più precisamente i resultati di uno studio anteriore concernente i fattori che influiscono sur rischio automobile.

Summary

The results of a sample inquiry of nearly 4,000 automobile policy-holders are described, in order to develop the findings of a preceding paper ([1]) studying the criteria that significantly affect the automobile risk.