Univerzita Hradec Králové

Přírodovědecká fakulta

bakalářská práce

2020 Martin Stolín

Univerzita Hradec Králové

Přírodovědecká fakulta

Katedra matematiky

Analýza přežití na datech v pojišťovnictví

bakalářská práce

Autor: Martin Stolín

Studijní program: Aplikovaná matematika

Studijní obor: Finanční a pojistná matematika

Vedoucí práce: Mgr. Jitka Kühnová, Ph.D.

Hradec Králové **květen 2020**

Poděkování:

Na tomto místě bych rád poděkoval Mgr. Jitce Kühnové, Ph.D., vedoucí mé bakalářské práce, za její ochotu, cenné rady a připomínky. Zároveň děkuji společnosti UNIQA pojišťovna, a.s. za poskytnutí potřebných dat.

Prohlášení:

Prohlašuji, že jsem bakalářskou práci vypracoval samostatně a že jsem v seznamu použité literatury uvedl všechny prameny, ze kterých jsem vycházel.

V Hradci Králové dne 13. května 2020 Martin Stolín

Anotace

STOLÍN, M. *Analýza přežití na datech v pojišťovnictví.* Hradec Králové, 2020. Bakalářská práce na Přírodovědecké fakultě Univerzity Hradec Králové. Vedoucí bakalářské práce Mgr. Jitka Kühnová, Ph.D. 36 s.

Práce se bude zabývat analýzou přežití. Nejprve bude popsána základní teorie (metody cenzorování, odhady funkce přežití) potřebná ke zpracování dat. Následně budou popsané poznatky aplikovány na reálných datech z oblasti pojišťovnictví.

Klíčová slova
Funkce přežití, Kaplanův-Meierův odhad, Log-rank test, Coxův model proporcionálních rizik

Annotation

STOLÍN, M. *Survival analysis on insurance data.* Hradec Králové, 2020. Bachelor Thesis at Faculty of Science University of Hradec Králové. Thesis Supervisor Mgr. Jitka Kühnová, Ph.D. 36 p.

The work will deal with the survival analysis. First, the basic theory (censoring methods, estimates of survival function) needed for data processing will be described. Subsequently, the described knowledge will be applied to real data from the field of insurance.

Keywords
Survival function, Kaplan-Meier estimate, Log-rank test, Cox Proportional Hazards Model

Obsah

Úvod

Analýza přežití je analýza tzv. *time-to-event* dat. Taková data popisují dobu od počátečního času po sledovaný konečný bod. Analýza se tak stala součástí mnoha oborů, příklady použití najdeme především v medicíně, biologii, sociologii, technických oborech, pojišťovnictví a bankovnictví.

Cílem této práce je čtenáři vyložit základní problematiku a používané metody. Teoretické poznatky budou doprovázeny ilustracemi a v závěru práce aplikovány na rozsáhlou skupinu dat z pojišťovny.

Práce je rozdělená do pěti kapitol. V první kapitole se budeme věnovat principu analýzy přežití, představíme cenzorování dat, díky kterému se analýza přežití odlišuje od ostatních oblastí statistiky. Zadefinujeme tři základní funkce, které charakterizují rozdělení času přežití a znázorníme vzájemné vztahy mezi funkcemi.

Druhá kapitola je věnována odhadům rozdělení času přežití pomocí funkce přežití. Mezi nejpoužívanější metody řadíme Kaplanovu-Meierovu metodu, či úmrtnostní tabulky.

Ve třetí kapitole se zaměříme na testy, které porovnávají rozdělení časů přežití ve skupinách dat.

Čtvrtá kapitola se zabývá základím představením Coxova modelu proporcionálních rizik, který našel uplatnění i mimo oblast, ve které vznikl – biostatistiku. V bankovnictví bývá zahrnut v credit-scoringových modelech.

Poslední kapitola je věnována aplikaci teoretických poznatků na data z pojišťovny. V kapitole zahrneme příkazy, které byly použity k analýze v softwaru RStudio.

1 Analýza přežití

Metody analýzy přežití se obvykle používají k analýze údajů shromážděných prospektivně (zaměřujících se na budoucnost) v čase, jako jsou data z prospektivní kohortové studie nebo data shromážděná pro klinické hodnocení.

Příkladem dat přežití mohou být dva soubory pacientů s určitým onemocněním a totožným podílem zemřelých po deseti letech od diagnózy, u nichž však přežití může mít naprosto odlišný průběh, kdy jedna skupina vykazuje vyšší úmrtnost brzy po začátku sledování (např. zahájení léčby) s následnou klesající tendencí, zatímco druhá skupina naopak vykazuje nižší úmrtnost na začátku sledování a její nárůst v průběhu času. Z toho plyne, že pro analýzu přežití není podstatná událost (smrt), ale čas, který uplynul do této události. Tento čas je potřeba definovat.

Čas přežití je doba, která uplynula do výskytu pozorované události. Značíme ji T. Aby mohl být jednoznačně určen čas přežití, je potřeba specifikovat počáteční čas tak, aby subjekty byly sledovány od určité události (charakteristické pro všechny subjekty). Například, pokud se studuje doba přežití pacientů s konkrétním typem rakoviny, mohl by být počáteční čas zvolen jako časový bod diagnózy tohoto typu rakoviny. Stejně důležité je, aby byl vhodně definován konečný bod nebo událost. Ve výše uvedeném příkladu by to mohla být smrt kvůli studované rakovině.

Jedním z důvodů, proč analýza přežití vyžaduje „speciální“ techniky (data nejsou vhodná k běžnému statistickému použití), je asymetričnost v rozdělení a kladné zešikmení. Je to způsobeno časem přežití, který nabývá pouze nezáporných hodnot. Proto není vhodné předpokládat, že taková data mají normální rozdělení.

Druhým důvodem je časté cenzorování dat. To znamená, že čas T je sledován částečně. Například jednotlivci mohou ze studie odstoupit, nebo se může stát jiná událost, například ve výše uvedeném příkladu smrt v důsledku nehody, která není součástí sledovaného parametru. Jinou možností je, že nastane čas konce studie a u některých subjektů nenastane událost, tudíž nebude jejich čas pozorován. Tato neúplná pozorování nelze ignorovat, ale je třeba s nimi zacházet odlišně.

1.1 Cenzorování a krácení dat

Rozlišujeme více druhů cenzorování. Nejběžnějším typem cenzury a nejjednodušší manipulací v analýze je cenzorování zprava.

Definice 1.1. Nechť T_i, $i = 1, \ldots, n$ je nezávislá doba přežití náhodné veličiny T a C_i, $i = 1, \ldots, n$, cenzorovaný čas přežití nezávislý na T_i. Doba přežití u i-tého subjektu je známa, pokud $T_i \leq C_i$ a *cenzorována zprava*, pokud $C_i < T_i$

Pozorovaný čas reprezentujeme pomocí dvojice náhodných veličin (t_i, δ_i), kde $t_i = \min(T_i, C_i)$ a $\delta_i = I(T_i < C_i)$. Náhodná veličina I udává hodnoty $(0, 1)$, kde pro $\delta_i = 1$ událost nastala a $\delta_i = 0$ událost cenzorována.

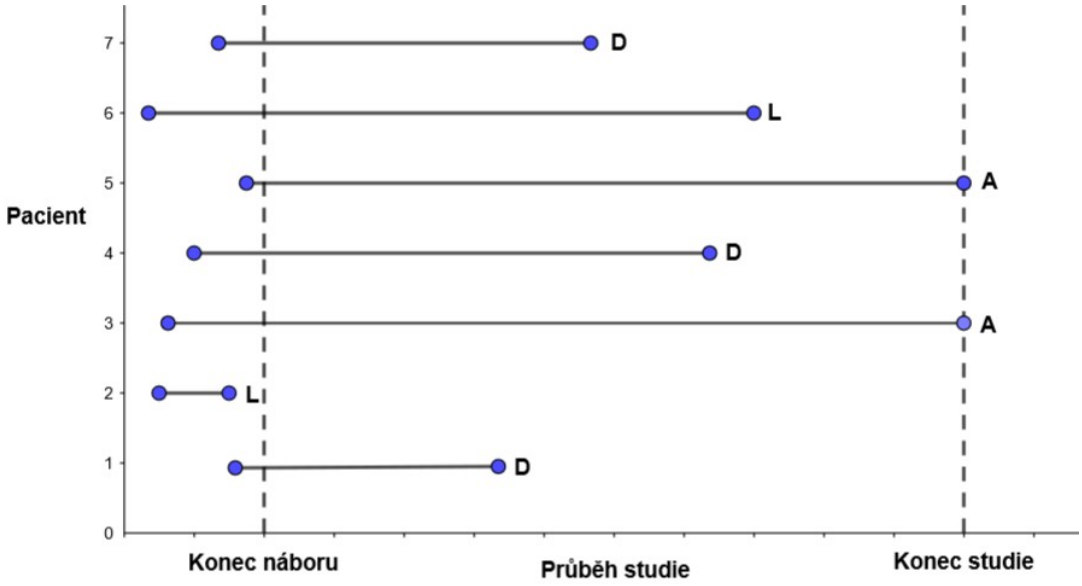
Dalším typem cenzorování je cenzorování zleva. Je to situace, kdy u subjektu nastala událost ještě před počátečním časem. S tímto problémem se setkáváme zřídkakdy.

Definice 1.2. Nechť T_i, $i = 1, \ldots, n$ je nezávislá doba přežití náhodné veličiny T a C_i, $i = 1, \ldots, n$, cenzorovaný čas přežití nezávislý na T_i. Doba přežití u i-tého subjektu je známa, pokud $T_i \geq C_i$ a *cenzorována zprava*, pokud $C_i > T_i$.

Opět pozorovaný čas reprezentujeme pomocí dvojice náhodných veličin (t_i, δ_i), kde $t_i = \max(T_i, C_i)$ a $\delta_i = I(T_i > C_i)$. Náhodná veličina I udává hodnoty $(0, 1)$, kde pro $\delta_i = 1$ událost nastala a $\delta_i = 0$ událost cenzorována.

Třetím typem je intervalové cenzorování, kdy je známo, že mezi dvěma po sobě jdoucími časy se vyskytla sledovaná událost. Příkladem jsou předem plánovaná pozorování ve studiích, protože subjekt může na další kontrolu přijít se změněným stavem. Z toho víme, že nastalá událost je větší než poslední doba pozorování (ve které ke změně nedošlo) a menší nebo rovna pozorovanému času ve kterém byla zjištěna událost. Podrobnější informace lze nalézt ve článku [2] a [5].

Pro představu cenzorování je použit obrázek 1.1. Na vodorovné ose je vyobrazen čas a na svislé jednotliví pacienti. Úsečka u jednotlivého pacienta představuje délku času přežití. Označení písmeny u každého pacienta vyjadřuje určitou informaci. Pacienti 1, 4 a 7, označení písmenem D jsou mrtví – nastala událost. Pacienti 2 a 6 označeni písmenem L byli "ztraceni", v průběhu studie ukončili spolupráci – byli cenzorováni. Pacienti 3 a 5 byli ke konci studie stále naživu (alive). Událost u nich mohla nastat po konci studie (to ovšem nevíme), proto i tito pacienti budou ve studii cenzorováni.



Obrázek 1.1: Příklad pro cenzorování zprava

1.2 Funkce přežití a riziková funkce

Čas přežití je náhodná veličina, která může nabývat pouze nezáporných hodnot. Rozdělení času přežití lze charakterizovat pomocí tří základních funkcí. První z nich je funkce přežití, poté riziková funkce a hustota pravděpodobnosti. Každá z těchto funkcí zobrazuje různé vlastnosti dat.

Předpokládejme, že náhodná veličina T má rozdělení pravděpodobnosti s hustotou $f(t)$. Ta nám udává pravděpodobnost výskytu sledované události v čase t, respektive v daném časovém intervalu na reálné ose.

$$f(t) = \lim_{\Delta t \to 0_+} \frac{P\left(t \leq T < t + \Delta t\right)}{\Delta t} \tag{1.1}$$

Distribuční funkce, tedy funkce udávající pravděpodobnost toho, že čas přežití bude menší než hodnota $t \in \langle 0, \infty)$, je pak definována jako

$$F(t) = P(T < t) = \int_0^t f(u)\mathrm{d}u. \tag{1.2}$$

Definice 1.3. *Funkce přežití (survival function)* náhodné veličiny T je definována jako

$$S(t) = P(T \geq t). \tag{1.3}$$

Funkce přežití je tedy pravděpodobnost, že subjekt přežije dobu větší nebo rovnou t. Z předchozího vztahu tedy plyne, že funkce přežití je doplňkem distribuční funkce. Lze ji

tedy vyjádřit jako

$$S(t) = 1 - F(t), \tag{1.4}$$

Ve spojitém případě pak

$$S(t) = \int_t^\infty f(u)\mathrm{d}u.$$

Funkce $S(t)$ nabývá hodnot mezi 0 a 1, kdy hodnotu 1 nabývá v čase $t(0)$ a hodnoty 0 nabývá limitně v nekonečnu. Funkce přežití je tedy nerostoucí funkce. Grafem funkce $S(t)$ je nazývána křivka přežití.

Definice 1.4. *Riziková funkce (hazard function)* nám udává pravděpodobnost, že jedinec zemře v čase t za podmínky, že subjekt přežil do času t. Riziková funkce představuje momentální úmrtnost. Pro náhodnou veličinu T definujeme jako

$$h(t) = \lim_{\Delta t \to 0_+} \frac{P\left[t \leq T < t + \Delta t \mid t \leq T\right]}{\Delta t}. \tag{1.5}$$

Riziková funkce může být rostoucí, klesající, konstantní, nebo může indikovat složitější proces. Riziková funkce jako okamžitá intenzita výskytu sledované události je velmi důležitou funkcí zejména v modelování přežití, nicméně pro věcný popis dosahovaného přežití v souboru subjektů je praktická spíše její kumulativní varianta, tedy kumulativní riziková funkce.

Definice 1.5. *Kumulativní rizikovou funkci (cumulative hazard function)* pro spojitou náhodnou veličinu T definujeme jako

$$H(t) = \int_0^t h(u)\mathrm{d}u \tag{1.6}$$

1.3 Vzájemné vztahy mezi funkcemi

Všechny výše definované funkce popisující pravděpodobnostní chování náhodné veličiny jsou matematicky ekvivalentní, neboť při znalosti jedné z nich lze dopočítat ostatní. Vzájemné výpočetní vztahy lze odvodit následující úvahou. Součástí definice rizikové funkce je podmíněná pravděpodobnost, respektive pravděpodobnost výskytu sledované události v intervalu $t + \Delta t$ za podmínky, že k ní nedošlo do času t.

Uvědomíme-li si, že $P(A \mid B)$ lze pomocí věty o podmíněné pravděpodobnosti vyjádřit jako $P(A \cap B)/P(B)$, můžeme podmíněnou pravděpodobnost v definici rizikové funkce

vyjádřit pomocí následujícího vztahu:

$$\frac{P(t \leq T < t + \Delta t)}{P(T \geq t)}.$$

Dostáváme:

$$h(t) = \lim_{\Delta t \to 0_+} \frac{\frac{P(t \leq T < t+\Delta t)}{P(T \geq t)}}{\Delta t} = \frac{\lim\limits_{\Delta t \to 0_+} \frac{P(t \leq T < t+\Delta t)}{\Delta t}}{P(T \geq t)}.$$

Pomocí funkce (1.1), která je ekvivalentní čitateli, a funkce (1.3) lze tento výraz přepsat jako

$$h(t) = \frac{f(t)}{S(t)}. \tag{1.7}$$

Dále podle vztahu (1.4) víme, že pravděpodobnostní hustota se rovná derivaci distribuční funkce

$$f(t) = \frac{\mathrm{d}}{\mathrm{d}t}\left[1 - S(t)\right] = -S'(t). \tag{1.8}$$

Dosazením tohoto výrazu do rovnice (1.7) získáme

$$h(t) = \frac{S'(t)}{S(t)} = -\left[\log S(t)\right]'. \tag{1.9}$$

Pokud tento výsledek dosadíme do definice kumulativní rizikové funkce dané vztahem (1.6), získáme vzorec dokumentující přímou souvislost funkce přežití a kumulativní rizikové funkce, který má tvar

$$H(t) = -\log S(t) \tag{1.10}$$

nebo

$$S(t) = \exp\{-H(t)\} = \exp\left\{-\int_0^t h(x)\mathrm{d}x\right\}. \tag{1.11}$$

Využitím rovnice (1.7) a (1.11) získáme

$$f(t) = h(t)\exp\{-H(t)\}. \tag{1.12}$$

Pro přehled jsou v tabulce 1.1 uvedeny vztahy (viz [7]).

$f(t)$	$=$	$-S'(t) = h(t)\exp\{-H(t)\}$	
$S(t)$	$=$	$1 - F(t) =$	$\exp\{-H(t)\}$
$h(t)$	$= -\left[\log S(t)\right]' =$		$\frac{f(t)}{S(t)}$
$H(t)$	$= -\log S(t) =$		

Tabulka 1.1: Vztahy pro různé funkce

2 Odhad funkce přežití

Statistické metody lze rozdělit na parametrické a neparametrické. Parametrické metody (parametric survival analysis) jsou metody, pro jejichž odvození je nutné specifikovat rozdělení náhodné veličiny T, zatímco neparametrické metody (nonparametric survival analysis) předpoklady ohledně rozdělení náhodné veličiny T nevyžadují. V analýze přežití se vyskytuje ještě třetí metoda, zvaná semiparametrická (semiparametric survival analysis). Jedná se o modelovací přístupy, které nejsou plně parametrické, protože nevyžadují předpoklad o znalosti rozdělení veličiny, nicméně jakožto modely s parametry, respektive regresními koeficienty, pracují. [6]

2.1 Parametrické modely

Parametrické metody *(parametric survival analysis)* jsou metody, které vyžadují specifikaci konkrétního rozdělení náhodné veličiny T. V analýze přežití je specifikace rozdělení velmi silný předpoklad, to má své výhody i nevýhody. Jestliže špatně zvolíme rozdělení, nebude se shodovat s napozorovanými daty a odhady též budou špatné. Výhodou bývá jednodušší odhad mediánu funkce přežití, nižší variabilita nebo přesnější odhad oproti Kaplanovu-Meierovu (kapitola (2.2.1)). Jedním ze základních, běžně používaných rozdělení v analýze přežití je např. Exponenciální nebo Weibullovo.

2.1.1 Exponenciální rozdělení

Exponenciální rozdělení vyjadřuje rozdělení délky časových intervalů mezi výskyty událostí, pokud se tyto události vyskytují nezávisle na sobě. Příkladem může být ve strojírenství použití rozdělení k měření času spojeného s přijetím vadné součásti na montážní lince. Ve financích se především často používá k měření pravděpodobnosti příštího selhání pro portfolio finančních aktiv.

Předpokládejme, že $T \sim Exp(\lambda)$, kde $\lambda > 0$. Hustota je pak definována jako

$$f(t) = \begin{cases} \lambda \exp\{-\lambda t\} & \text{kde } t \geq 0, \\ 0 & \text{kde } t < 0 \end{cases}.$$

Pomocí vztahů vyjádřených v kapitole 1.3 lze odvodit funkci přežití a rizikovou funkci

$$S(t) = \exp\{-\lambda t\}$$
$$h(t) = \lambda.$$

2.1.2 Weibullovo rozdělení

Předpokládejme, že $T \sim Weib(\lambda, \gamma)$ s parametry $\lambda > 0, \gamma > 0$. Weibullovo rozdělení pravděpodobnosti je zobecněním exponenciálního rozdělení, které navrhl Weibull pro popis životnosti materiálů. Na rozdíl od exponenciálního Weibullovo rozdělení nepředpokládá konstantní riziko výskytu sledované události v čase, ale uvažuje monotónní rizikovou funkci (tedy s časem monotónně rostoucí nebo klesající funkci), z čehož plyne také jeho širší uplatnění v praxi.

$$f(t) = \begin{cases} \lambda\gamma(\lambda t)^{\gamma-1} \exp\{-(\lambda t)^{\gamma}\} & \text{kde } t \geq 0, \\ 0 & \text{kde } t < 0 \end{cases}$$

Pomocí vztahů vyjádřených v kapitole 1.3 lze odvodit funkci přežití a rizikovou funkci

$$S(t) = \exp\{-(\lambda t)^{\gamma}\},$$
$$h(t) = \lambda\gamma(\lambda t)^{\gamma-1}. \tag{2.1}$$

Parametr γ určuje tvar rozdělení a parametr λ škálování. Z rovnice (2.1) lze vypozorovat, že riziková funkce silně závisí na parametru γ. Platí

- pro $\gamma < 1$ je riziková funkce náhodné veličiny T klesající,
- pro $\gamma = 1$ je riziková funkce náhodné veličiny T konstantní,
- pro $\gamma > 1$ je riziková funkce náhodné veličiny T rostoucí.

2.2 Neparametrické modely

Neparametrické metody jsou velmi jednoduché metody, které nekladou žádné nároky na rozdělení náhodné veličiny T. Neparametrické modely jsou užitečné pro shrnutí údajů o přežití a pro jednoduchá srovnání, ovšem nikoli pro komplexnější situace. Nevýhodou oproti parametrickým metodám je nepřesnost ve výsledcích, už z principu absence znalosti o rozdělení dat. Nejznámnějším a nejpoužívanějším neparametrickým modelem pro určení funkce přežití se pro jeho jednoduchost stal Kaplanův-Meierův odhad funkce přežití (viz zdroj [3], [1]). Jako dalším odhadem funkce přežití se budeme zabývat odhadem pomocí úmrtnostních tabulek.

2.2.1 Kaplanův-Meierův odhad

Kaplanovu-Meierovu křivku lze použít v jednoduchých analýzách, jejichž cílem je porovnat doby přežití dvou nebo více skupin. Tato metoda je vhodná pro cenzorovaná data.

Nechť $t_1 < t_2 < \cdots < t_k$ jsou seřazené napozorované časy a n velikost vzorku dat. Nechť d_j je počet událostí v čase t_j, kde $j = 1, \ldots, k$ a m_j je počet cenzorovaných subjektů v intervalu $\langle t_j, t_{j+1})$. Potom $n_j = (m_j + d_j) + \cdots + (m_k + d_k)$ je počet subjektů v riziku před časem t_j. Časový interval $\langle t_j, t_{j+1})$ zahrnuje alespoň jednu událost. Protože je zde n_j přeživších před časem t_j a d_j událostí v čase t_j, pravděpodobnost nastání události během intervalu $\langle t_j, t_{j+1})$ je odhadnuta jako d_j/n_j. Odhad pravděpodobnosti přežití je potom $(n_j - d_j)/n_j$. Předpokládáme, že jednotlivé sledované události jsou na sobě nezávislé. Potom Kaplanův-Meierův odhad funkce přežití je definovaný jako součin jednotlivých odhadů pravděpodobností přežití do času t_j:

$$\hat{S}(t) = \hat{P}[T > t] = \prod_{j:\, t_j \leq t} \frac{n_j - d_j}{n_j} = \prod_{j:\, t_j \leq t} \left(1 - \frac{d_j}{n_j}\right). \tag{2.2}$$

Kaplanův-Meierův odhad $\hat{S}(t)$ je kroková funkce se skoky v čase události t_j. Velikost kroků závisí na počtu událostí d_j a počtu ohrožených osob n_j v odpovídajícím čase. Z výše uvedeného zavedení je vidět, že cenzorovaná data skok nezpůsobují, ovšem navyšují počet osob n_j. Pokud jsou poslední data cenzorovaná, odhad funkce se nedostane k 0. Pokud by se v datech nevyskytovala cenzorovaná pozorování, pak Kaplan-Meierův odhad je totožný s tzv. *empirickým odhadem*, který je jednodušší, ale pro většinu dat přežití nepostačující.

Jako ukázkový příklad pro Kaplanův-Meierův odhad jsem zvolil dataset „ovarian“ z programu R. Vychází z randomizované studie porovnávající dva typy léčby rakoviny vaječníku na 26 pozorovaných subjektech. Zde jsou uvedeny další klinické informace:

- **number** - číslo pacienta
- **futime** - doba do smrti, nebo cenzorování
- **fustat** - údaj o tom, zda byl pacient cenzorován
- **age** - věk v letech
- **resid.ds** - přítomnost zbytkové nemoci
- **rx** - léčebná skupina
- **ecog.ps** - výkonostní stav

Pro zjednodušení je vidět v tabulce 2.1 výběr pacientů ze skupiny A. Tito pacienti jsou již seřazeni podle času přežití.

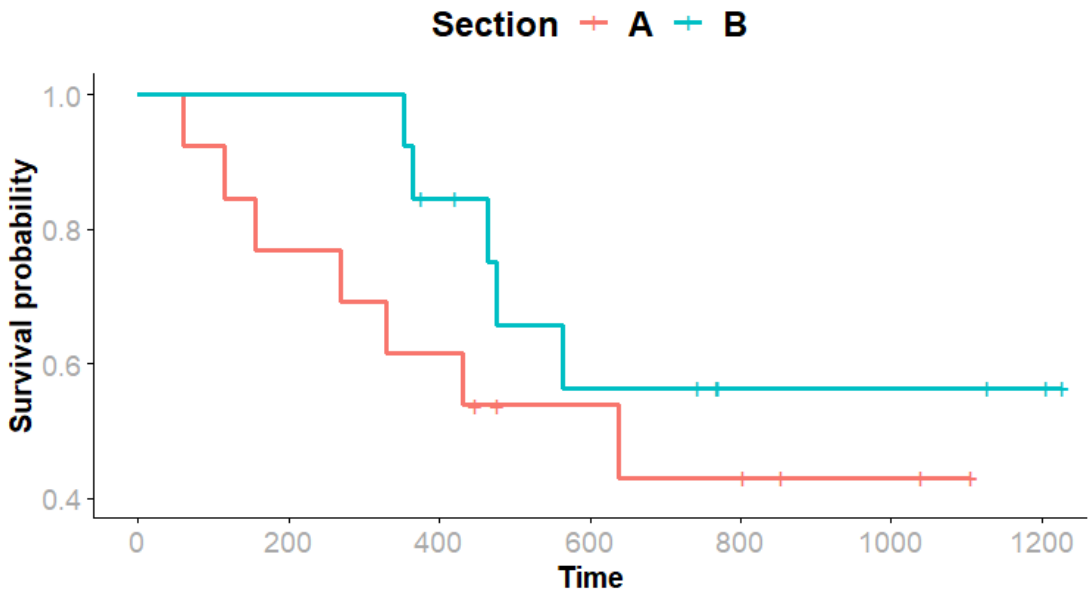
number	futime	fustat	age	resid.ds	rx
1	59	1	72.3315	yes	A
2	115	1	74.4932	yes	A
3	156	1	66.4658	yes	A
22	268	1	74.5041	yes	A
23	329	1	43.1370	yes	A
5	431	1	50.3397	yes	A
6	448	0	56.4301	no	A
9	477	0	64.1753	yes	A
11	638	1	56.7562	no	A
15	803	0	39.2712	no	A
16	855	0	43.1233	no	A
17	1040	0	38.8932	yes	A
18	1106	0	44.6000	no	A

Tabulka 2.1: Výčet pacientů ze skupiny A

Odhad funkce přežití se počítá postupně podle vzorce (2.2).
Takto získané hodnoty zakreslíme do grafu spolu se skupinou B (Obrázek 2.1).

čas t_i	n_i	d_i	$\hat{S}(t_i)$
59	13	1	0.9230769
115	12	1	0.8461538
156	11	1	0.7692308
268	10	1	0.6923077
329	9	1	0.6153846
431	8	1	0.5384615
448	7	0	0.5384615
477	6	0	0.5384615
638	5	1	0.4307692
803	4	0	0.4307692
855	3	0	0.4307692
1040	2	0	0.4307692
1106	1	0	0.4307692

Tabulka 2.2: Hodnoty odhadu funkce přežití skupiny A - K-M odhad



Obrázek 2.1: Kaplanův-Meierův odhad skupin A a B

2.2.2 Úmrtnostní tabulky

Tabulka úmrtnosti, nebo také *life-table*, je jedna z nejstarších a nejpoužívanějších statistických metod pro zjištění úmrtnosti. Tabulky úmrtnosti se v pojišťovnictví často používají ke stanovení pojistného.

Cílem je podobně jako u Kaplanova-Meierova odhadu, vyjádřit odhad funkce přežití jako součin podmíněných pravděpodobností v určitých intervalech rozdělením doby pozorování. Většinou se tato metoda používá, když neznáme přesný vznik události.

Protože pracujeme s delšími časovými intervaly, budou nám stačit souhrnné údaje pro jednotlivé časové intervaly. Označme d_j počet sledovaných událostí v j-tém intervalu, kde $j = 1, \ldots, k$, který se rozkládá od času t_j k t_{j+1}. Následně c_j označme jako počet cenzorovaných dob přežití. Předpokládáme, že cenzorované doby přežití se v daném intervalu vyskytují rovnoměrně. Proto n_j, jako počet subjektů na začátku k-tého intervalu, je často upravován tak, aby odrážel průměrný počet subjektů v riziku výskytu sledované události takto:

$$n'_j = n_j - \frac{c_j}{2}. \tag{2.3}$$

Definice 2.1. Odhad funkce přežití pomocí tabulek úmrtnosti je pak definován jako

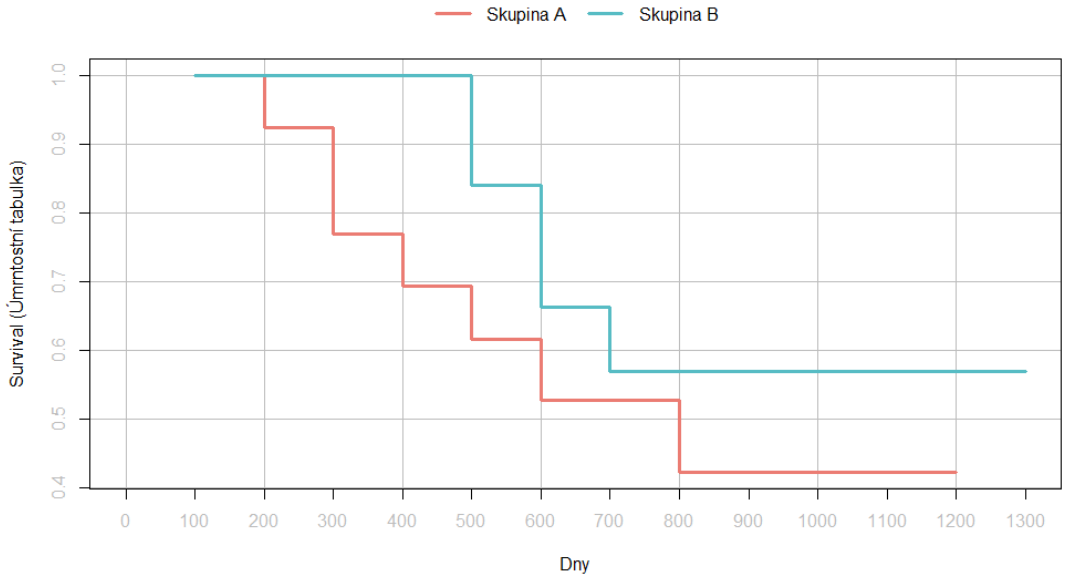
$$\hat{S}(t) = \prod_{j=1}^{k} \left(1 - \frac{d_j}{n'_j}\right).$$

Na data „ovarian“ použijeme tentokrát metodu úmrtnostních tabulek. Z tabulky 2.3, pro skupinu A, vidíme intervalové rozdělení, které si můžeme v programu libovolně nastavit. Všimněme si, že díky cenzorovaným hodnotám může průměrný počet subjektů n'_j nabývat ne-celočíselných hodnot.

Obrázek 2.2 vyjadřuje odhad pomocí úmrtnostních tabulek skupiny A a B. Pokud porovnáme tento graf s grafem 2.1, vidíme u skupiny A rozdílné odhady. U K-M odhadu cenzorované údaje nezpůsobovaly „skoky“, pouze zvyšovaly počet subjektů v riziku. U úmrtnostních tabulek cenzorované údaje ovlivňují velikost „skoků“.

interval	n'_i	d_i	$\hat{S}(t_i)$
0–100	13.0	1	1.0000
100–200	12.0	2	0.9230
200–300	10.0	1	0.7692
300–400	9.0	1	0.6923
400–500	7.0	1	0.6153
500–600	5.0	0	0.5274
600–700	5.0	1	0.5274
700–800	4.0	0	0.4219
800–900	3.0	0	0.4219
900–1000	2.0	0	0.4219
1000–1100	1.5	0	0.4219
1100–1200	0.5	0	0.4219

Tabulka 2.3: Hodnoty odhadu funkce přežití skupiny A – úmrtnostní tabulky



Obrázek 2.2: Odhad funkce přežití pomocí úmrtnostních tabulek – dataset „ovarian“

3 Porovnání rozdělení

V předchozí kapitole jsme si ukázali jak odhadnout distribuci časů T. V této kapitole si ukážeme jak takové odhady distribuce porovnat. V analýze přežití se může porovnávat více odhadů, jimiž jsou funkce přežití. Testujeme tedy hypotézu

$$H_0 : S_1 = S_2 = \cdots = S_k, \tag{3.1}$$

kde S_i je funkce přežití v i-té skupině pro $i = 1, 2, \ldots, k$.

Pokud se v pozorováních nevyskytují cenzorovaná data, lze použít klasické neparametrické testy, např. Mannův-Whitneyův test, či Kruskalův-Wallisův. Pro data s cenzorovanými pozorováními byly vyvinuty speciální testy.

3.1 Log-rank test

Log-rank test je jedním z nejpoužívanějších statistických testů pro porovnání distribuce dvou, či více skupin. Tento test vychází z Mantelova-Haenszelova testu jako modifikace pro analýzu stratifikovaných kontingenčních tabulek. Logiku testu si odvodíme pro porovnání dvou skupin. Pro více skupin by platily vztahy obdobně.

Předpokládejme ve dvou skupinách necenzorované časy přežití $t_1 < t_2 < \cdots < t_m$. V každém z těchto m časů lze sestrojit kontingenční tabulku shrnující pozorované přežití. Předpokládáme ve dvou skupinách počet $n_{j,1}, n_{j,2}$ jedinců v nebezpečí úmrtí před časem t_j a celkový rozsah výběru $n_j = n_{j,1} + n_{j,2}$. Obdobně pak ve dvou skupinách $d_{j,1}, d_{j,2}$ jako počty sledovaných událostí (úmrtí) v čase t_j. Příkladem kontingenční tabulky shrnující přežití ve skupinách 1 a 2 je tabulka 3.1.

skupina	událost nastala	událost nenastala	celkem
1	$d_{j,1}$	$n_{j,1} - d_{j,1}$	$n_{j,1}$
2	$d_{j,2}$	$n_{j,2} - d_{j,2}$	$n_{j,2}$
celkem	d_j	$n_j - d_j$	n_j

Tabulka 3.1: Pozorované počty událostí v čase t_j

Na základě hypotézy $H_0 : S_1 = S_2$ tedy $H_1 : S_1 \neq S_2$ a její platnosti lze ukázat (viz [1]), že náhodná veličina $d_{j,1}$ má hypergeometrické rozdělení pravděpodobnosti se střední hodnotou

$$E(d_{j,1}) = \frac{n_{j,1} d_j}{n_j}$$

a rozptylem $d_{j,1}$

$$var(d_{j,1}) = \frac{n_{j,1} n_{j,2} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}.$$

Zavedeme statistiku U

$$U = \sum_{j=1}^{m} (d_{j,1} - E(d_{j,1})),$$

která vyjadřuje rozdíl mezi celkovým a očekávaným počtem sledovaných událostí ve skupině 1. Pro její rozptyl platí

$$var(U) = \sum_{j=1}^{m} var(d_{j,1}) = \sum_{j=1}^{m} \frac{n_{j,1} n_{j,2} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}.$$

Má-li statistika U přibližně normální rozdělení pravděpodobnosti s nulovou střední hodnotou, pak statistika

$$L = \frac{U}{\sqrt{var(U)}} \sim N(0, 1)$$

má přibližně normální rozdělení pravděpodobnosti s nulovou střední hodnotou a jednotkovým rozptylem (aproximace je tím lepší, čím více máme pozorovaných událostí) [6]. Dále víme, že pokud náhodnou veličinu se standardizovaným normálním rozdělením umocníme na druhou, získáme náhodnou veličinu, která se řídí rozdělením chí-kvadrát

$$L^2 = \frac{U^2}{var(U)} \sim \chi^2(1). \tag{3.2}$$

Statistika L^2 nám vyjadřuje odchýlení pozorovaných dob přežití od očekávaných, pro dvě skupiny. Hypotézu H_0, kdy není přepokládán rozdíl mezi skupinami, zamítáme pro vysoké hodnoty statistiky L^2. Testová statistika L^2 bude mít pro k skupin χ^2 rozdělení s $k-1$ stupni volnosti. V případě zamítnutí nulové hypotézy víme, že mezi skupinami existuje nějaký rozdíl. K přesnému určení rozdílu mezi skupinami musíme udělat post-hoc testy mnohonásobného porovnání.

3.2 Wilcoxonův test

Wilcoxonův test je velmi podobný log-rankovému testu. Při stejné nulové hypotéze je pak Wilcoxonův text založen na statistice

$$U = \sum_{j=1}^{m} n_j(d_{j,1} - E(d_{j,1})),$$

ve které je navíc uvažována váha rozdílu pozorovaných a očekávaných dob všech skupin v čase t_j počtem jedinců n_j. Rozptyl statistiky U je dán jako

$$var(U) = \sum_{j=1}^{m} \frac{n_{j,1}n_{j,2}d_j(n_j - d_j)}{n_j - 1}$$

a Wilcoxonova testová statistika

$$W = \frac{U^2}{var(U)} \sim \chi^2(1)$$

má za platnosti nulové hypotézy chí-kvadrát rozdělení o jednom stupni volnosti.

Z těchto dvou testů je log-rank test silnější, pokud dvě sledované skupiny mají vzájemně proporcionální rizikové funkce. Lze tedy v odhadech funkcí přežití obou skupin pozorovat narůstající rozdíl. V ostatních případech je lepší použití Wilcoxonova testu, protože oproti log-rank testu dává větší váhu počátečním hodnotám a ne tehdy, kdy je počet žijících jedinců malý.

Pro nejjednodušší rozhodnutí o tom, jaký test použít, nám dobře poslouží vyobrazení dvou skupin pomocí Kaplan-Meierova odhadu. Při křížení funkcí přežití je vhodné použít Wilcoxonův test, protože nevyžaduje konzistentní poměr rizika, ale vyžaduje, aby jedna skupina měla trvale vyšší riziko. Pro tyto a další informace o porovnáních rozdělení je možné najít ve zdroji [6] nebo [5] v kapitole 5.

4 Coxův model proporcionálních rizik

V předchozích kapitolách jsme se věnovali neparametrickým metodám, které jsou vhodné pro porovnávání dvou či více skupin. Nicméně u většiny studií bývá doplněno mnoho dalších informací u každého subjektu. Příkladem mohou být lékařské studie, které u pacientů kromě informaci o pohlaví, nabízejí informace o životním stylu – kouření, stravovací návyky. Ty mohou mít ovšem významný dopad na délku přežití pacientů. Pro zpracování těchto informací slouží regresní modely, které hodnotí vztahy mezi vysvětlovanou proměnnou a vysvětlujícími proměnnými (informace o životním stylu). V analýze přežití se vlivem cenzorování vyskytují kompletní i nekompletní údaje, proto nelze použít klasické postupy při regresním modelování.

Tato kapitola tedy pojednává o použití regresních modelů v analýze přežití, konkrétně *Coxova modelu proporcionálních rizik* – často využívaného ve financích a pojišťovnictví. Model vychází ze složité teorie, které se nebudeme věnovat. Popíšeme tento model, ukážeme jeho aplikaci a význam. Podrobnější informace o Coxovu modelu lze nalézt ve zdroji [1] kapitola 3, [5] kapitola 12 nebo [4].

Definice 4.1. *Coxův model proporcionálních rizik* je vyjádřen pomocí rizikové funkce vztahem

$$h(t|x) = h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p) = h_0(t)\exp(\mathbf{x}'\boldsymbol{\beta}),$$

kde x_i, $i = 1, \ldots, p$ jsou ovlivňující faktory (kovariáty) vysvětlujících proměnných, β_i, $i = 1, \ldots, p$ jsou odhadované parametry příslušné jednotlivým kovariátům.

Funkce $h_0(t)$ se nazývá **základní riziková funkce** (*baseline hazard function*), která představuje změny rizika v čase. Tuto funkci není potřeba znát, protože postup odhadu regresních koeficientů modelu není závislý na podobě této funkce. Výraz $\exp(\mathbf{x}'\boldsymbol{\beta})$ vyjadřuje tzv. **poměr rizik**, to je poměr rizikové funkce k základní rizikové funkci, který je v průběhu času konstantní. Použití exponenciální funkce zajišťuje, že nebezpečí je pozitivní. Poměr

rizik lze vyjádřit pro dvě proměnné takto:

$$\frac{h(t|x_1)}{h(t|x_2)} = \frac{h_0(t)\exp(\mathbf{x_1'}\boldsymbol{\beta})}{h_0(t)\exp(\mathbf{x_2'}\boldsymbol{\beta})} = \exp((\mathbf{x_1} - \mathbf{x_2})'\boldsymbol{\beta}).$$

Míra rizika, která se odhaduje z Coxova modelu proporcionálních rizik, se interpretuje spíše jako relativní než absolutní riziko. Předpokládá se, že kovariáty x_i mají aditivní účinek na přirozený logaritmus rizikové funkce. Interpretace parametrů β_i je taková, že pro každé zvýšení kovarianty x_i o jednotku,se zvýší riziko násobkem $\exp\beta_i$. Kladná hodnota keoficientu znamená, že riziko události je větší s vyšší hodnotou vysvětlující proměnné. Záporný koeficient nám říká, že vysvětlující proměnná má s vyšší hodnotou nižší riziko výskytu události.

$$\exp\beta_i = \frac{h(t, x_1, x_2, \ldots, x_{i+1}, \ldots, x_p)}{h(t, x_1, x_2, \ldots, x_i, \ldots, x_p)} \tag{4.1}$$

V případě binární proměnné nabývající hodnot 0 a 1 tato hodnota vyjadřuje, kolikrát větší riziko výskytu sledované události má riziková skupina subjektů proti skupině referenční (za předpokladu, že jsou obě skupiny srovnatelné s ohledem na ostatní faktory).

Odhad jednotlivých regresních koeficientů probíhá metodou *parciální věrohodnosti*, která je obdobou metody maximální věrohodnosti. Místo standardní funkce věrohodnosti je maximalizována tzv. parciální věrohodnostní funkce:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{\sum_{l=1}^{n} \exp(\mathbf{x}_l'\boldsymbol{\beta})},$$

kde n udává počet zkoumaných objektů a k počet necenzorovaných subjektů. V praxi se obvykle pracuje s logaritmem věrohodnostní funkce, jejíž maximum se hledá. Pro nalezení odhadů jednotlivých regresních koeficientů se provede k parciálních derivací, které se položí rovno 0, a vyřeší se příslušný systém rovnic.

5 Analýza dat

V této kapitole představíme data z pojišťovny. Pro analýzu přežití jsme vybrali data o likvidaci pojistných událostí havarijního pojištění. Budeme tedy zkoumat dobu od hlášení pojistné události po dokončení likvidace pojistné události.

Havarijní pojištění, tzv. KASKO (*Casualty and Collision*), spadá pod neživotní pojištění. Toto pojištění kryje škody na motorových vozidlech, ať už ji řidič způsobil, či nikoli. Kryje rizika škody na vozidle, dále rizika živelní, riziko odcizení a vandalství. Cenu pojištění ovlivňuje výše spoluúčasti (částka, kterou se pojistník podílí na výši škody), pořizovací cena vozidla, typ vozidla, obsah, věk řidiče nebo počet osob využívajících vozidlo.

5.1 Úprava a reprezentace dat

K dispozici máme 6914 případů pojistných událostí. Surová data musela být upravena pro další použití v analýze. Musel být vypočítán čas přežití – tj. čas do doby likvidace, doplněny údaje o cenzorování a vhodně rozdělit původní proměnné do kategorií. Všechny úpravy probíhaly v aplikaci Excel. Následná podoba dat obsahuje 6914 řádků, tedy pojistné události, 9 sloupců poskytujících dodatečné informace k události.

- ANRD: *1 - muž, 2 - žena, 4 - firma*
- time: čas přežití ve dnech
- indicator: *1 - čas přežití, 0 - cenzorovaný čas přežití*
- PSC: *1-7 - Struktura PSČ*
- cechy_morava: proměnná PSC rozdělena na Čechy a Moravu
- obdobi: jaro, léto, podzim, zima
- obdobi_2: jaro+léto a podzim+zima

- vyroba: kategorie rozdělující rok výroby auta - *1990_2000, 2000_2010, 2010_2017*
- vehicle_age: *nove* - auta mladší osmi let; *stare* - stará osm a více let

Data z pojišťoven a bank se oproti jiným oblastem liší v cenzorování a krácení dat. Většinou k cenzorování dochází z časových důvodů. Cenzorovanými údaji byly uzavřené, ale nevyplacené události. Proměnná `PSC` byla rozdělena do kategorií podle první číslice, ta přibližně odpovídá starému členění Československa na kraje. Proměnná `obdobi` vyjadřuje rozdělení ročního období podle datumu nahlášení škody.

Po zpracování v Excelu byla data importována do aplikace RStudio. Obrázek 5.1 představuje ukázku dat datasetu `likvidaceW`.

	ANRD	time	indicator	PSC	cechy_morava	obdobi	vyroba	vehicle_age	obdobi_2
1	1	0	1	1	1	podzim	2010_2017	nove	zima
2	4	0	0	2	1	zima	2000_2010	nove	zima
3	2	0	1	3	1	leto	2000_2010	stare	leto
4	4	0	0	3	1	leto	2010_2017	nove	leto
5	1	0	1	7	2	leto	2010_2017	nove	leto
6	1	0	0	6	2	zima	2000_2010	nove	zima
7	2	0	0	2	1	jaro	2000_2010	stare	leto
8	1	0	1	1	1	podzim	2010_2017	nove	zima
9	2	0	0	2	1	leto	2010_2017	nove	leto
10	1	0	1	1	1	leto	2000_2010	stare	leto
11	4	1	1	6	2	zima	2010_2017	nove	zima
12	2	1	1	2	1	zima	2000_2010	nove	zima
13	4	1	1	1	1	leto	2010_2017	nove	leto

Obrázek 5.1: Náhled dat

5.2 Kaplanův-Meierův odhad

Nejprve začneme čistým odhadem funkce přežití, která představuje dobu do likvidity. Pro tyto účely budeme využívat funkcionalitu balíku `survival`[1]. Událost je tedy v našem případě likvidita a riziko udává šanci, že bude událost v dalším okamžiku zlikvidována. Pokud chceme v analýze dat použít Kaplanův-Meierův odhad, aplikujeme na data funkci `survfit`, která sdružuje odhady funkce přežití pomocí různých metod, dle zadadné formule (předpisu). Pokud chceme prostý K-M odhad, pak se volí závislost na nespecifické proměnné `1`. Pokud se jedná o odhady pomocí proměnných z datasetu, volíme závislost jedné

[1] https://cran.r-project.org/web/packages/survival/index.html

proměnné. Tato funkce obsahuje dva hlavní argumenty – `survival object` (tj. formule) a `dataset`. `Survival object` se vytvoří pomocí funkce `Surv`, která obsahuje proměnné čas přežití a cenzorovanou proměnnou. Jako výsledek analýzy se zobrazí základní informace o modelu znázorněné ve výstupu 5.1 doplněné o příkazy. Máme tedy 6914 záznamů, z toho je 6299 událostí. K cenzorování došlo v 615 případech. Polovina případů bude vyřešena do 43 dnů od začátku řízení (medián) a s 95% odhadem 42 až 44 dnů.

```
> surv_object <- Surv(time = likvidaceW$time, event = likvidaceW$
    indicator)
> fit_KM <- survfit(surv_object ~ 1, data = likvidaceW)
> print(fit_KM)
Call: survfit(formula = surv_object ~ 1, data = likvidaceW)

      n  events  median 0.95LCL 0.95UCL
   6914    6299      43      42      44
```

Výpis 5.1: Základní informace o odhadu

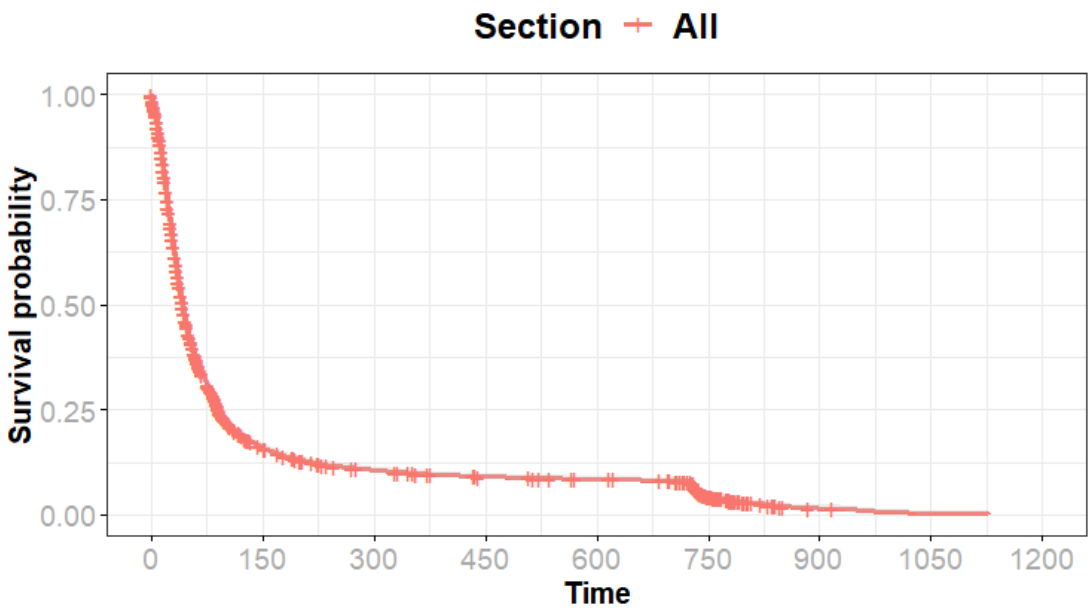
Podrobnější informace o K-M odhadu nám poskytuje funkce `summary`, která vypíše informace o prvních deseti dnech pozorování. Ve výstupu 5.2 v prvním sloupci se vyskytuje vzestupně seřazený čas ve dnech, který uvádí čas do likvidace. V druhém sloupci počet hlášených událostí. Následuje třetí sloupec s počtem zlikvidovaných a vyplacených škod. Sloupec `survival` vyjadřuje odhad funkce přežití a poslední dva sloupce jeho 95% interval spolehlivosti.

```
> summary(fit_KM)
Call: survfit(formula = surv_object ~ 1, data = likvidaceW)

 time n.risk n.event survival  std.err lower 95% CI upper 95% CI
    0   6914       5  0.99928 0.000323     0.998643      0.99991
    1   6904      33  0.99450 0.000890     0.992758      0.99625
    2   6842      63  0.98534 0.001448     0.982510      0.98818
    3   6754      54  0.97747 0.001790     0.973964      0.98098
    4   6691      39  0.97177 0.001998     0.967859      0.97569
    5   6637      49  0.96459 0.002231     0.960231      0.96898
    6   6580      50  0.95726 0.002443     0.952488      0.96206
    7   6521      57  0.94890 0.002661     0.943695      0.95413
    8   6458      89  0.93582 0.002964     0.930029      0.94165
    9   6361      89  0.92273 0.003231     0.916415      0.92908
   10   6267      79  0.91109 0.003445     0.904367      0.91787
```

Výpis 5.2: Podrobnější informace o odhadu

Obrázek 5.2 vyobrazuje podobu Kaplanova-Meierova grafu pomocí funkce `ggsurvplot`. Ta obsahuje nepřeberné množství způsobů, jak graficky znázornit odhadovanou funkci přežití – ukázka v podkapitole 5.3. Funkce přežití na začátku prudce klesá. To je pravděpodobně zapříčiněno povinností do 90 dnů zlikvidovat pojistnou událost. V polovině prvního roku křivka zvolňuje a v tomto pozvolném trendu pokračuje dále.



Obrázek 5.2: Kaplanův-Meierův graf

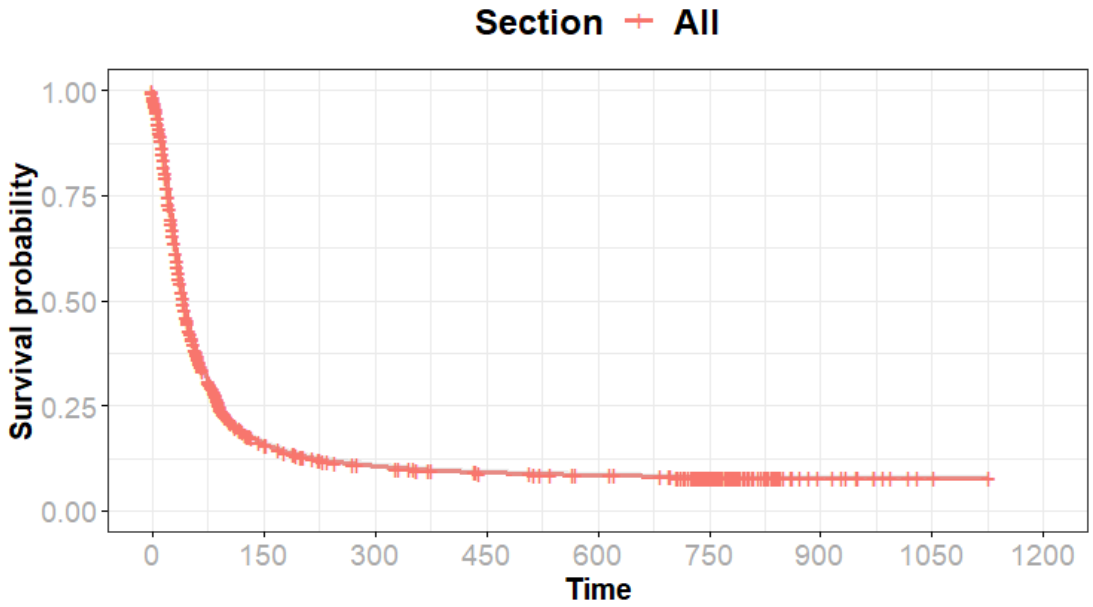
Po dvou letech je vidět propad zapříčíněný velkým množstvím zlikvidovaných pojistných událostí. Tento nesoulad je zapříčiněn procházením starých případů událostí a jejich uzavřením, pokud jsou nelikvidní. Protože je tento krok systematický, mohl by ovlivňovat výsledky naší analýzy. Musíme tedy zlikvidované události, delší dvou let cenzorovat.

Odhad funkce přežití již vypadá sourodě (obrázek 5.3). S takto upravenými daty můžeme pokračovat dále.

Zajímavou informaci o tomto grafu nám poskytne funkce `quantile`. Vidíme, že čtvrtina událostí bude zlikvidována do 22 dnů, polovina do 43 dnů a tři čtvrtiny do tří měsíců.

```
> quantile(fit_KM, probs = c(0.25, 0.5, 0.75))
$quantile
25 50 75
22 43 92
```

Výpis 5.3: Kvantily K-M odhadu



Obrázek 5.3: Upravený Kaplanův-Meierův odhad

5.3 Dvouvýběrový test

Z předchozí podkapitoly již víme, jak v RStudiu vytvořit odhad funkce přežití. Pomocí dvouvýběrových testů můžeme na datech zkoumat rozdíly v rozdělení času přežití například pro fyzické (FO) a právnické osoby (PO). V proměnné `ANRD` byly spojeny hodnoty `Muži` a `Ženy`, čímž pak vznikla kategorie fyzických osob v novém datasetu `data_FOvsPO`. Použijeme stejné funkce `Surv`, `survfit` a uděláme odhad vztažený na proměnnou `ANRD`, jak je vidět z příkazů ve výstupu 5.4, ve kterém je zobrazen i výsledek odhadu.

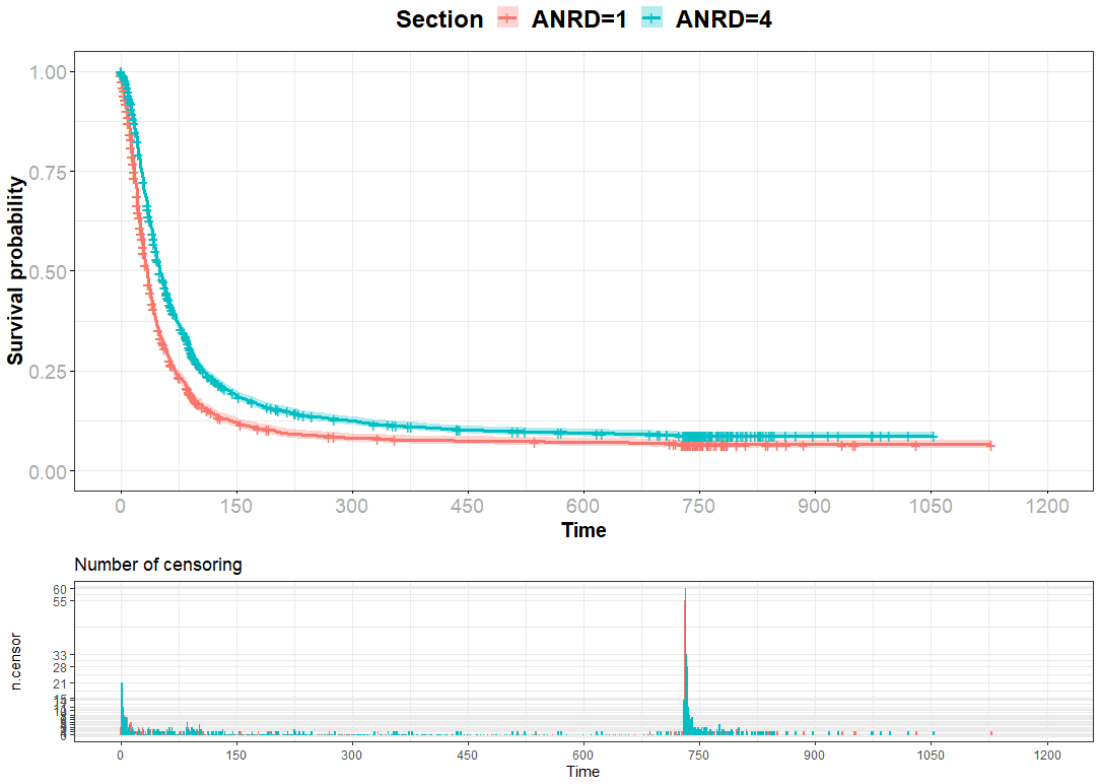
```
> fit_ANRD <- survfit(surv_object ~ ANRD, data = data_FOvsPO)
> print(fit_ANRD)
Call: survfit(formula = surv_object ~ ANRD, data = data_FOvsPO)

            n events median 0.95LCL 0.95UCL
ANRD=FO  2915   2608     34      32      36
ANRD=PO  3999   3494     51      49      54
```

Výpis 5.4: Základní informace pro obě skupiny

Již z přehledu je vidět, že se doby likvidace pro fyzické a právnické osoby budou výrazně lišit. Medián pro FO je 34, pro PO je 51. Tento rozdíl bude ještě výrazněji vidět, pokud si odhady funkcí přežití vykreslíme pomocí funkce `ggsurvplot`. Tato funkce nabízí do výstupu zahrnout další část grafu, například počet cenzorovaných událostí v průběhu času

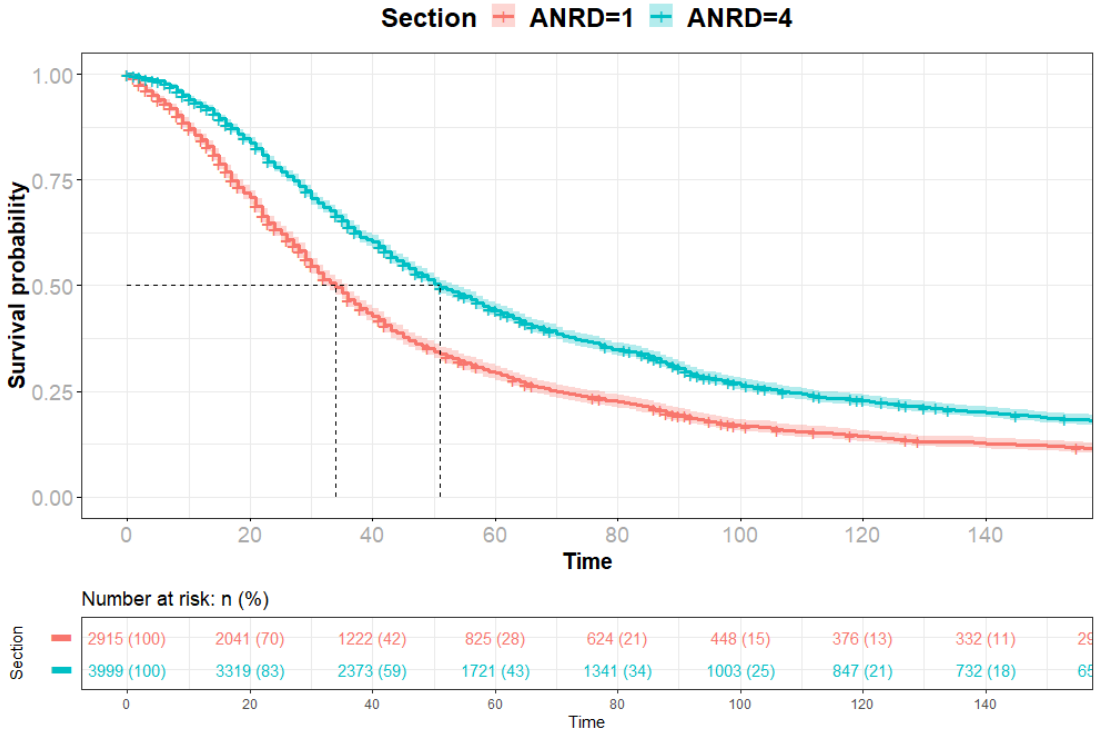
(obrázek 5.4). Vidíme zde vysoký počet cenzorovaných událostí v období dvou let, to jsou přesně ty časy přežití, které způsobovaly propady v odhadech přežití.



Obrázek 5.4: K-M odhad pro fyzické a právnické osoby

Z obrázku 5.4 bohužel nevidíme tak dobře podstatnou část průběhu odhadované funkce. Většina škodních událostí byla zlikvidována do půl roku, proto byl graf upraven do podoby obrázku 5.5. Z něho je možné jednoduše vypozorovat medián jednotlivých skupin, popřípadě horní i dolní kvantily. V tabulce pod funkcemi je vyobrazen jak počet aktuálně nezlikvidovaných škodních událostí, tak i procentuální zastoupení.

Z pohledu na graf 5.5 lze předpokládat, že rozdělení času přežití obou skupin se liší. Pro potvrzení tohoto předpokladu, vzhledem k nekřížení odhadů přežití, použijeme základní Log-rankový test. Pro tento test použijeme funkci `survdiff`, která obsahuje parametry `formula, data` a `rho`, které značí typ testu. Pro hodnotu 0 je to Log-rankový test, pro hodnotu 1 je ekvivalentní Peto-Petoův modifikovaný Gehanův-Wilcoxonův test. Z výstupu 5.5 zjistíme hodnotu testové statistiky a p-hodnotu, která je téměř nulová. Na základně p-hodnoty hypotézu H_0 zamítáme, vzhledem k 5% hladině významnosti. Rozdělení časů přežití se tedy významně liší. Delší čas likvidity má skupina právnických osob. Mohlo by to



Obrázek 5.5: Zkrácený odhad přežití

být zapříčiněno hromadnými pojištěními, kdy je přeci jen složitější administrativa. Nebo je to zapříčiněno tím, že právnické osoby, oproti fyzickým, nespěchají tolik na opravu a nepotřebují peníze hned, jsou schopni nějakou chvíli čekat.

```
> survdiff(surv_object ~ ANRD, data = data_FOvsPO, rho = 0)
Call:
survdiff(formula = surv_object ~ ANRD, data = data_FOvsPO, rho = 0)

            N Observed Expected (O-E)^2/E (O-E)^2/V
ANRD=FO  2915     2608     2144     100.6       158
ANRD=PO  3999     3494     3958      54.5       158

 Chisq= 158  on 1 degrees of freedom, p= <2e-16
```

Výpis 5.5: Log-rankový test

Pro příklad zde uvedeme Log-rankový test pro více položek. Jako proměnnou do funkce `survdiff` použijeme `PSC`. Z úvodu víme, že tato čísla označují první místo v PSČ starého rozdělení Československa na kraje:

- 1 - PSČ Prahy,

- 2 - PSČ Středočeského kraje,
- 3 - PSČ Jihočeského a Západočeského kraje,
- 4 - PSČ Severočeského kraje,
- 5 - PSČ Východočeského kraje a části Jihomoravského kraje,
- 6 - PSČ Jihomoravského kraje,
- 7 - PSČ Severomoravského kraje.

Z výstupu 5.6 opět zamítáme nulovou hypotézu, víme tedy, že mezi skupinami exituje nějaký rozdíl. Abychom určili konkrétně, mezi jakými skupinami jsou signifikantní rozdíly, použijeme funkci `pairwise_survdiff`.

```
> survdiff(surv_object ~ PSC, data = data_FOvsPO, rho = 0)
Call:
survdiff(formula = surv_object ~ PSC, data = data_FOvsPO, rho = 0)

          N Observed Expected (O-E)^2/E (O-E)^2/V
PSC=1 2724     2373     2670   32.9721   59.5838
PSC=2  849      761      671   12.1521   13.8593
PSC=3  699      613      595    0.5197    0.5845
PSC=4  747      646      640    0.0521    0.0591
PSC=5  566      515      471    4.1002    4.5109
PSC=6  677      598      579    0.6431    0.7211
PSC=7  652      596      476   30.1411   33.2406

 Chisq= 82  on 6 degrees of freedom, p= 1e-15
```

Výpis 5.6: Log-rankový test pro více položek

Tato funkce spadá pod testy mnohonásobného porovnání. Z výstupu 5.7 vidíme tabulku p-hodnot, vyjadřující výsledek testů všech dvojic.

```
> res <- pairwise_survdiff(Surv(time=time,event =indicator) ~ PSC,
+                          data = data_FOvsPO)
> res

  Pairwise comparisons using Log-Rank test

data:  data_FOvsPO and PSC


```

```
  1       2       3       4       5       6
2 1.9e-08 -       -       -       -       -
3 0.0023  0.1352  -       -       -       -
4 0.0090  0.0741  0.7771  -       -       -
5 5.1e-05 0.5944  0.4049  0.2713  -       -
6 0.0023  0.1514  0.9251  0.7771  0.4281  -
7 6.5e-14 0.1674  0.0023  0.0021  0.0572  0.0037
```

Výpis 5.7: Post-hoc test

Pro zjednodušení jsme pomocí funkce `symnum` nahradili p-hodnoty symboly, které určují významnosti testů. Z toho plyne, že hypotézy zamítáme mezi hlavním městem Prahou a všemi kraji. Rozdělení časů přežití mezi Prahou a ostatními kraji se liší. Dále vidíme, že hypotézy zamítáme mezi všemi kraji (až na výjimku Jihomoravského kraje) a Severomoravským krajem. Z toho plyne, že nebylo dobré sloučit proměnnou `PSC` na `cechy_morava`. Lepší bude sloučit proměnnou na Prahu, Čechy a Moravu – proměnná `PSC2`.

```
> symnum(res$p.value, cutpoints = c(0, 0.0001, 0.001, 0.01, 0.05, 0.1, 1)
    ,
+         symbols = c("****", "***", "**", "*", "+", " "),
+         abbr.colnames = FALSE, na = "")
  1    2    3    4    5    6
2 ****
3 **
4 **   +
5 ****
6 **
7 ****  **   **   +    **
attr(,"legend")
[1] 0 '****' 1e-04 '***' 0.001 '**' 0.01 '*' 0.05 '+' 0.1 ' ' 1 \t    ##
    NA: ''
```

Výpis 5.8: Vizualizace post-hoc testu

5.4 Coxův model proporcionálních rizik

V dalším kroku nás zajímá, jaký vliv mají ostatní proměnné na vybraná data. Použijeme funkci `coxph`, kde k `ANRD` přidáme další proměnné – `obdobi_2, PSC2, vehicle_age`. Výsledek zobrazíme pomocí funkce `summary`, viz výstup 5.9. Všimněme si, že model označil právnické osoby, zimní část roku, Moravu a Prahu, jako signifikantní. Hledané parametry β_i, které ovlivňují hodnotu přirozeného logaritmu rizikové funkce, přestavují sloupec `coef`.

Proměnná `ANRD$PO` podle parametru snižuje hodnotu rizikové funkce. Sloupec `exp(coef)` vyjadřuje poměry rizika.

```
> cox <- coxph(surv_object ~ ANRD + obdobi_2+ cechy_morava +vehicle_age,
+                    data = data_FOvsPO)
> summary(cox)
coxph(formula = surv_object ~ ANRD + obdobi_2 + PSC2 + vehicle_age,
    data = data_FOvsPO)

  n= 6914, number of events= 6102

                     coef exp(coef) se(coef)       z Pr(>|z|)
ANRDPO           -0.29872   0.74177  0.02852 -10.475   <2e-16 ***
obdobi_2zima      0.05522   1.05677  0.02567   2.151   0.0314 *
PSC2Morava        0.07693   1.07997  0.03515   2.188   0.0286 *
PSC2Praha        -0.07557   0.92721  0.03082  -2.452   0.0142 *
vehicle_agestare -0.02751   0.97287  0.03211  -0.857   0.3916
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Výpis 5.9: Souhrn informací o modelu

Další informace o modelu poskytuje výstup 5.10. Hodnota *concordance* – shoda, nám říká na kolik tyto proměnné vysvětlují chování tohoto modelu. Likehood ratio test, Wald test a Score test jsou testy regresních koeficientů.

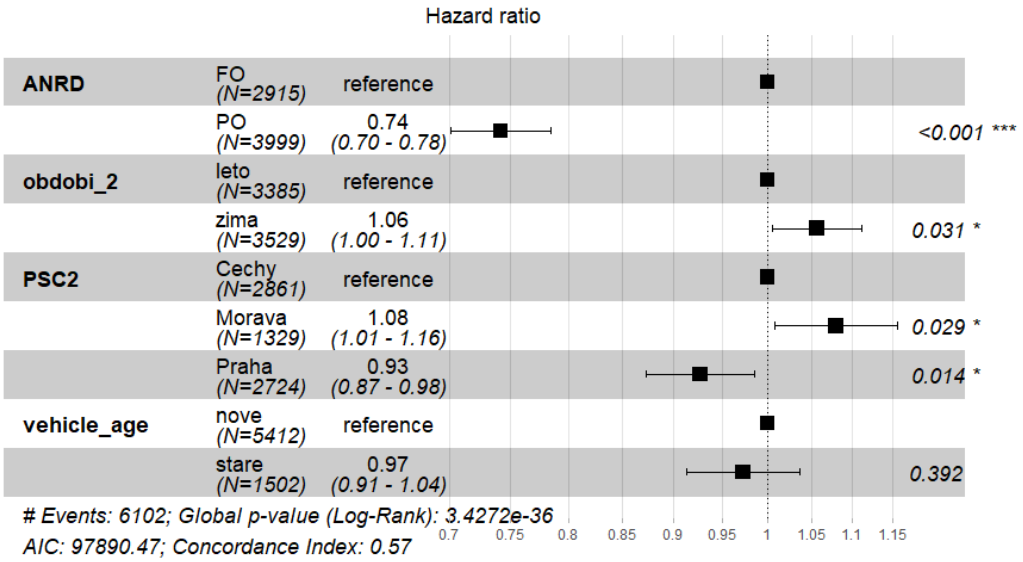
```
                 exp(coef) exp(-coef) lower .95 upper .95
ANRDPO              0.7418     1.3481    0.7014    0.7844
obdobi_2zima        1.0568     0.9463    1.0049    1.1113
PSC2Morava          1.0800     0.9260    1.0081    1.1570
PSC2Praha           0.9272     1.0785    0.8729    0.9849
vehicle_agestare    0.9729     1.0279    0.9135    1.0361

Concordance= 0.568  (se = 0.004 )
Likelihood ratio test= 176.2  on 5 df,   p=<2e-16
Wald test            = 179.3  on 5 df,   p=<2e-16
Score (logrank) test = 180.8  on 5 df,   p=<2e-16
```

Výpis 5.10: Druhá část výstupu

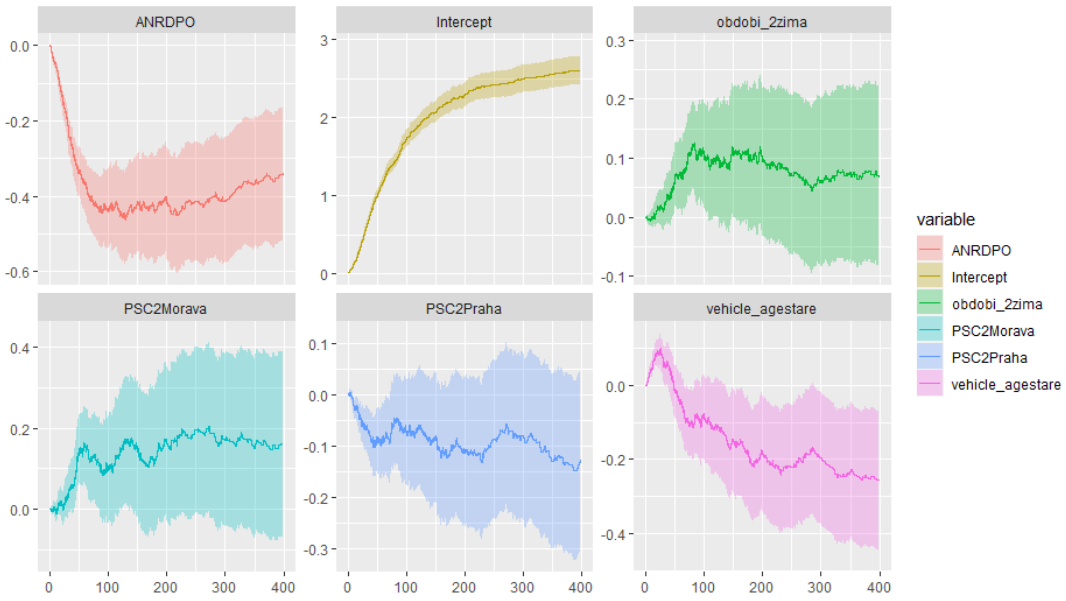
Následně tyto výsledky modelu vyzualizujeme pomocí funkce `ggforest`. Tento typ vyobrazení nám přehledně ukáže již zmíněné poměry rizik (HR) pro všechny proměnné, které jsou zahrnuty v modelu. $HR > 1$ naznačuje zvýšené riziko nastání události, v našem případě dokončení likvidace pojistné události. V opačném případě $HR < 1$ je riziko menší. Každý poměr rizik představuje relativní riziko dokončení pojistné události, které porovnává

jednu část kováritu s další částí. Například poměr rizika 0,74 pro skupinu fyzických a právnických osob z obrázku 5.6 nám říká, že právnické osoby, mají snížené riziko nastání sledované událost (dokončení likvidace pojistné události) ve srovnání s fyzickými osobami. Interval spolehlivosti je $0,70$, až $0,78$, tento výsledek je významný. Na obrázku je hezky vidět poměr rizik u proměnné `PSC2`, kde Praha riziko snižuje a Morava prozměnu zvyšuje. U Prahy by se to dalo vysvětlit zvýšenou nehodovostí vzhledem ke zbytku ČR.



Obrázek 5.6: Poměr rizik

Nad rámec představené teorie je možné využít vestavěné funkce pro Aalenův aditivní regresní model, který zahrnuje do modelu časovou složku. Znázorňuje, jak se účinky kovariát mění v čase. Pokud se z obrázku 5.7 zaměříme na kovariátu `ANRD$PO`, vidíme hned ze začátku prudký pokles účinku, který se zastavil až po sto dnech. Interpretace tohoto výsledku je složitá, protože změna koficientů β_i v průběhu času nám nemůže posloužit ke kvantifikování vlivu na rizikovou funkci. K tomu slouží již ukázané testování statisticky významných kovariát, ale Aalenovy grafy jsou jediným způsobem, jak interpretovat parametry v čase.



Obrázek 5.7: Aalenův aditivní regresní model

Závěr

Analýza přežití našla uplatnění v mnoha oborech. Díky tomu se stala součásti mnoha statistických programů, nevyjímaje jazyka R, který je jedním z nejrozšířenějších a nejpoužívanějších pro statistické analýzy.

V teoretické části jsme se věnovali nějčastěji používaným základním metodám, které jsou pomocí balíčků obsažené i v programu RStudio. Ukázali jsme si různé metody odhadu funkce přežití a vysvětlili základní testy pro porovnání rozdělení času přežití ve dvou skupinách. Teoretickou část jsme zakončili výkladem o semi-parametrikém modelu, hojně využivaného ve financích.

Tyto teoretické poznatky jsme aplikovali na reálných datech z pojišťovny. Při analýze dat, jsme pomocí vizualizace odhadu funkce přežití narazili na nesoulad v datech, který by jinak mohl ovlivňovat výsledky. Pomocí dvouvýběrových testů jsme byli schopni rozhodnout o rozdělení času přežití dvou skupin a následně ukázali, jakým nedostatkem trpí Log-rank test při porovnání vícero skupin. V poslední části jsme pomocí Coxova modelu rozhodovali o významnosti jednotlivých kovariátů a přehledně vizualizovali poměry rizik.

Literatura

[1] David Collett. *Modelling survival data in medical research*. CRC press, 2015.

[2] Jian Huang et al. Efficient estimation for the proportional hazards model with interval censoring. *The Annals of Statistics*, 24(2):540–568, 1996.

[3] Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

[4] Christiana Kartsonaki. Survival analysis. *Diagnostic Histopathology*, 22(7):263–270, 2016.

[5] Elisa T Lee and John Wang. *Statistical methods for survival data analysis*, volume 476. John Wiley & Sons, 2003.

[6] Tomáš Pavlík. Aplikovaná analýza přežívání. `https://portal.matematickabiologie.cz/index.php?pg=aplikovana-analyza-klinickych-a-biologickych-dat--aplikovana-analyza-preziti`, (20. února 2020).

[7] Marta Sestelo. A short course on survival analysis applied to the financial industry. `https://bookdown.org/sestelo/sa_financial/`.