Validation of the selected factors impact on the insured accident

Ověření vlivu vybraných faktorů na vznik pojistné události

Martina Borovcová [1]

Abstract

Paper is focused on application possibilities of the logistic regression in the insurance sector. There are generally defined the areas of the possible application of the logistic regression, there is in detail the area of the non-life property insurance analyzed, more precisely accident insurance. By applying logistic regression are key factors defined with influence on the occurrence probability of insured accident, where influences of binary, categorial and continuous variables are analyzed. Impacts of selected factors are within one dimensional analysis quantified, as well.

Key words

Regression analysis, Logistic Regression, Generalized Linear Models, Binary variables, Categorial variables, Continuous variables,

JEL Classification: C20**,** G22

1 Úvod

Nejčastěji využívané přístupy k analýze dat nejrůznější povahy souvisí s regresní analýzou. Její využití v oblasti financí, řízení a rozhodování finančních institucí a tedy i pojišťovnictví je rovněž reálné. Při vyslovení slova regrese se zpravidla vybaví regrese lineární, méně často nelineární nebo logistická, i když právě logistická regrese je již nejméně tři desetiletí standardní metodou v západoevropské a americké vědě včetně společenské.

Cílem příspěvku je proto ověření vlivu vybraných faktorů na vznik pojistné události při využití logistické regrese. Na konkrétním příkladu je pomocí logistické regrese zhodnoceno, zda vybrané faktory jsou určujícími pro vznik pojistné události a tento vztah je dále kvantifikován. V článku je nejprve definována a vysvětlena regresní analýza a stručně je popsán regresní model. Následně je zmíněna podstata logistické regrese, včetně odhadu koeficientů metodou maximální věrohodnosti a poté je provedena analýza vlivu vybraných faktorů na vznik pojistné události ve smyslu havárie motorového vozidla.

2 Regresní analýza

Statistické metody, pomocí nichž odhadujeme hodnotu určité náhodné veličiny na základě znalosti veličin jiných, označujeme jako regresní analýzu. Přitom náhodná veličina, jejíž

[1] Ing. Martina Borovcová, Ph.D., VŠB-TU Ostrava, Ekonomickáfakulta, katedra financí, Sokolská třída 33, 702 21 Ostrava, martina.borovcova@vsb.cz. Tento příspěvek vznikl v rámci řešení projektu 2011 " Modelování a predikce pojistných rizik ve výuce Pojišťovnictví na Ekf-VŠB TU Ostrava"

hodnota je odhadována, může být označena také jako závisle proměnná, cílová proměnná, proměnná vysvětlovaná nebo také odezva, regresand. Naproti tomu veličina, jejíž znalost již máme, je nezávisle proměnná, proměnná vysvětlující, regresor. Ne vždy je použita nezávisle proměnná jediná. Často vystupuje regresorů několik, přičemž může jít o další veličiny, nebo funkce menšího počtu veličin.

Modelování vztahů mezi vysvětlující a vysvětlovanou proměnnou patří mezi základní aktivity, se kterými je možné se setkat ve statistice. Obvyklý je předpoklad, že závisle proměnná je náhodnou veličinou s normálním rozdělením. Pro odvození modelu je pak zpravidla použita metoda nejmenších čtverců.

Je-li však závisle proměnná znakem binárním, nikoli spojitým statistickým znakem, může nastat problém. V takovém případě by k odhadu parametrů bylo použití regresní analýzy s odhadem regresních koeficientů prostřednictvím metody nejmenších čtverců problematické.

Podstatou řešení regrese je pak stanovení nejlepšího regresního modelu, spočívající v určení matematické rovnice, která bude popisovat závislost y na x, stanovení parametrů modelu, související se stanovením nejlepších odhadů parametrů β, stanovení statistické významnosti modelu, související s určením, zda nalezený model přispěje ke zpřesnění odhadu závisle proměnné oproti použití pouhého průměru, či interpretace výsledků zjištěných modelem z hlediska zadání.

3 Logistická regrese

Cílem analýzy, která využívá metodu regrese, je nalézt co nejlepší, nejúspornější a současně věcně smysluplný model, který popíše vztah mezi závislou proměnnou a skupinou nezávislých proměnných. Je-li vysvětlovaná proměnná spojitá, obracíme se k regresi lineární, není-li spojitá, pak k regresi logistické. Metoda logistické regrese není omezená jen na případ, kdy vysvětlovaná proměnná je binární. I když pro tuto situaci byla logistická regrese původně vyvinuta a je interpretačně, ale i jinak nejsnazší. Existují však metody a také programy, které pracují s případy, kdy kategorizovaná závislá proměnná není binární, a dokáží respektovat požadavek, aby ji považovaly za ordinální.

3.1 Možnost využití logistické regrese v oblasti pojišťovnictví

S ohledem na výše uvedenou podstatu logistické regrese je šíře jejího využití v oblasti pojišťovnictví evidentní. Ať už se jedná o měření solventnosti pojišťoven, hospodaření pojišťoven, hodnocení úrovně pojistného trhu, odhad výše technických rezerv, stanovení pojistného v jednotlivých odvětvích pojištění a další, ve všech případech je možná snaha o vytvoření modelu a nalezení vztahu mezi konkrétními závislými a nezávislými proměnnými.

3.2 Formulace modelu

Předpokládejme, že máme binární veličinu Y_i charakterizující kladnou a zápornou variantu v kontextu vzniku pojistné události i-tého pojistníka, tedy

$$Y_i = \begin{cases} 1 \text{ pro pozitivní variantu (pojistná událost nastane),} \\ 0 \text{ pro negativní variantu (pojistná událost nenastane),} \end{cases} \text{pro } i = 1,...,n,$$

kde n je počet pojistníků. Každý tento pojistník je charakteristický vektorem $\mathbf{x}_i = (1, x_{1i}, x_{2i}, \ldots, x_{ki})$ obsahujícím k prvků, Strišš, Valečková, Valecký (2010, str. 206-207).

Pravděpodobnost vzniku pojistné události i-tého pojistníka $P_i = P(Y_i = 1)$ na základě jeho charakteristického vektoru $\mathbf{x}_i$ lze vyjádřit funkcí $F(\boldsymbol{\beta}; \mathbf{x}_i)$, jenž je monotónně rostoucí

$F'(\boldsymbol{\beta};\mathbf{x}_i)\geq 0$ a má definiční obor $(-\infty,\infty)$ a obor hodnot (0,1). Platí tedy, že $F(-\infty)=0$ a $F(+\infty)=1$ a funkci pravděpodobnosti odpovědi lze psát jako

$$P_i=F(\boldsymbol{\beta};\mathbf{x}_i), \tag{1}$$

kde $\boldsymbol{\beta}$ je vektor parametrů $(\beta_0, \beta_1,\ldots, \beta_k)$.

Tyto vlastnosti jsou splněny kumulativní distribuční funkcí logistického rozdělení ve tvaru

$$P_i = P(Y_i = 1) = F(\boldsymbol{\beta};\mathbf{x}_i) = \frac{e^{\beta'\mathbf{x}_i}}{1+e^{\beta'\mathbf{x}_i}}, \tag{2}$$

která je zároveň funkcí pravděpodobnosti vzniku pojistné události. Pravděpodobnost negativní varianty, nevzniknutí pojistné události, lze pak vyjádřit ve tvaru

$$1-P_i = P(Y_i = 0) = 1-F(\boldsymbol{\beta};\mathbf{x}_i) = \frac{1}{1+e^{\beta'\mathbf{x}_i}}. \tag{3}$$

Definujme dále podíl pravděpodobnosti vzniku a nevzniknutí pojistné události známé také jako šance (odds) ve tvaru

$$\frac{\pi}{1-\pi} = \frac{P(Y_i = 1)}{P(Y_i = 0)} = e^{\beta'\mathbf{x}_i}, \tag{4}$$

a dále tzv. logitovou transformaci (log-odds, logit) vztahu (4)

$$\ln\left[\frac{\pi}{1-\pi}\right] = \boldsymbol{\beta}'\mathbf{x}_i = g(\mathbf{x}_i). \tag{5}$$

Odhad parametrů modelu

K odhadu neznámých parametrů $\boldsymbol{\beta}$ je nejčastěji používána metoda maximální věrohodnosti. Tato metoda spočívá v nalezení věrohodnostní funkce $l(\ )$, která je posléze maximalizována. Mějme pravděpodobnost kladné odpovědi i-tého respondent charakteristického vektorem $\mathbf{x}_i$, tedy

$$P(Y_i = 1|\mathbf{x}_i) = \pi(\mathbf{x}_i), \tag{6}$$

a dále pravděpodobnost negativní varianty, nevzniknutí pojistné události

$$P(Y_i = 0|\mathbf{x}_i) = 1-P(Y_i = 1|\mathbf{x}_i) = 1-\pi(\mathbf{x}_i). \tag{7}$$

Sdružená pravděpodobnost kladných a záporných variant vzniku pojistné události lze poté vyjádřit ve tvaru

$$P(Y_i|\mathbf{x}_i) = \pi(\mathbf{x}_i)^{Y_i}\left[1-\pi(\mathbf{x}_i)\right]^{(1-Y_i)}. \tag{8}$$

Jsou-li jednotlivá pozorování nezávislá, pak věrohodnotstní funkce je určena jako součin sdružených pravděpodobností pro všechny pojistníky, tedy

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{N} \pi(\mathbf{x}_i)^{Y_i} \left[1 - \pi(\mathbf{x}_i)\right]^{(1-Y_i)}. \tag{9}$$

Odhad parametrů metodou maximální věrohodnoti je získán maximalizací logaritmu rovnice (9) ve tvaru

$$L(\boldsymbol{\beta}) = \ln l(\boldsymbol{\beta}) = \sum_{i=1}^{N} Y_i \cdot \ln(\pi(\mathbf{x}_i)) + (1 - Y_i) \cdot \ln(1 - \pi(\mathbf{x}_i)), \tag{10}$$

za podmínek

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} = 0 \text{ pro } j = 1,...,k. \tag{11}$$

Odhad metodou maximální věrohodnoti bývá prováděn pomocí iteračních algoritmů, přičemž nejčastěji je používána Newton-Raphsonova metoda. Princip této metody spočívá v aproximaci logaritmu věrohodnoctví funkce v okolí počátečního odhadu pomocí prvních tří členů Taylorova rozvoje, viz Pecáková (2007), přičemž počáteční odhad lze získat například metodou nejmenších čtverců.

4 Ověření vlivu vybraných faktorů na vznik pojistné události

V této části příspěvku jsou analyzovány vybrané faktory, a zjištěna jejich statistická významnost na dané hladině spolehlivosti.

Analýza je provedena pomocí dat získaných z pojistného kmene konkrétní pojišťovny a skládá se z datového vzorku 61 897 smluv. Smlouvy, tvořící pojistný kmen, jsou uzavřeny na produkt neživotního pojištění, pojištění majetku (vozu) ve smyslu havarijního pojištění. Jsou použita data za rok 2008, přičemž časová expozice smlouvy je jeden rok. V souboru jsou obsaženy údaje o velikosti a počtu škod, ceně vozu, věku a pohlaví pojistníka, o lokalitě bydliště pojistníka a o průměrném věku obyvatelstva v dané lokalitě, dále také o stáří dopravního prostředku, objemu motoru, výkonu dopravního prostředku, druhu spalovaného paliva, o uplatnění zvýhodněných balíčků, o způsobu používání vozidla či intenzitě jeho využívání a další. Zdrojem pro veškeré výstupy uvedené v této subkapitole je program STATA.

4.1 Popis jednotlivých zkoumaných znaků z datového souboru

Prvním z dostupných údajů v datovém souboru je nárok, *claim*. V datovém souboru je zachyceno, zda ke vzniku nároku dochází či nikoli. Dalšími zkoumanými znaky, uvedenými v datovém souboru jsou typ paliva (*fuel*), typ pojištění (*ins*), počet obyvatelstva v regionu (*nocit*), průměrný věk obyvatelstva v regionu (*avgagereg*), výše spoluúčasti (*excess*), stáří vozidla (*agecar*), věk pojistníka (*ageman*), pohlaví pojistníka (*gender*), využití dopravního prostředku k podnikání (*company*), výkon vozidla (*kw*) a objem motoru (*volume*).

Tabulka 1: Popis jednotlivých zkoumaných znaků z datového souboru

claim

```
              type:  numeric (byte)
             label:  Claim

             range:  [0,1]                        units:  1
       unique values:  2                       missing .:  0/61897

        tabulation:  Freq.   Numeric  Label
                     57146         0  no
                      4751         1  yes
```

fuel

```
              type:  numeric (byte)
             label:  Fuel

             range:  [0,4]                        units:  1
       unique values:  5                       missing .:  0/61897

        tabulation:  Freq.   Numeric  Label
                     43044         0  petrol
                     18334         1  diesel
                         8         2  p-butane
                        36         3  other
                       475         4  no fuel
```

ins

```
              type:  numeric (byte)
             label:  Ins_product

             range:  [0,2]                        units:  1
       unique values:  3                       missing .:  0/61897

        tabulation:  Freq.   Numeric  Label
                     61171         0  A
                       715         1  B
                        11         2  C
```

nocit

```
              type:  numeric (long)

             range:  [41255,1249026]              units:  1
       unique values:  77                       missing .:  780/61897

              mean:   527387
         std. dev:   519558

        percentiles:        10%       25%       50%       75%       90%
                          92903    116730    186681   1.2e+06   1.2e+06
```

avgagereg

```
              type:  numeric (float)

             range:  [38.2,42]                    units:  .1
       unique values:  29                       missing .:  780/61897

              mean:   40.8283
         std. dev:   .820363

        percentiles:        10%       25%       50%       75%       90%
                          39.8      40.3      40.9      41.6      41.6
```

excess

```
              type:  numeric (byte)

             range:  [5,30]                       units:  1
       unique values:  5                        missing .:  0/61897

        tabulation:  Freq.  Value
                     46166  5
                     13736  10
                      1047  15
                       908  20
                        40  30
```

agecar

type:	numeric (byte)				
range:	[0,43]		units:	1	
unique values:	31		missing .:	0/61897	
mean:	4.89623				
std. dev:	3.23353				
percentiles:	10%	25%	50%	75%	90%
	1	2	4	7	10

ageman

type:	numeric (int)				
range:	[0,99]		units:	1	
unique values:	81		missing .:	0/61897	
mean:	32.0119				
std. dev:	26.0075				
percentiles:	10%	25%	50%	75%	90%
	0	0	38	55	64

gender

type:	numeric (byte)		
label:	Gender		
range:	[0,1]	units:	1
unique values:	2	missing .:	0/61897

tabulation:	Freq.	Numeric	Label
	48789	0	male
	13108	1	female

company

type:	numeric (byte)		
label:	Company		
range:	[0,1]	units:	1
unique values:	2	missing .:	0/61897

tabulation:	Freq.	Numeric	Label
	22003	0	yes
	39894	1	no

kw

type:	numeric (int)				
range:	[1,850]		units:	1	
unique values:	278		missing .:	475/61897	
mean:	72.1535				
std. dev:	33.3764				
percentiles:	10%	25%	50%	75%	90%
	44	50	65	83	110

volume

type:	numeric (int)				
range:	[49,15928]		units:	1	
unique values:	708		missing .:	475/61897	
mean:	1664.57				
std. dev:	667.182				
percentiles:	10%	25%	50%	75%	90%
	1198	1289	1498	1896	2198

Dle hodnot uvedených v tabulce 1 je možné konstatovat, že v rámci sledovaného souboru dat u převážné většiny pojištěných dopravních prostředků nedochází ke vzniku pojistného nároku. Převážná většina dopravních prostředků při svém provozu spaluje benzín. V početně nejvyšším zastoupení je sjednáno pojištění dopovídající stupni A, což představuje pojištění kryjící široký rozsah rizik. Průměrný počet obyvatel v regionech (okresech) činí 527 387 osob, přičemž minimální počet obyvatel v regionu byl zjištěn v počtu 41 255 obyvatel a maximální počet obyvatel v regionu činí 1 249 026 obyvatel. Průměrný věk obyvatelstva se pohybuje v rozmezí 38,2 až 42 let. Výše spoluúčasti pojistníka na případném pojistném plnění je

v rámci sledovaného souboru dat udávána v rozmezí pěti až třiceti procent, přičemž častější je výskyt spoluúčasti ve výši pěti procent. Nejstarším vozidlem v souboru je dopravní prostředek ve stáří 43 let, opakem je pak zcela nový vůz, jehož stáří nedosahuje ani jednoho roku. Průměrné stáří pojištěných vozů je 4,9 let. Maximální věk pojistníka činí 99 let, přičemž průměrný věk činí 32 let. Zcela zřetelné je rozdělení pojistníků dle pohlaví, kdy zastoupení mužů je více než dvoutřetinové. Většina pojištěných vozů není používána k podnikání. Průměrný výkon pojištěných vozů je vyčíslen ve výši 72,15 kw a průměrný objem motoru je 1 664,57 cm^3.

4.2 Jednofaktorová analýza a test významnosti kategorií

V rámci této podkapitoly je postupně hodnocen vztah jednotlivých vysvětlujících proměnných vzhledem k vysvětlované proměnné a to za předpokladu, že zbylé vysvětlující veličiny nabývají nulových hodnot. Je-li proměnná rozdělena na podkategorie, pak jsou dílčí kategorie zahrnuty souběžně.

Tabulka 2: Jednofaktorová analýza a test významnosti kategorií

claim	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
fuel_2	.4609755	.0311271	14.81	0.000	.3999674	.5219836
fuel_3	(omitted)					
fuel_4	.2395702	.6033318	0.40	0.691	-.9429383	1.422079
fuel_5	-1.565079	.3812673	-4.10	0.000	-2.312349	-.8178085
_cons	-2.637465	.0193091	-136.59	0.000	-2.675311	-2.59962

claim	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
ins_2	-.2890097	.1597675	-1.81	0.070	-.6021482	.0241289
ins_3	2.302744	.6057202	3.80	0.000	1.115555	3.489934
_cons	-2.485066	.0151739	-163.77	0.000	-2.514806	-2.455326

claim	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
nocit	2.68e-07	2.84e-08	9.42	0.000	2.12e-07	3.24e-07
_cons	-2.640469	.0226626	-116.51	0.000	-2.684887	-2.596052

claim	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
avgagereg	.1152978	.0192607	5.99	0.000	.0775476	.153048
_cons	-7.20193	.787755	-9.14	0.000	-8.745901	-5.657958

claim	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
excess2	-.2210374	.038222	-5.78	0.000	-.2959511	-.1461236
excess3	-.9151291	.1704576	-5.37	0.000	-1.24922	-.5810383
excess4	-.9224725	.1835408	-5.03	0.000	-1.282206	-.5627391
excess5	-1.243525	1.012889	-1.23	0.220	-3.22875	.7416996
_cons	-2.42005	.0169952	-142.40	0.000	-2.45336	-2.38674

claim	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
agecar	-.0618286	.0049914	-12.39	0.000	-.0716115	-.0520456
_cons	-2.200667	.0265733	-82.81	0.000	-2.25275	-2.148584

claim	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
ageman	-.0095528	.0005848	-16.33	0.000	-.0106991	-.0084065
_cons	-2.20743	.0216794	-101.82	0.000	-2.249921	-2.164939

claim	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
gender	-.0797871	.0376983	-2.12	0.034	-.1536744	-.0058998
_cons	-2.470628	.0168872	-146.30	0.000	-2.503726	-2.437529

claim	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
company	-.4414676	.0304382	-14.50	0.000	-.5011254	-.3818098
_cons	-2.221907	.0226949	-97.90	0.000	-2.266388	-2.177426

claim	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
kw	.0073012	.0003558	20.52	0.000	.0066038	.0079985
_cons	-3.035043	.0322253	-94.18	0.000	-3.098204	-2.971883

claim	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
volume	.0002298	.0000161	14.28	0.000	.0001983	.0002613
_cons	-2.874297	.0321697	-89.35	0.000	-2.937348	-2.811246

Dle výše uvedených výstupů je považováno za vhodné začlenit do odhadovaného modelu téměř všechny nezávislé veličiny. Některé z podkategorií se sice jeví jako nevýznamné, ale odstranění takových vybraných podkategorií z modelu není možné, neboť je vždy nutné současné zahrnutí veškerých dílčích podkategorií.

5 Závěr

Cílem příspěvku bylo ověření vlivu vybraných faktorů na vznik pojistné události. Tedy zjištění, zda jsou vybrané faktory určujícími pro vznik pojistné události či nikoli a tento vztah dále kvantifikovat a to pomocí logistické regrese na konkrétním příkladu datového vzorku smluv konkrétního pojistitele. V článku byla nejprve definována a vysvětlena regresní analýza a stručně byl popsán regresní model. Následně byla zmíněna podstata logistické regrese, včetně odhadu koeficientů metodou maximální věrohodnosti a poté byla provedena analýza vlivu vybraných faktorů na vznik pojistné události ve smyslu havárie motorového vozidla.

Z výše uvedených výsledků vyplývá, že vznik pojistné události je determinován typem paliva, typem pojištění, počtem obyvatelstva v regionu, průměrným věkem obyvatelstva v regionu, výší spoluúčasti, stáří vozidla, věkem pojistníka, pohlavím pojistníka, využitím dopravního prostředku k podnikání, výkonem vozidla a objemem motoru.

References

[1] BOROVCOVÁ, M. 2011. Application possibilities of the logistic regression in the insurance sector. *Financial management of firms and financial institutions*. VŠB-TU Ostrava, pp. 32-39.

[2] BOROVCOVÁ, M. 2011. Analýza vlivu vybraných faktorů na vznik pojistné události. *Aktuárska veda v teórii a praxi*. Ekonomická univerzita v Bratislave, pp. 6-11.

[3] FOJTÍKOVÁ, A. 2012. *Konstrukce modelu stanovení pojistného na bázi metody regresní analýzy*. Diplomová práce. VŠB-TU Ostrava.

[4] HARDIN, J. W., HILBE J. M., 2007. *Generalized Linear Models and Extensions*. Texas: Stata Press.

[5] HOSMER, D.W., LEMESHOW, S., 2000. *Applied Logistic Regression.* New Jersey: John Wiley & Sons.

[6] PECÁKOVÁ, I., 2007. Logistická regrese s vícekategoriální vysvětlovanou proměnnou. *Acta Oeconomica Pragensia*, roč. 15, č. 1, pp. 86-96.

[7] STRIŠŠ, J., VALEČKOVÁ, J., VALECKÝ, J., 2010. Aplikace logistické regrese v měření spokojenosti zákazníků. *Rozvoj marketingu v teórii a praxi*, Žilinská univerzita v Žilině, pp. 205-210.

[8] ŠIMURDA, M., 2008. Zobecněný lineární model (GLM). Dostupné na: http://www.actuaria.cz/upload/GLM_SMM_MFF_web.pdf.

[9] WEISBERG, S., 2005. *Applied Linear Regression.* New Jersey: John Wiley & Sons.

[10] ZMEŠKAL, Z., 2004. *Finanční modely*. Praha: EKOPRESS.

Summary

Příspěvek je zaměřen na možnosti aplikace metody logistické regrese v oblasti pojišťovnictví, konkrétně její využití při ověření vlivu vybraných faktorů na vznik pojistné události. Obecně jsou definovány oblasti možného využití logistické regrese, podrobněji je analyzována oblast neživotního pojištění majetku, konkrétně havarijní pojištění. Pomocí logistické regrese jsou identifikovány klíčové faktory ovlivňující pravděpodobnost vzniku pojistné události, přičemž jsou analyzovány vlivy binárních, kategoriálních i spojitých veličin. Vlivy vybraných faktorů jsou v rámci jednorozměrné analýzy rovněž kvantifikovány.