Selection procedures of regression analysis applied to automobile insurance

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

Objekttyp: **Article**

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

Band (Jahr): **77 (1977)**

PDF erstellt am: **22.07.2024**

Persistenter Link: https://doi.org/10.5169/seals-967015

Selection Procedures of Regression Analysis Applied to Automobile Insurance

By Jean Lemaire

Abstract

We expose here some of the findings of a large survey of the automobile portfolio of a Belgian company. More than 100,000 policies were observed during a year, and the values taken by 24 variables recorded. A classical test of equality of means compelled us to reject the hypothesis of independence between the variables "number" and "average amount" of the claims, so that we had to study them separately. After the analysis of the claim frequencies and pure premiums, we applied three selection techniques of regression analysis to determine the significant criteria, and studied the regression- and correlation coefficients of this "optimal" tariff.

1. Statistical Data

106,974 policies – the entire automobile third party liability portfolio of a Belgian company – were observed during the period running from the first of July 1975 to the 30th of June 1976. This was a period of relative stability of the claim process, following a law enforcing the compulsory use of the safety belts, and reducing the speed limits and the drinking tolerances. This law proved to be very effective since the over-all claim frequency dropped to a bit more than 10%.

Only the policies in force during the whole period were taken into account. This omission of the new drivers slightly reduces the claim frequency (.1011 instead of .1098); this is only of trifling importance, since we want to study the relative influence of the different factors.

The policies were classified according to the values taken by the following 24 variables x_i (between brackets the observed percentage of policies in the portfolio);

x_1 = *number of accidents with full or partial responsability;*
x_2 = *number of accidents without any responsability;*
x_3 = *total amount of the claims* (or latest estimation);
x_4 = *mean amount of a claim;*
x_5, x_6, x_7 = dichotomous variables characterizing the type of vehicle;

$$x_5 = \begin{cases} 1 & \textit{if the car belongs to the category "tourism and business", i.e. if it is never used for the transportation of goods} \ (96.99\%), \\ 0 & \textit{otherwise;} \end{cases}$$

$$x_6 = \begin{cases} 1 & \textit{if the car may be used for the transportation of goods} \ (\text{"mixed category"}) \ (2.75\%), \\ 0 & \textit{otherwise;} \end{cases}$$

$$x_7 = \begin{cases} 1 & \textit{if the automobile is a sports car} \ (.26\%), \\ 0 & \textit{otherwise;} \end{cases}$$

despite their different classification, cars of the categories "tourism and business" and "mixed" belong to the same tariff class. Sports cars on the other hand have a special rating.

$$x_8 = \begin{cases} 1 & \textit{if the policy-holder is sedentary (i.e. if he uses his car only to and from work and for pleasure purposes)} \ (86.3\%), \\ 0 & \textit{otherwise;} \end{cases}$$

sedentaries benefit of a 15% discount in Belgium;

x_9 = *age of the main driver on the first of january 1976;*
x_{10} = *level of premium in the Belgian merit-rating system on july the first, 1975* (see for instance [2] for the presentation of this system);
x_{11} = *price of the car;*
x_{12} = *number of horse-power of the car;*
x_{13} = *cylinder volume of the car;*
x_{14} = *age of the car on january the first, 1976;*

$$x_{16} = \begin{cases} 1 & \textit{if the car has a diesel engine,} \\ 0 & \textit{otherwise;} \end{cases}$$

$$x_{17} = \begin{cases} 1 & \textit{if the main driver is a male} \ (87.17\%), \\ 0 & \textit{otherwise;} \end{cases}$$

$$x_{18} = \begin{cases} 1 & \textit{if the main driver is a female} \ (9.61\%), \\ 0 & \textit{otherwise;} \end{cases}$$

$x_{19} = \begin{cases} 1 & \textit{if the car belongs to a company } (3, 22\%), \\ 0 & \textit{otherwise;} \end{cases}$

$x_{20} = \begin{cases} 1 & \textit{if the policy-holder speaks French,} \\ 0 & \textit{if he speaks Dutch;} \end{cases}$

x_{21} = *effective third party liability premium paid, at the price of the first of july 1975* (in Belgium the premiums are bound to a price index);

$x_{22} = \begin{cases} 1 & \textit{if the policy-holder has subscribed to a full coverage insurance } (6.69\%), \\ 0 & \textit{if he only paid for the compulsory third party liability;} \end{cases}$

x_{23}, x_{24}, x_{25} = dichotomous variables characterizing the geographical location or territory of the car.

$x_{23} = \begin{cases} 1 & \textit{if the main driver lives in a town of more than 40,000 inhabitants} \\ & (17.06\%), \\ 0 & \textit{otherwise;} \end{cases}$

$x_{24} = \begin{cases} 1 & \textit{if } x_{23} \textit{ and } x_{25} = 0 \ (48.90\%), \\ 0 & \textit{otherwise;} \end{cases}$

$x_{25} = \begin{cases} 1 & \textit{if the main driver lives in a village of less than 5,000 inhabitants} \\ & (34.04\%), \\ 0 & \textit{otherwise.} \end{cases}$

Only the variables x_7, x_8, x_{10} and x_{12} appear in the Belgian tariff.
Of course some data are missing, like the age of the driver if the car belongs to a company, or the age of the car if the policy-holder purchased a new one during the observed period. This accounts for small discrepancies in the calculation of the average claim frequencies of § 3.
Unfortunately the company does not possess vital information like annual mileage, marital status, profession, ..., but we are performing for the moment a sample inquiry in order to determine the influence of those characteristics.

2. Independence Between Claim Number and Claim Amounts

If we want to apply regression techniques, two of the preceding variables can be considered as dependent: x_1, the number of the claims, and x_3, the total amount of the claims. In most actuarial papers, x_1 and the average amount of a claim are postulated to be independent as a first approximation, so that one can limit the

study to x_1. Yet the intuition suggests that this assumption might not be true, that for instance the citizen of the towns cause more accidents than the countrymen, but with lesser damages. To check for this hypothesis we have computed the observed conditional means of x_4, given the values of x_1.

Accidents	Absolute Frequency	Conditional Mean	Conditional Variance
1	9,240	36,621	3.72×10^{10}
2	740	40,797	3.39×10^{10}
3	43	14,620	4.95×10^{8}
4	9	14,387	7.24×10^{7}

The preceding table shows that the drivers involved in three or more accidents present a considerably lower claim amount average than the other groups. A classical test for equality of means rejects the null hypothesis at every usual level of probability. One might however object [1] that since very few people can afford to destroy four cars in a single year, the independence assumption should be tested using the distribution of the amount of the first accident instead of the mean claim cost. The preceding result is not altered, since the mean cost of the first accident of the group of the policy-holders causing 3 or 4 claims is 13,040 (variance 2.06×10^8), significantly different from 36,621. So we are forced to abandon the independence hypothesis and to study x_1 and x_3 separately.

3. Claim Frequencies, Mean Accident Cost, Pure Premium

The following tables summarize the results for the main variables. They give for each class or set of classes

1. the claim frequency,
2. the mean accident cost, and
3. the pure premium, which results from (1) and (2).

Since we are interested only in the relative value of those amounts between the classes, we have set the mean claim cost and the pure premium to 1,000 Belgian Francs for one class. The mean cost of an accident is 37,400 B. F. (100 Belgian francs $\simeq$ 7 Swiss francs), the mean pure premium is nearly 4,000 B. F. The huge difference between this last figure and the mean effective premium (10,000 B. F.)

[1] We thank Prof. *H. Bühlmann* for this valuable suggestion.

is explained by the commissions, the expenses, and the heavy taxes (17.25%) imposed to the Belgian drivers. Classes marked with an * contain less than 150 cases.

1. By age class

x_9	(1)	(2)	(3)
*18–19	0.1389	303	403
19–20	0.3554	481	1,636
20–21	0.2445	1,611	3,767
21–22	0.1932	543	1,003
22–23	0.2035	1,072	2,086
23–24	0.1950	1,451	2,707
24–25	0.1736	834	1,384
25–30	0.1321	982	1,242
30–35	0.1075	694	714
35–40	0.1090	1,179	1,229
40–45	0.1028	1,081	1,063
45–50	0.1046	1,000	1,000
50–55	0.0980	1,230	1,153
55–60	0.0910	1,151	1,002
60–65	0.0902	1,560	1,345
65–70	0.0980	411	385
+70	0.1284	2,794	3,433

Mean 39.15 years.
Standard deviation 13.2 years.

With exception of the first class (whose absolute frequency is very weak), one notices very high claim frequencies for the young drivers, a steep decrease until the late fifties then a moderate rise.

2. By merit-rating class

Class	(1)	(2)	(3)
1	0.0662	1,000	1,000
2	0.0779	795	936
3	0.0903	947	1,293
4	0.1009	854	1,302
5	0.1171	814	1,442
6	0.1573	943	2,241
7	0.1437	737	1,601
8	0.1652	747	1,865
9–10	0.1648	1,040	2,589
11	0.1748		
12	0.1654		
13	0.1858		
*14	0.1261	834	2,097
*15	0.1327		
*16	0.1731		
*17	0.1852		
*18	0.1500		

Mean 3.57.
Standard deviation 2.44.

In the classes with sufficiently many observations, one observes a quasi-linear dependence between the claim frequency and the merit-rating class. This seems to indicate that the system succeeds quite well in discriminating the policy-holders. Remark the small absolute frequency in the upper classes.

3. By class of power, cylinder volume and price

x_{12}	(1)	(2)	(3)
10– 30	0.0840	1,000	1,000
30– 40	0.0890	836	886
40– 50	0.0945	804	905
50– 60	0.0981	640	748
60– 70	0.1052	731	916
70– 80	0.1104	749	984
80– 90	0.1141	831	1,129
90–100	0.1130	1,123	1,511
100–150	0.1316	952	1,490
+150	0.1461	366	635

x_{11}	(1)	(2)	(3)
0– 80	0.0990	1,452	1,115
80–120	0.1135	632	556
120–160	0.1289	1,000	1,000
160–200	0.1303	570	576
200–300	0.1317	418	427
+300	0.1042	364	294
(x1,000 *F*)			

x_{13}	(1)	(2)	(3)
0– 500	0.0762	1,539	1,192
500– 600	0.0828	505	425
600– 700	0.0779	631	499
700– 800	0.0976	484	482
800– 900	0.0912	1,012	938
900–1000	0.0984	1,000	1,000
1000–1100	0.0971	1,064	1,049
1100–1200	0.0997	897	909
1200–1300	0.0966	1,006	987
1300–1400	0,1190	1,286	1,556
1400–1500	0.0940	829	792
1500–1600	0.1147	997	1,162
1600–1700	0.1140	1,347	1,561
1700–1800	0.1084	815	897
1800–1900	0.1042	1,302	1,379
1900–2000	0.1051	1,270	1,356
2000–2500	0.1318	840	1,124
2500–3000	0.1377	1,062	1,486
3000–4000	0.1224	779	969
+4000	0.1436	404	590

Mean power 61.6 HP.
Standard deviation 21.3 HP.

Mean cylinder volume 1309 cc.
Standard deviation 436 cc.

Mean price 118,500 B. F.
Standard deviation 26,500 B. F.

The dependence between the claim frequency and the number of horse-power is nearly linear, increasing however for the very powerful cars. The dependence concerning the cylinder volume and the price of the car is less apparent.

4. By age of vehicle

x_{14}	(1)	(2)	(3)
0– 2	0.1243	1,000	1,000
2– 3	0.1272	1,156	1,183
3– 4	0.1216	1,418	1,387
4– 5	0.1184	1,158	1,103
5– 6	0.1297	1,012	1,055
6– 7	0.1276	772	792
7– 8	0.1491	966	1,159
8– 9	0.1705	1,204	1,650
9–10	0.1601	828	1,066
10–11	0.1595	979	1,256
11–12	0.1754	514	725
12–13	0.1635	686	903
13–14	0.1709	875	1,203
14–15	0.1447	298	346
+ 15	0.0752	404	244

Mean 4.64 years.
Standard deviation 2.77 years.

No clear relationship emerges from the data, with the exception of a frequency increase for the old cars.

5. By effective premium paid

x_{21}	(1)	(2)	(3)
0– 6,000	0.0891	1,000	1,000
6,000– 7,000	0.0859	1,138	1,098
7,000– 8,000	0.0916	801	823
8,000– 9,000	0.0942	708	748
9,000–10,000	0.0971	691	754
10,000–11,000	0.1024	821	944
11,000–12,000	0.1058	741	880
12,000–13,000	0.1168	1,061	1,392
13,000–14,000	0.1325	414	616
+ 14,000	0.1490	1,353	2,264

Mean 10,018 B. F.
Standard deviation 1,661 B. F.

The strong positive linear relationship indicates (fortunately!) that the claim frequency is positively correlated with the premium. No such dependence emerges for the pure premium.

6. By type, use, sex

	(1)	(2)	(3)
Tourism and business	0.1009	1,000	1,000
Mixed	0.1073	881	936
Sports cars	0.1099	2,842	3,097
Sedentaries	0.1003	1,000	1,000
Non-sedentaries	0.1063	1,113	1,180
Males	0.1002	1,000	1,000
Females	0.1066	1,015	1,080
Companies	0.1073	1,146	1,227

The claim frequency rises by about 6% for the non-sedentaries, the ladies and the "mixed" category.

7. By territory, language, type of coverage

	(1)	(2)	(3)
Towns	0.1208	1,000	1,000
Suburbs	0.1043	1,111	959
Villages	0.0865	1,593	1,140
French	0.1058	1,000	1,000
Dutch	0.0915	840	727
Third party liability	0.0928	1,000	1,000
Full coverage	0.1387	969	1,441

The observations confirm the fact that the accidents in towns are more frequent, but less severe, so that the pure premium is more or less unchanged. The two last results are unexpected: the strong rise (48.7%) in the claim frequency of the persons who purchased full coverage (this proves that those drivers judge themselves very well) and the better results of the Dutch group.

One notices immediately that the influence of each variable is more apparent when one studies the number of claims instead of the pure premium. This is due to the importance of the very large claims: the 39 accidents (3.6‰) causing damages of more than a million B. F. account for more than 32% of the total claim amount. Eliminating those accidents would only transfer the problem.

Many companies base their tariffication on the preceding tables. They select a few discriminating criteria, determine the surcharges to apply to each class, and add or multiply them. This procedure is nevertheless incorrect if the explaining variables are not independent, since the conditional means do not take into account the interrelationships or interactions between the variables. Adding surcharges for young drivers *and* for used cars *and* for townsmen is unfair if younger persons are more likely to live in towns and to buy used cars than the others. Adding surcharges for sports cars *and* for powerful cars *may* be counting twice for the same factor, since both variables are certainly strongly positively correlated. To base a premium solely on loss-ratios without accounting for interrelationships may cause some absurdities or contradictions in the tariff: an American company found that the persons purchasing an insurance with a 100 U. S.-$ deductible should pay more than the 50$-deductible group. The reason for this paradox is that young drivers, usually less wealthy than their elders, turn themselves towards cheaper forms of insurance and thus provoke an antiselection.

Those examples clearly demonstrate the necessity of the use of techniques able to analyse the correlations between the variables and to isolate the effects of each factor. So we shall turn ourselves to regression analysis. The use of this technique implies a linearity hypothesis whose validity might be questioned, for instance for the effects of the variable «age of the driver». The working conditions and the urgency of this first over-all investigation compelled us to choose between

(i) applying regression techniques to a very large number of cases, and

(ii) applying more refined models to a sample of policies.

We hope to proceed to step (ii) when we shall possess more detailed information about the policy-holders.

4. Selection of the Significant Criteria – Regression with x_1

Let

$$x_1 = b_0 + \sum_{j \in Q} b_j x_j$$

be the regression equation, where Q denotes the set of the explaining variables. Let $R_{x_1(Q)}$ be the multiple correlation coefficient of this regression. Its square is the ratio of the sum of the squares due to regression to the total sum of squares. Finally let us denote by β_i the regression coefficient of the variable x_i in the population corresponding to the observed b_i.

To sort out the significant variables, we have applied three different selection procedures. The main tool of these techniques is the partial Fisher-Snedecor F test of significance of

$$H_0: \beta_i = 0 \text{ against } H_1: \beta_i \neq 0.$$

Under the null hypothesis, the expression

$$F = \frac{R^2_{x_1}(Q) - R^2_{x_1}(Q \setminus \{x_i\})}{1 - R^2_{x_1}(Q)} \cdot (n - q - 1)$$

admits a Fisher-Snedecor distribution with 1 and $n-q-1$ degrees of freedom, where n is the number of cases and q the number of variables of Q.

The three selection procedures are

a) The backward selection
In this procedure the variables are eliminated one by one until all remaining variables are significant.
1. Start with all the variables.
2. Compute the observed F for all the variables in the regression.
3. Select the smallest F.
4. Apply the partial F-test to the variable selected.
If H_0 is rejected, all the variables in the regression are significant and the selection is terminated.
If H_0 is not rejected, delete the variable and go to 2.

b) The forward selection
Here all the significant variables are entered one by one in the regression equation.

1. Select the variable x_i which is the most correlated to x_1:

$$|r_{x_1x_i}| = \max_j |r_{x_1x_j}|.$$

2. Compute all the partial correlation coefficients

$$r_{x_1x_j.x_i} = \frac{r_{x_1x_j} - r_{x_1x_i}.r_{x_jx_i}}{\sqrt{1-r^2_{x_1x_i}}\ \sqrt{1-r^2_{x_jx_i}}} \qquad \forall_j \neq i.$$

3. Select the highest partial correlation (in absolute value)

$$|r_{x_1x_k.x_i}| = \max_{j \neq i} |r_{x_1x_j.x_i}|.$$

4. Apply the partial F-test to x_k.
If H_0 is not rejected, no variable significantly increases the multiple correlation and the selection is over.
If H_0 is rejected, introduce x_k in the regression.

5. Compute all the second-order partial correlation coefficients

$$r_{x_1x_j.x_ix_k}.$$

6. Select the highest $|r_{x_1x_j.x_ix_k}|$ and apply the partial F-test to this variable.

c) The stepwise selection
In the forward selection, once a variable is entered, it remains in the regression equation until the end of the procedure, although the later introduction

of a strongly correlated variable may render the former one useless. Hence we can refine the method by testing all the variables in the regression, in order to possibly eliminate one of them. At each step of the selection, we shall thus

1. insert the variable with the highest partial correlation,
2. compute the observed F for all the variables in the regression,
3. apply the partial F-test to the smallest observed F.

   If H_0 is rejected, proceed to next step.

   If H_0 is not rejected, suppress the corresponding variable.

   If it is the variable just entered, stop the procedure

   If it is another one, proceed to next step.

Applied to our insurance sample, all three methods lead to the same solution, which consists of nine significant criteria. The procedures are summarized in the following table.

Step	Variable	Significance	Multiple Correlation	Step	
1	x_{10}	0.000	0.09836		
2	x_2	0.000	0.10686		
3	x_{12}	0.000	0.11346		
4	x_{14}	0.000	0.11956		
5	x_9	0.000	0.12296		Selected variables
6	x_{22}	0.000	0.12551		
7	x_{25}	0.000	0.12781		
8	x_{20}	0.000	0.13102		
9	x_{24}	0.000	0.13221		
	x_7	0.134	0.13230	9	
	x_8	0.255	0.13235	8	
	x_{11}	0.479	0.13236	7	
	x_{21}	0.533	0.13238	6	Eliminated variables
	x_5	0.618	0.13239	5	
	x_{19}	0.658	0.13240	4	
	x_{13}	0.763	0.13240	3	
	x_{16}	0.759	0.13240	2	x_6, x_{17}, x_{23} were not
	x_{18}	0.969	0.13240	1	introduced in the regression to avoid multicollinearity

The regression coefficients of the optimal regression equation are:

Variable	Coefficient	Confidence Interval ($\alpha = 5\%$)
x_{10}	0.002328	(0.002151; 0.002506)
x_2	0.067980	(0.0571; 0.07886)
x_{12}	0.000538	(0.000442; 0.000633)
x_{14}	0.004827	(0.00408; 0.005574)
x_9	−0.000799	(−0.000636; −0.000962)
x_{22}	0.030123	(0.021795; 0.038451)
x_{25}	−0.033466	(−0.027282; −0.039649)
x_{20}	−0.022079	(−0.017669; −0.026489)
x_{24}	−0.016578	(−0.010837; −0.022319)
Constant	−0.048171	(−0.02793; −0.068412)

The nine significant variables are thus:

1. x_{10}, *the level of merit-rating*. The claim frequency rises by 2.3% [2] by merit-rating point;
2. x_2, *the number of accidents without any responsability*. It is quite surprising that this criteria is the most important after merit-rating. Each claim without fault should be penalized by an increase of 41.7% of the premium! This unexpected result naturally rise unanswered questions. Is this increase of the claim frequency due to the fact that some bad drivers induce claims without responsability through a too nervous behaviour on the road? Or is this positive correlation spurious and due to the absence of another variable, like annual mileage for instance? Some previous studies ([1]) indeed indicate that the policy-holders with big annual mileage cause more accidents. Since they spend more time on the road, they are also more likely to undergo claims without responsability.
3. x_{12}, *the power of the car*. We notice an increase of the claim frequency of .000538 per HP. This means 2.48% between 60 HP and 70 HP, 2.26% between 100 HP and 110 HP, instead of 11.14% and 2.8% respectively in the present statutory Belgian tariff.
4. x_{14}, *the age of the car*. This variable enters in all selection procedures before the age of the driver. The premium should moderately increase (2.79% each year) with the age of the car.

[2] Unless otherwise specified, the reductions or surcharges concern the example of a French speaking 40-year old policy-holder, living in a large town and driving in class 10 of the merit-rating system a new 60 HP-car.

5. x_9, *the age of the driver*. A 50-year old driver causes 13.4% less accidents than a 20-year old policy-holder. The analysis of the claim frequencies suggest however that the validity of the linearity hypothesis might be seriously questioned.
6. x_{22}, *the type of coverage*. The regression analysis confirms that the drivers purchasing full coverage are more likely to provoke claims and should be penalized by a 18.5% increase of the premium.
7. x_{20}, *the language of the driver*. The most unexpected result of this study is that the Dutch group represents a far better risk than the French group (13.5%). More information is required in order to distinguish between the possible causes of this variation: difference in the driving ability, in the quality of the roads, in the annual mileage, or simply the fact that most foreign workers, coming from countries where the claim frequency is higher than 10%, usually fill their form in French.

8 and 9. x_{24} and x_{25}, *the territory*. Comparing to the citizen of large towns, the countrymen are entitled to a 21.5% discount, while the people living in the suburbs stand about half-way between those two extremes.

Notice the absence of two variables of the Belgian tariff: x_8 (sedentarity) and x_7 (sports cars). The 6% lower claim frequency and the significant negative zero-order correlation of the sedentaries result from a spurious relationship. The dependence can be fully explained by the fact that the sedentaries own on the average less powerful cars. Adding surcharges for non-sedentaries and for mighty cars amounts to counting twice for the same factor. The same remark applies to the sports cars: the positive zero-order correlation between x_1 and x_7 becomes slightly negative (although non significantly) when one accounts for the power of the car. It is thus unfair to impose a surcharge to the owners of sports cars.

Remark also the disappearance of the sex factor. It is not necessary to rate the females differently once the merit-rating is introduced. The partial correlation between x_1 and x_{18}, controlling for the effects of merit-rating, is exactly zero: the ladies are sufficiently penalized by a higher average merit-rating class.

Finally notice that the 48.7% rise in claim frequency for the full coverage group reduces to 18.5% when one accounts for all the significant criteria: the policy-holders of this group are indeed younger, they live in more crowded areas, they drive more powerful cars.

We must conclude this paragraph by observing the very low value of the multiple correlation coefficient (.13221), which proves that the regression equation is a poor predictor of the claims. However, the introduction of the nine variables would increase by more than 56% the discriminating power of the Belgian tariff, measured by the explained percentage of the variance of the number of claims.

5. Selection of the Significant Criteria – Regression with x_3

The three methods converge to the same solution which consists of a set of four variables (x_2, x_{10}, x_{20}, x_{21}). If, however, we apply the stepwise selection procedure with the statutory tariff as a starting point (i. e. x_8, x_{10} and x_{12} in the initial regression equation), the solution is the set (x_2, x_{10}, x_{12}, x_{20}). The difference between both solutions (the power replaces the effective premium) is very light, since the correlation coefficient between x_{12} and x_{21} is .88. The first set is slightly more efficient ($R_{x_3}(Q)$ = 04454 versus .04401), but the greater easiness of interpretation of the second set overcomes the small difference.

Step	Variable	Significance	Candidate Variable	Multiple Correlation
	x_8	0.637		
1	x_{10}	0.000	x_2 (S = 0.000)	0.01755
	x_{12}	0.019		
	x_2	0.000		
2	x_{10}	0.000	x_{20} (S = 0.013)	0.04331
	x_{12}	0.032		
	x_2	0.000		
3	x_{10}	0.000	x_{25} (S = 0.294)	0.04401
	x_{12}	0.026		
	x_{20}	0.013		

Variable	Coefficient	Confidence Interval (α = 5%)
x_2	14,733.93	(12,430.22; 17,037.64)
x_{10}	78.89	(43.83; 113.96)
x_{20}	−1,148.59	(−2,057.48; −239.71)
x_{12}	22.90	(2.8; 43.01)
Constant	−2,122.8	(−5,272.86; +1,027.26)

In the "ideal" tariff should thus appear 4 variables:
- *the number of claims without responsability,*
- *the merit-rating system,*
- *the language of the main driver, and*
- *the power of the car.*

Notice the disappearance of the age and territory factors (the greater average claim amount of the accidents in the villages balances the smaller claim frequency for instance).
Note also the very low value of the multiple correlation coefficient. We hope that the introduction of more variables will increase this value significantly.

Bibliography

[1] *Lantelli, G.*: Novelties in Swedish automobile rating. *The ASTIN Bulletin. Vol. II* (1962), pp. 96–101.
[2] *Lemaire, J.*: La soif du bonus. *The ASTIN Bulletin. Vol. IX* (1977), pp. 181–190.

Jean Lemaire
Institut de Statistique
Université Libre de Bruxelles
CP. 210 Campus de la Plaine
Boulevard du Triomphe
B-1050 Bruxelles

Acknowledgment

We gratefully acknowledge the important contribution of Marc Hallin (Université Libre de Bruxelles) to this work. His nonparametric generalization of the step-wise selection procedure, presented at the latest ASTIN Colloquium, will eliminate most of the objections that can be addressed to the distributional assumptions of regression analysis.

Zusammenfassung

Die Gesamtzahl der Autoversicherungen einer belgischen Versicherungsgesellschaft (106974 Policen) wurde während eines Jahres beobachtet. Nachdem man festgestellt hat, dass «Zahl» und «Betrag» der Unfälle voneinander abhängig sind, und man die Häufigkeit der Unfälle und die Beiträge für alle Klassen des Hauptvariablen berechnet hat, wenden wir drei verschiedene Selektionsmethoden der Regressionsanalyse an, um die Kriterien zu bestimmen, die einen bedeutenden Einfluss auf das Risiko haben.

Résumé

Le portefeuille automobile d'une compagnie d'assurances belge (comprenant 106974 polices) a été observé pendant un an. Après avoir démontré que les variables «nombre» et «montant» des sinistres ne sont pas indépendantes, et calculé les fréquences de sinistres et les primes pures pour toutes les classes des principales variables, nous appliquons trois méthodes de sélection de l'analyse de régression pour déterminer les critères influençant significativement le risque.

Riassunto

Il portafoglio automobilistico di una compagnia di assicurazioni belga (106974 polizze) è stato osservato durante un anno. Dopo aver dimostrato che le variabili «numero» e «importo» dei sinistri non sono indipendenti e dopo aver calcolato le frequenze dei sinistri e i premi puri per tutte le classi dei principali variabili, abbiamo applicato tre metodi di selezione dell'analisi di regressione per determinare i criteri che influenzano in modo significativo il rischio.

Summary

The entire portfolio of a large Belgian company (106,974 policies) was observed during a year. After the demonstration of the dependence between the variables "number" and "amount" of the claims, we compute the different claim frequencies and loss ratios for the major variables, and apply three selection procedures of regression analysis to sort out the significant criteria.