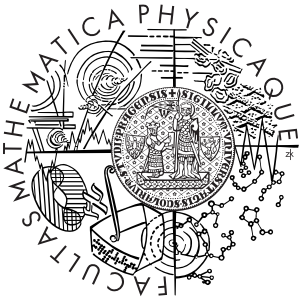
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Rudolf Kostka

Omezené a cenzorované vysvětlované proměnné

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce:	Mgr. Přemysl Bejda
Studijní program:	Matematika
Studijní obor:	Finanční matematika

Praha 2014

Zde bych chtěl především poděkovat svému školiteli za velmi vstřícný přístup, odborné poznámky a ochotu pomoci při psaní této práce. Své rodině za podporu během celého mého studia. Firmě SC Johnson za poskytnutí dat. MFF za možnost bezplatně používat program Mathematica. Knihovně MFF, kde jsem získal většinu potřebných materiálů. Všem, kteří se mnou při psaní této práce měli trpělivost.

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V Praze dne 22.května 2014 Rudolf Kostka

Název práce: Omezené a cenzorované vysvětlované proměnné

Autor: Rudolf Kostka

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: Mgr. Přemysl Bejda, Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Nejprve se práce zaměřuje na teorii týkající se různých možností zacházení s omezenými a cenzorovanými vysvětlovanými proměnnými. Začneme s diskrétními proměnnými a ukážeme teorii k binárním a ordinálním proměnným. Poté vysvětlíme užití modelů logit a probit na praktickém příkladě a zároveň provedeme jejich srovnání. Třetí kapitola se zabývá omezenými proměnnými, konkrétně cenzorovanými, useknutými a proměnnými představující nějakou dobu trvání. V poslední kapitole popíšeme některé funkce, které lze využít při programování na vykreslení funce přežití a to v softwarech R a Mathematica. Pro srovnání uvádíme i možnosti Excelu, které jsou ale značně omezené. Ukázané funkce poté demonstrujeme na konkrétním příkladě se získanými reálnými daty.

Klíčová slova: Omezené a cenzorované vysvětlované proměnné, logit, analýza přežití, R.

Title: Limited and censored explained variables

Author: Rudolf Kostka

Department: Department of Probability and Mathematical Statistics

Supervisor: Mgr. Přemysl Bejda, Department of Probability and Mathematical Statistics

Abstract: In this thesis at first we focus on theory of dealing with limited and censored explained variables. We begin with discrete variables and show the theory of binary and categorical variables. Later we explain utility of models logit and probit and demonstrate it at a practical example. We also provide a comparison of these two models. Third chapter deals with limited explained variables, specifically censored, truncated and variables representing some time to event. In the last chapter we describe some functions, which might be used to plot a graph of a survival function using softwares R or Mathematica. Some options in Excel are also mentioned, but they are very limited. Described functions are then demonstrated in use at a practical example with our gained data.

Keywords: Limited and censored explained variables, logit, survival analysis, R.

Obsah

Úvod

Tato bakalářská práce se zabývá omezenými a cenzorovanými vysvětlovanými proměnnými. Nejprve se zaměřuje na teoretickou část a poté ukazuje praktické využití v různých situacích a s různými softwary.

V první kapitole se nejprve jen teoreticky zabýváme diskrétnimi proměnnými, konkrétně binárními a ordinálními. Poté v následující kapitole ukážeme modely logit a probit, řekneme kdy se začaly poprvé používat a využijeme je v konkrétním typickém příkladě.

Třetí kapitola se týká omezených vysvětlovaných proměnných. Je zde také ukázán příklad s reálnými daty, která nám zapůjčila firma SC Johnson. Dále se porovnává, který model na data použít z předcházejcích popsaných modelů.

V poslední 4. kapitole ukážeme několik výpočtu s různými daty a v různých softwarech a pokusíme se o srovnání možností v těchto softwarech pro analýzu přežití.

Kapitola 1

Diskrétní vysvětlované proměnné

Nejprve se zabývejme diskrétními vysvětlovanými proměnnými. Zatímco v klasickém regresním modelu

$$y_t = \boldsymbol{x}_{t\cdot}\boldsymbol{\beta} + \varepsilon_t,$$

předpokládáme, že regresory $x_1, ..., x_p$ mohou ovlivňovat jen střední hodnotu y, nyní si rozebereme více obecnou situací a to sice, že regresory skrz lineární kombinaci $\sum_{t=1}^{p} x_t\beta_t$ ovlivňují distribuční funkci y [7]. Nejprve si to ukážeme na $0-1$ proměnné.

1.1 Binární vysvětlovaná proměnná

Častým typem diskrétní proměnné je *binární proměnná (binary dependent variable)*, jejíž hodnoty nabývájí jen jedniček a nul. Pro tyto veličiny jsou typické 2 následující situace, které mohou nastat:

- Jde o tzv. dummy proměnnou, která může nabývat jen dvou hodnot, typicky muž/žena, daná situace nastala/nenastala atd.

- Druhá varianta je, že proměnná sice může nabývat více hodnot, ale my si stanovíme určitou hranici a podle ní jen posuzujeme, zda ji hodnota překročila či nikoli. Např. věk osoby je více/méně než 40 let atd.

Ukažme si model pro diskrétní vysvětlované proměnné. Můžeme postupovat jako v případě spojité proměnné, ale pak je třeba se zamyslet nad možnou interpretací takového modelu [2].

Nechť tedy máme model pro binární vysvětlovanou proměnnou y_t (v čase t), kde 1 znamená, že nějaká situace nastala a 0 znamená, že nenastala. Předpokládejme, že je $\boldsymbol{X}$ pevně dané, tedy není náhodné. Pak můžeme zapsat model následovně:

$$\mathsf{P}(y_t = 1) = 1 - F(-\boldsymbol{x}_{t\cdot}\boldsymbol{\beta}), \qquad t = 1, \ldots, T,$$

kde $F(\cdot)$ je vhodně zvolená spojitá distribuční funkce.

Pokud naopak $\boldsymbol{X}$ není pevné, bude to znamenat, že pravděpodobnosti jsou podmíněné hodnotou, kterou $\boldsymbol{X}$ nabývá.

$$\mathsf{P}(y_t = 1|\boldsymbol{x}_{t\cdot}, \boldsymbol{\beta}) = 1 - F(-\boldsymbol{x}_{t\cdot}\boldsymbol{\beta}), \qquad t = 1, \ldots, T, \tag{1.1}$$

nebo ekvivalentně

$$\mathsf{P}(y_t = 0|\boldsymbol{x}_{t\cdot}, \boldsymbol{\beta}) = F(-\boldsymbol{x}_{t\cdot}\boldsymbol{\beta}), \qquad t = 1, \ldots, T. \tag{1.2}$$

Je-li distribuční funkce symetrická, resp. její hustota funkce sudá, pak lze (1.1) a (1.2) přepsat do tvaru

$$\mathsf{P}(y_t = 1|\boldsymbol{x}_{t\cdot}, \boldsymbol{\beta}) = F(\boldsymbol{x}_{t\cdot}\boldsymbol{\beta}), \qquad \mathsf{P}(y_t = 0|\boldsymbol{x}_{t\cdot}, \boldsymbol{\beta}) = 1 - F(\boldsymbol{x}_{t\cdot}\boldsymbol{\beta}).$$

Poznámka 1.1 *Interpretaci modelu s diskrétní vysvětlovanou proměnnou můžeme řešit více přístupy. Nyní uveďme 3 z nich:*

1. Můžeme použít latentní neboli tzv. nepozorovatelnou proměnnou y^*, která je provázána s regresory $\boldsymbol{X}$ lineárním modelem ,

   $$y_t^* = \boldsymbol{x}_{t\cdot}\boldsymbol{\beta} + \varepsilon_t, \tag{1.3}$$

   kde ε_t jsou iid náhodné veličiny s nulovou střední hodnotou. Jde tedy o běžný model lineární regrese se spojitou vysvětlovanou proměnnou. My ale pozorujeme binární vysvětlovanou proměnnou s hodnotami 0 a 1, u které sledujeme, zda je pod nulou či nikoli a její hodnotu určíme následovně

   $$y_t = \begin{cases} 1 & \text{pro } y_t^* > 0 \\ 0 & \text{pro } y_t^* \leq 0. \end{cases}$$

   Odtud dostaneme

   $$\mathsf{P}(y_t = 1|\boldsymbol{x}_{t\cdot}, \boldsymbol{\beta}) = \mathsf{P}(y_t^* > 0|\boldsymbol{x}_{t\cdot}, \boldsymbol{\beta}) = \mathsf{P}(\boldsymbol{x}_{t\cdot}\boldsymbol{\beta} + \varepsilon_t > 0) = 1 - F(-\boldsymbol{x}_{t\cdot}\boldsymbol{\beta}).$$

   Nyní je ovšem $F(\cdot)$ distribuční funkce reziduální složky ε modelu (1.3). Jakou zvolíme nulovou úroveň prahu, není podstatné, pokud model obsahuje intercept.

2. V dalším přístupu využijeme podmíněnou střední hodnotu

$$\begin{aligned}\mathsf{E}(y_t|\boldsymbol{x}_{t\cdot},\boldsymbol{\beta}) &= 1\cdot\mathsf{P}(y_t=1|\boldsymbol{x}_{t\cdot},\boldsymbol{\beta}) + 0\cdot\mathsf{P}(y_t=0|\boldsymbol{x}_{t\cdot},\boldsymbol{\beta})\\ &= \mathsf{P}(y_t=1|\boldsymbol{x}_{t\cdot},\boldsymbol{\beta}) = 1-F(-\boldsymbol{x}_{t\cdot}\boldsymbol{\beta}).\end{aligned}$$

   Pokud budeme psát

$$y_t = \big(1-F(-\boldsymbol{x}_{t\cdot}\boldsymbol{\beta})\big)+\varepsilon_t,$$

   pak ε představuje odchylku náhodné veličiny y od její podmíněné střední hodnoty a platí pro ni

$$\mathsf{E}(\varepsilon_t|\boldsymbol{x}_{t\cdot},\boldsymbol{\beta})=0, \qquad \mathsf{var}(\varepsilon_t|\boldsymbol{x}_{t\cdot},\boldsymbol{\beta}) = F(-\boldsymbol{x}_{t\cdot}\boldsymbol{\beta})\big(1-F(-\boldsymbol{x}_{t\cdot}\boldsymbol{\beta})\big).$$

   Rozptyl stačí spočítat pro y_t, , protože $1-F(-\boldsymbol{x}_{t\cdot}\boldsymbol{\beta})$ je díky podmíněnosti pouze konstanta.

3. Nejjednodušší možný zápis modelu by byl ve tvaru $y_t = \boldsymbol{x}_{t\cdot}\boldsymbol{\beta}+\varepsilon_t$. Vzhledem k nulové střední hodnotě residuální složky bychom ho mohli interpretovat $\boldsymbol{x}_{t\cdot}\boldsymbol{\beta} = \mathsf{E}(y_t) = 0\cdot\mathsf{P}(y_t=0)+1\cdot\mathsf{P}(y_t=1)=\mathsf{P}(y_t=1)$. Pak by ale bylo nutné přidat podmínku $0\le\boldsymbol{x}_{t\cdot}\boldsymbol{\beta}\le 1$. Tuto metodu proto nepoužíváme.

Pro interpretaci parametrů β_i, nemůžeme použít stejný přístup jako je tomu v lineárním modelu a to sice ztotožnit je s marginálním vlivem příslušného regresoru na vysvětlovanou proměnnou. Nicméně platí, že

$$\frac{\partial\mathsf{E}(y_t|\boldsymbol{x}_{t\cdot},\boldsymbol{\beta})}{\partial x_{ti}} = \frac{\partial\mathsf{P}(y_t=1|\boldsymbol{x}_{t\cdot},\boldsymbol{\beta})}{\partial x_{ti}} = f(-\boldsymbol{x}_{t\cdot}\boldsymbol{\beta})\cdot\beta_i,$$

kde $f(\cdot)$ je hustota odpovídající nějaké distribuční funkci $F(\cdot)$. Potom platí:

$$\frac{\partial\mathsf{E}(y_t|\boldsymbol{x}_{t\cdot},\boldsymbol{\beta})/\partial x_{ti}}{\partial\mathsf{E}(y_t|\boldsymbol{x}_{t\cdot},\boldsymbol{\beta})/\partial x_{tj}} = \frac{\beta_i}{\beta_j},$$

tedy podíl parametrů, lze určit jako podíl parametrů odpovídajících těmto regresorům. Občas se také může používat *preferenční poměr (odds ratio)*

$$\frac{\mathsf{P}(y_t=1|\boldsymbol{x}_{t\cdot},\boldsymbol{\beta})}{\mathsf{P}(y_t=0|\boldsymbol{x}_{t\cdot},\boldsymbol{\beta})} = \frac{1-F(-\boldsymbol{x}_{t\cdot}\boldsymbol{\beta})}{F(-\boldsymbol{x}_{t\cdot}\boldsymbol{\beta})} = \frac{F(\boldsymbol{x}_{t\cdot}\boldsymbol{\beta})}{1-F(\boldsymbol{x}_{t\cdot}\boldsymbol{\beta})},$$

který je vyjádřen podílem pravděpodobností, že daný jev nastane či nikoli. Poslední rovnost lze použít jen tehdy, když distribuční funkce $F(\cdot)$ je symetrická.

$F(\cdot)$ se volí tak, abychom s ní v praxi mohli pracovat a proto se používají jen některá rozdělení. Zde si ukážeme 3 nejčastější možné modely pro naší $F(\cdot)$.

1. *Probit:*

$$\mathsf{P}(y_t = 1|\boldsymbol{x}_{t\cdot}, \boldsymbol{\beta}) = 1 - F(-\boldsymbol{x}_{t\cdot}\boldsymbol{\beta}) = 1 - \Phi(-\boldsymbol{x}_{t\cdot}\boldsymbol{\beta}) = \Phi(\boldsymbol{x}_{t\cdot}\boldsymbol{\beta}).$$

   Za $F(\cdot)$ použijeme distribuční funkci normálního rozdělení $\Phi(\cdot)$, zcela přesně použijeme standardizované normální rozdělení $N(0, 1)$.

2. *Logit:*

$$\mathsf{P}(y_t = 1|\boldsymbol{x}_{t\cdot}, \boldsymbol{\beta}) = 1 - F(-\boldsymbol{x}_{t\cdot}\boldsymbol{\beta}) = 1 - \frac{\mathrm{e}^{-\boldsymbol{x}_{t\cdot}\boldsymbol{\beta}}}{1 + \mathrm{e}^{-\boldsymbol{x}_{t\cdot}\boldsymbol{\beta}}} = \frac{\mathrm{e}^{\boldsymbol{x}_{t\cdot}\boldsymbol{\beta}}}{1 + \mathrm{e}^{\boldsymbol{x}_{t\cdot}\boldsymbol{\beta}}}$$

   Tento model je založený na distribuční funkci logistického rozdělení a často z něj dostaneme výsledky podobné předchozímu modelu. Hustota logistického rozdělení je $f(x) = \frac{\mathrm{e}^x}{(1+\mathrm{e}^x)^2}$ a preferenční poměr je $\exp(\boldsymbol{x}_{t\cdot}\boldsymbol{\beta})$.

3. *Gompit:*

$$\begin{aligned}\mathsf{P}(y_t = 1|\boldsymbol{x}_{t\cdot}, \boldsymbol{\beta}) &= 1 - F(-\boldsymbol{x}_{t\cdot}\boldsymbol{\beta}) = 1 - \big(1 - \exp\big(-\mathrm{e}^{-\boldsymbol{x}_{t\cdot}\boldsymbol{\beta}}\big)\big) \\ &= \exp\big(-\mathrm{e}^{-\boldsymbol{x}_{t\cdot}\boldsymbol{\beta}}\big).\end{aligned}$$

   Používá se stejná distribuční funkce jako má náhodná veličina s extremálním rozdělením typu I. Toto rozdělení slouží k modelování extremální hodnoty. Od normálního a logistického rozdělení se také liší tím, že je nesymetrické a má nenulovou šikmost.

Odhad parametru $\boldsymbol{\beta}$ se většinou provádí metodou maximální věrohodnosti, tzv. ML metodou. Tuto metodu také používají softwary k výpočtu. O metodě maximální věrohodnosti viz [1, str. 146]. Naše věrohodnostní funkce

$$l(\boldsymbol{\beta}) = \prod_{t=1}^{T} \big(1 - F(-\boldsymbol{x}_{t\cdot}\boldsymbol{\beta})\big)^{y_t} \big(F(-\boldsymbol{x}_{t\cdot}\boldsymbol{\beta})\big)^{1-y_t}$$

má po zlogaritmování tvar

$$L(\boldsymbol{\beta}) = \sum_{t=1}^{T} y_t \ln\big(1 - F(-\boldsymbol{x}_{t\cdot}\boldsymbol{\beta})\big) + \sum_{t=1}^{T}(1 - y_t) \ln\big(F(-\boldsymbol{x}_{t\cdot}\boldsymbol{\beta})\big). \tag{1.4}$$

Platí-li, že máme symetrickou distribuční funkci, můžeme psát (pro logit a probit)

$$L(\boldsymbol{\beta}) = \sum_{t=1}^{T} \ln\big(F(\boldsymbol{x}_{t\cdot}\boldsymbol{\beta})\big) + \sum_{t=1}^{T}(1 - y_t) \ln\big(1 - F(\boldsymbol{x}_{t\cdot}\boldsymbol{\beta})\big).$$

Toto nyní budeme maximalizovat přes $\boldsymbol{\beta}$ a získáme tak odhad $\hat{\boldsymbol{\beta}}$.

Kvalita těchto odhadů se v praxi posuzuje pomocí tzv. *McFaddenova koeficientu* R^2_{McFadden}. Používá se podobně jako koeficient determinace. Je založen na věrohodnostním poměru

$$R^2_{\text{McFadden}} = 1 - \frac{L_U}{L_R},$$

kde L_U je maximální hodnota logaritmické věrohodnostní funkce (1.4) a L_R je také její maximální hodnota, ale při omezení $\beta_1 = \beta_2 = \cdots = \beta_k = 0$.

Zmíněné modely můžeme použít, když budeme chtít odhadnout hodnotu vysvětlované proměnné y_t pro nějaké dané vysvětlující proměnné (regresory) $\boldsymbol{x_t}$. Předpověď z modelu nám říká, že daný jev nastane (tj. $\bar{y} = 1$), když

$$\hat{\bar{\mathsf{P}}}^* = 1 - F(-\bar{\boldsymbol{x}}^\top \hat{\boldsymbol{\beta}}) \geq 0,5.$$

V softwarech je často uvedeno, pro kolik hodnot v rámci napozorovaných dat by odhadnutý model předpověděl správně, že se uvažovaný jev vyskytne ($\bar{y} = 1$).

1.2 Ordinální vysvětlovaná proměnná

Binární vysvětlovanou proměnnou můžeme zobecnit, aby již nenabývala jen dvou hodnot $(0, 1)$, ale více (přesto konečně mnoha). Pak mluvíme o *multinomické vysvětlované proměnné*.

Speciálním případem je *ordinální vysvětlovaná proměnná*, neboli uspořádaná multinomická proměnná, která nám navíc říká, jaké je pořadí hodnot u těchto proměnných. Předpokládáme, že hodnoty jsou nějak uspořádané a dá se toto pořadí určit. Např. stáří nějakého produktu v letech nebo velikost platu atd. Obvykle kategorie značíme $0, \ldots, R$.

Použijeme latentní vysvětlovanou proměnnou y^*, která je provázaná s regresory $\boldsymbol{X}$ v modelu

$$y_t^* = \boldsymbol{x}_{t\cdot}\boldsymbol{\beta} + \varepsilon_t, \tag{1.5}$$

kde ε_t jsou iid náhodné veličiny s nulovou střední hodnotou.

Vysvětlovaná proměnná y, kterou pozorujeme, je multinomická a její vztah k latentní vysvětlované proměnné y^* je následující

$$y_t = \begin{cases} 0 & \text{pro } y_t^* \le m_1, \\ 1 & \text{pro } m_1 < y_t^* \le m_2, \\ 2 & \text{pro } m_2 < y_t^* \le m_3, \\ \vdots & \\ R & \text{pro } m_R < y_t^*. \end{cases}$$

Kromě parametrů $\beta_1, \ldots, \beta_k$ a reziduálního rozptylu jsou dalšími neznámými parametry v modelu prahy $m_1, \ldots, m_R$. V tomto případě platí

$$\mathsf{P}_r = \begin{cases} \mathsf{P}(y_t = 0|\boldsymbol{x}_{t\cdot}, \boldsymbol{\beta}, \boldsymbol{m}) = F(m_1 - \boldsymbol{x}_{t\cdot}\boldsymbol{\beta}), \\ \mathsf{P}(y_t = 1|\boldsymbol{x}_{t\cdot}, \boldsymbol{\beta}, \boldsymbol{m}) = F(m_2 - \boldsymbol{x}_{t\cdot}\boldsymbol{\beta}) - F(m_1 - \boldsymbol{x}_{t\cdot}\boldsymbol{\beta}), \\ \mathsf{P}(y_t = 2|\boldsymbol{x}_{t\cdot}, \boldsymbol{\beta}, \boldsymbol{m}) = F(m_3 - \boldsymbol{x}_{t\cdot}\boldsymbol{\beta}) - F(m_2 - \boldsymbol{x}_{t\cdot}\boldsymbol{\beta}), \\ \vdots \\ \mathsf{P}(y_t = R|\boldsymbol{x}_{t\cdot}, \boldsymbol{\beta}, \boldsymbol{m}) = 1 - F(m_R - \boldsymbol{x}_{t\cdot}\boldsymbol{\beta}), \end{cases} \tag{1.6}$$

kde $F(\cdot)$ je distribuční funkce reziduí v (1.5). Obdobně jako v předchozím odstavci můžeme rozlišovat modely logit, probit a gompit. Používání kategorií $0, \ldots, R$, není podstatné. Mohli bychom je označit libovolně. Stačí jen dodržet uspořádání. Tedy že $y_s < y_t$ je ekvivalentní s $y_s^* < y_t^*$.

Pro interpretaci jednotlivých parametrů β_i, platí

$$\frac{\partial \mathsf{P}(y_t = r)}{\partial x_{ti}} = \frac{\partial F(m_{r+1} - \boldsymbol{x}_{t\cdot}\boldsymbol{\beta})}{\partial x_{ti}} - \frac{\partial F(m_r - \boldsymbol{x}_{t\cdot}\boldsymbol{\beta})}{\partial x_{ti}}, \qquad r = 1, \ldots, R-1.$$

Pomocí výpočtu z (1.6) lze ověřit, že v koncových bodech $r = 0$ a $r = R$ se pravděpodobnost $\mathsf{P}(y_t = r)$ zřejmě mění podle znaménka u parametru β_i. Tedy pro $\mathsf{P}(y_t = 0)$ platí, že se mění v opačném směru a pro $\mathsf{P}(y_t = R)$ je tomu naopak, tedy pravděpodobnost se mění ve stejném směru jako je u znaménka u parametru β_i. Obecně ale nelze říci, jaký vliv má změna regresoru x_{ti} na pravděpodobnost P_t, např. na základě znaménka β_i.

Odhad se v příslušném modelu opět provádí typicky metodou maximální věrohodnosti. Budeme chtít odhadnout parametry $\boldsymbol{\beta}$ a $\boldsymbol{m}$. Pak má logaritmická věrohodnostní funkce tvar

$$L(\boldsymbol{\beta}, \boldsymbol{m}) = \sum_{t=1}^{T} \sum_{r=0}^{R} I_r(y_t) \cdot \ln\big(\mathsf{P}(y_t = r|\boldsymbol{x}_{t\cdot}, \boldsymbol{\beta}, \boldsymbol{m})\big),$$

kde $I_r(y_t)$ je indikátor toho, zda $y_t = r$, pak $I_r(y_t) = 1$ a jinak $I_r(y_t) = 0$

Nyní ještě uveďme příklad logaritmické věrohodnostní funkce pro model logit.

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{m}) &= \sum_{t=1}^{T} I_0(y_t) \cdot \ln\left(\frac{\mathrm{e}^{m_0 - \boldsymbol{x}_{t\cdot}\boldsymbol{\beta}}}{1 + \mathrm{e}^{m_0 - \boldsymbol{x}_{t\cdot}\boldsymbol{\beta}}}\right) \\ &+ \sum_{t=1}^{T} \sum_{r=1}^{R-1} I_r(y_t) \cdot \ln\left(\frac{\mathrm{e}^{m_{r+1} - \boldsymbol{x}_{t\cdot}\boldsymbol{\beta}}}{1 + \mathrm{e}^{m_{r+1} - \boldsymbol{x}_{t\cdot}\boldsymbol{\beta}}} - \frac{\mathrm{e}^{m_r - \boldsymbol{x}_{t\cdot}\boldsymbol{\beta}}}{1 + \mathrm{e}^{m_r - \boldsymbol{x}_{t\cdot}\boldsymbol{\beta}}}\right) \\ &+ \sum_{t=1}^{T} I_R(y_t) \cdot \ln\left(1 - \frac{\mathrm{e}^{m_R - \boldsymbol{x}_{t\cdot}\boldsymbol{\beta}}}{1 + \mathrm{e}^{m_R - \boldsymbol{x}_{t\cdot}\boldsymbol{\beta}}} \cdot\right) \end{aligned}$$

Tak jako v binárním případě můžeme pro modely (probit,logit atd.) zkoumat úspěšnost předpovědí v tomto odhadnutém modelu, lze totéž provádět pro ordinální vysvětlované proměnné.

Kapitola 2

Historie modelů logit, probit a jejich užití

Následující kapitola je převážně převzatá z [3]. Zkusme se teď na chvíli podívat trochu blíže na historii a první použití modelu logit, probit atd. Idea použít model probit byla poprvé publikována v díle *Science* od Chestera Blisse v roce 1934. Bliss pracoval jako etnomolog a jeho hlavní záměr bylo nalézt pesticid, který by efektivně neutralizoval hmyz na listech grepů. Porovnáním reakce hmyzu na různé druhy pesticidu mohl pozorovat, že každý pesticid ovlivňoval hmyz jinak. Nicméně neměl statistický nástroj na porovnání těchto rozdílů. Logický postup by byl proložit regresí odezvu hmyzu proti koncentraci pesticidu. Ale vztah odezvy k dávce byl esovitého tvaru a v té době se regrese používala jen na lineární data. Proto dostal nápad, transformovat esovitou křivku závislosti na dávce na rovnou čáru. Tuto myšlenku dále rozvíjel profesor Finney, který v roce 1952 napsal knihu *Probit analysis*. Odhady pomocí tohoto modelu se staly praktické v sedmdesátých letech 20.století, když bylo možné pomocí počítačových softwarů vyřešit problém s nelineární maximalizací. A stalo se tak hlavním postupem v analýze diskrétních problémů. Dnes je stále probit model preferovaná statistická metoda k pochopení vztahů u reakce nějakých veličin na nějakou dávku (léku, protijedu atd.).

Příklad 2.1 *Stěžejní příklad uváděný téměř v každé literatuře, co se zabývá diskrétními (konkrétně binárními) vysvětlovanými proměnnými a kterým se také zabýval Bliss ve svém úplně prvním dile je studie efektu očkování na pneumokoka aplikovaného na myši. Každou myš nejdříve infikovali bakterií pneumokoka a poté dostala určitou dávku očkování proti pneumokokovi. U všech myší, které umřely během 7 dní, byla zkoumána přítomnost pneumokoka. Myši co přežily 7.*

Síla dávky	Počet úmrtí ze 40
0.0028	35
0.0056	21
0.0112	9
0.0225	6
0.0450	1

Tabulka 2.1: Tabulka ukazuje počet mrtvých myší na pneumokoka podle síly dávky.

den, považovali za přeživší a dále nezkoumali. Naše binární proměnná je, zda-li nastala smrt na pneumokoka v 7 dnech či nenastala. V tabulce (2.1) *vidíme, kolik myší ze 40 zemřelo, podle toho, kolik mg dávky očkování jim bylo aplikováno.*

Ačkoli proporce mrtvých myší a dávky očkování indikují, že se zvýšenou dávkou očkování, počet úmrtí klesá, kompletní náhled na vztah mezi pravděpodobností úmrtí a síly dávky je dán statistickým modelem. Tento model by poté mohl být použit na určení potřebné síly očkování, která by dostatečně ochránila určitý počet myší. Teď ještě ukažme, jak bychom postupovali při proložení našich dat přímkou. Tedy půjde o klasický model lineární regrese. Nechť Y_i je náhodná veličina představující počet mrtvých z n_i v i-té skupině myší, $i = 1, 2, ..., 5$, a model lineární regrese

$$\mathsf{E}\frac{Y_i}{n_i} = p_i = \beta_0 + \beta_1 d_i,$$

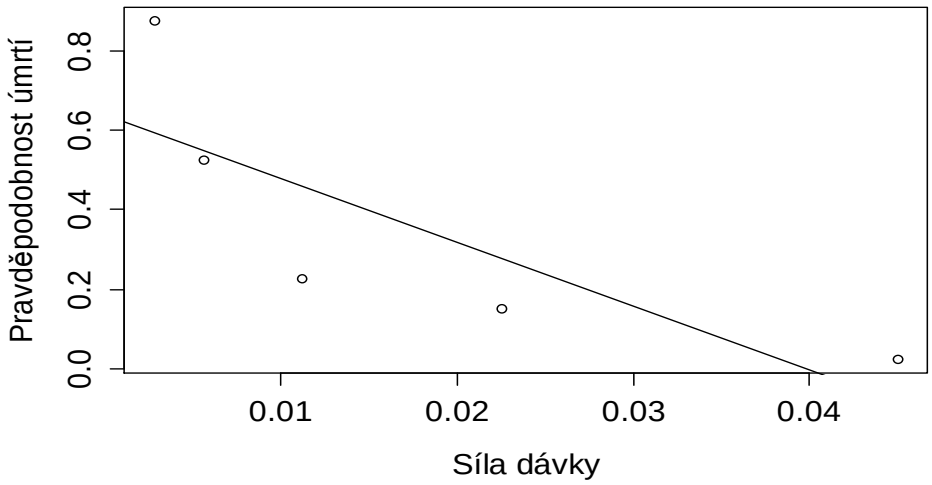
se pak použije pro proložení dat přímkou, metodou nejmenších čtverců, za předpokladu, že $\frac{Y_i}{n_i}$ má konstantní rozptyl. Odhadnutá pravděpodobnost úmrtí pro dávku o síle d_i je poté

$$\hat{p_i} = 0.64 - 16.08 d_i$$

a odhadnutá pravděpodobnost úmrtí z tohoto modelu pro myše s dávkou silnou 0.045 *pak je*

$$\hat{p_i} = 0.64 - 16.08 * 0.045 = -0.084.$$

Záporná pravděpodobnost je samozřejmě nepřijatelná. Ukažme si nyní graf naší proložené přímky.



Obrázek 2.1: Proložení dat přímkou z tabulky 2.1.

Snadno můžeme vidět, že aby pravděpodobnost úmrtí dosáhla 1*, musela by být dávka negativní a naopak, pro dávky větší jak* $0,04$*, je pravděpodobnost negativní. Kromě těchto nedostatků je taky zřejmé, že přímka zkrátka neprokládá data.*

Kvadratický výraz u síly dávky by v nějakém rozsahu proložil data lépe, ale zase by to pravděpodobně vedlo k dosažení pravděpodobnosti větší než jedna pro málo silné dávky. Nicméně hlavním problémem by bylo, že zatímco by pravděpodobnost úmrtí nejprve s rostoucí silou dávky klesala, v určitém bode by začala růst ,což je zcela v rozporu s naším modelem a biologickou představou. Konkrétně by růst začal pro dávku 0.034*. Toto všechno nám dává najevo, že klasická metoda lineárních modelů, kde přímku prokládáme pomocí metody nejmenších čtverců je zkrátka nevhodná pro modelováni reakcí u binárních dat.*

Místo používání lineárního modelu pro závislost pravděpodobnosti úspěchu (též nazývána šance - odds ratio) na vysvětlujících proměnných (regresorech), je nejprve pravděpodobnostní škála transformována z rozsahu $(0, 1)$ *na* $(-\infty, \infty)$*. Lineární model je poté adaptován na transformovanou hodnotu pravděpodobnosti úspěchu. Tato procedura zajišťuje, že po proložení bude pravděpodobnost mezi*

jedničkou a nulou. A takové transformace jsou právě logit a probit.

2.1 Porovnání modelů logit a probit

Modely logit a probit jsou si celkem velmi podobné, ale z pohledu náročnosti na výpočet, se logit jeví jako vhodnější. Existují ještě 3 další důvody, proč se logit jeví jako vhodnější než probit.

1. Má přímou interpretaci, co se týká logaritmu poměru šancí.

2. Modely založené na transformaci logit jsou také vhodné pro analýzu dat, která byla sbírána retrospektivně, např. taková, která se používají na studii kontrolování nějákého případu.

3. Poslední důvod je jaksi technický, ale lze ukázat, že binární data mohou být sumarizovány ve smyslu kvantity, známý jako postačující statistika *(sufficient statistics)*, pokud je použita logistická transformace. Bližší detaily k tomuto tématu mohou být nalezeny v kapitole 9 [3].

Příklad 2.2 *Dodejme ještě k předchozímu příkladu, že je velmi často naším zájmem odhadnout dávku, která by byla účinná pro nějaký požadovaný podíl z celého vzorku. Např. v literatuře se často uvádí tzv.* medián efektivní dávka*, která je účinná pro* 50 *procent. Nazývána též* ED_{50}*. Pokud je obsahem studii smrt, pak někdy též* medián smrtící dávka LD_{50} (lethal dose)*. Typicky by nás ovšem více zajímala situace* ED_{90}*, tedy dávka, po níž* 90 *procent myší bude v pořádku a výpočet takové potřebné dávky si nyní ukažme.*

Předpokládejme, že lineární logistický model je použit, aby popsal vztah mezi pravděpodobností reakce p *a hodnoty* d *z vysvětlující proměnné "dávka". Takže,*

$$logit(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 d.$$

Dávka, pro kterou je $p = 0.5$ *je hodnota* ED_{50} *a vzhledem k tomu, že* $logit\,(0.5) = \log 1 = 0$*, hodnota* ED_{50} *pak splňuje rovnici*

$$\beta_0 + \beta_1 ED_{50} = 0,$$

takže $ED_{50} = -\beta_0/\beta_1$*. Po proložení lineárním logistickým modelem získáme odhady* β_0, β_1 *a tak hodnota* ED_{50} *je odhadnuta*

$$\hat{ED}_{50} = -\frac{\hat{\beta}_0}{\hat{\beta}_1}.$$

Pro náš případ, abychom získali ED_{90}*, položíme* $p = 0.90$*, tedy*

$$\log\left(\frac{0.9}{0.1}\right) = \beta_0 + \beta_1 ED_{90},$$

tedy hodnota ED_{90} *je* $(2.1972 - \beta_0)/\beta_1$*, která může být odhadnuta*

$$\hat{ED}_{90} = \frac{(2.1972 - \beta_0)}{\beta_1}.$$

Kdybychom raději místo charakteristiky dávka použili $\log$*(dávka) jako vysvětlující proměnnou, rovnice modelu by poté musela být upravena. A to sice*

$$logit(p) = \beta_0 + \beta_1 \log(d).$$

Hodnota ED_{50} *se pak získá z*

$$\beta_0 + \beta_1 \log\left(ED_{50}\right) = 0,$$

tedy $ED_{50} = \exp\left(-\beta_0/\beta_1\right)$*. Což se odhadne*

$$\hat{ED}_{50} = \exp\left(-\frac{\hat{\beta}_0}{\hat{\beta}_1}\right).$$

Podobně se pak odhadne hodnota ED_{90}

$$\hat{ED}_{90} = \exp\left(\frac{2.1972 - \beta_0}{\beta_1}\right).$$

$\hat{\beta}_0$ *a* $\hat{\beta}_1$ *se pak dopočítají metodou maximální věrohodnosti a dojdeme k tomu, že požadovaná dávka, která efektivně ochrání* 90 *procent myší je* 0.0219*.*

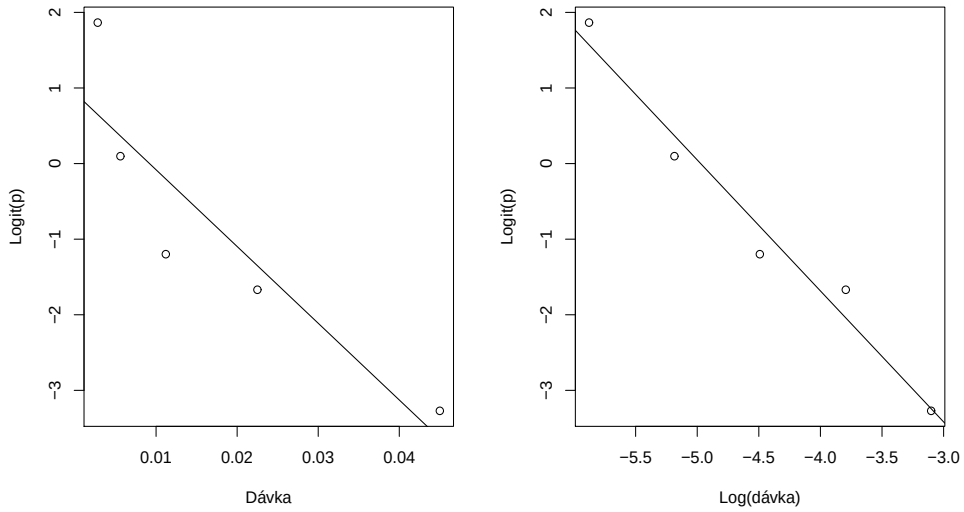
Odhady ED_{50} *a* ED_{90} *mohou být získány podobně i v modelu probit.*

Pokud bychom jako závislou proměnnou použili $\log$*(dávka), vypadal by model pro probit následovně*

$$probit(p) = \Phi^{-1}(p) = \beta_0 + \beta_1 \log(d).$$

Pro $p = 0.5$ *je* $probit(0.5) = 0$ *a tak stejně jako pro model logit je hodnota* ED_{50} *odhadnuta* $\hat{ED}_{50} = \exp\left(-\hat{\beta}_0/\hat{\beta}_1\right)$*.*

Když $p = 0.9$*,* $probit(0.9) = 1.2816$ *a tak hodnota* ED_{90} *se odhadne jako* $\exp\left((1.2816 - \hat{\beta}_0)/\hat{\beta}_1\right)$ *a* $\hat{\beta}_0$*,* $\hat{\beta}_1$ *bychom odhadli v modelu probit.*



Obrázek 2.2: Porovnání použití vysvětlujících proměnných dávka a log(dávka).

Poznámka 2.1 *Při odhadování veličin jako* ED_{50} *a podobně nás může zajímat standartní odchylka a interval spolehlivosti, aby náš odhad mohl být posouzen. Postupy pro tyto výpočty lze nalézt v kapitole 4.2.1 a 4.2.2 knihy [3].*

Nechť y_i *je počet mrtvých z* n_i *v i-té skupině myší,* $i = 1, 2, ..., 5,$. *Graf empirických logitů,* $\log\left((y_i + 0.5)/(n_i - y_i + 0.5)\right)$ *k vysvětlující proměnné dávka a* log*(dávka) je ukázán* (2.2).

(2.2) *jednoznačně naznačuje, že vztah mezi logit(p) a* log*(dávka) je více lineární, než u logit(p) a dávka. Nicméně toto si teď potvrdíme analýzou odchylky* (analysis of deviance). *Odchylka v proložení pozorovaných dat v modelu logistické regrese, pokud jako vysvětlující proměnnou použijeme* log*(dávka), je 2.81 se 3 stupni volnosti. Vzhledem k tomu, že je odchylka tak blízko k počtu stupňů volnosti, na kterých je založena, zdá se, že model data prokládá dobře. Pro porovnání, pokud bychom v modelu místo* log*(dávka) použili dávka, odchylka by byla 15.9 na 3 stupních volnosti.*

Spolu s tímto a s předchozím obrázkem, kde je zřejmé, že regresní model, který využívá vysvětlující proměnnou dávka místo log*(dávka), prokládá data hůře, je náš závěr, že by tento druhý model byl nepostačující.*

Výše jsme tedy ukázali, že jako nejvhodnější model pro proložení dat je

v tomto případě lineární logistický model pro pravděpodobnost úmrtí, kde se vyskytuje logaritmus dávky jako vysvětlující proměnná.

Tedy proložený model je

$$logit(p) = 9.19 - 1.83 \log(d).$$

a hodnota ED_{50} *se pak odhadne*

$$\hat{ED}_{50} = \exp\left(-9.19/1.83\right) = 0.0066cc.$$

Na závěr je ještě vždy dobré porovnat odhadnutou hodnotu ED_{50} *s napozorovanými daty, zda-li se odhad jeví správný. V tomto případě tomu tak je, protože při dávce* 0.0056 *cc je proporce úmrtí* $21/40$*, což je velmi podobné našemu výpočtu a to sice, že dávce* 0.0066 *odpovídá proporce úmrtí* $20/40$ *tedy* ED_{50}.

Kapitola 3

Omezené vysvětlované proměnné

Nyní budeme zkoumat spojité vysvětlované proměnné, kde ale nebudeme znát celou informaci z dat.

Budeme vycházet z následujícího modelu

$$y_t^* = \boldsymbol{x}_{t\cdot}\boldsymbol{\beta} + \sigma \cdot \varepsilon_t, \tag{3.1}$$

kde proměnnou y^* budeme pozorovat skrz nějakou proměnnou y.

Vysvětlované proměnné mohou nabývat hodnot, které jsou jen částečně známé, např. pouze víme, že něčí plat přesahuje hranici 100 000, ale již nevíme o kolik. V takovém případě všechny tyto platy budeme považovat za rovny 100 000 a mluvíme o *cenzorované vysvětlované proměnné*. Takové označení má svůj smysl, protože zachovává alespoň částečně důležitou informaci, která pak může být použita. Např. vhodný kandidát na koupi drahého vozu nebo bychom naopak mohli cenzorovat platy zleva a dostali bychom nevhodné kandidáty na koupi drahého vozu.

Jiný případ by nastal, pokud bychom neznámou informaci chtěli zcela vypustit z našeho vzorku. To si můžeme představit u zkoumaného vzorku lidí, kterým je nabídnuto investovat do nějakého projektu. Kdybychom zkoumali výši investic, tak bychom určitě dospěli k tomu, že mnoho lidí by vůbec neinvestovalo a jak tedy s tím naložit. Takové klienty bychom zcela vypustili ze zkoumaného vzorku, jako bychom „odsekli“ všechny, jejichž investice je rovna 0 a mluvíme o *useknuté vysvětlované proměnné*.

3.1 Cenzorované veličiny

Jak už jsme zmínili v předchozím textu, cenzorovaná veličina se tedy mimo své hraniční body intervalu nahrazuje danými body, což si nyní formálně zapíšeme.

Definice 3.1 (Cenzorovaná veličina) *Nechť y^* je spojitá latentní veličina, ale my pozorujeme jen veličinu y. Dále mějme meze $d_t < h_t$ pro každé pozorování t. Pak cenzorovanou veličinou nazveme veličinu, která se chová následovně*

$$y_t = \begin{cases} d_t & pro\ y_t^* \leq d_t, \\ y_t^* & pro\ d_t < y_t^* < h_t \\ h_t & pro\ h_t \geq y_t^*. \end{cases}$$

Častým případem je, že $d_t = d$ a $h_t = h$ pro všechna $t \in T$. Poté označíme meze jen d a h.

Když $d = -\infty$ říkáme, že neprovádíme cenzorování zleva. Pokud $h = \infty$ neprovádíme cenzorování zprava.

Důležitý případ nastane,

$$y_t = \begin{cases} 0 & \text{pro } y_t^* \leq 0, \\ y_t^* & \text{pro } y_t^* > 0 \end{cases}$$

Tedy pokud $d = 0, h = \infty,$. Pokud je reziduální složka normálně rozdělena, mluvíme o *modelu tobit.* Viz [3].

Podívejme se nyní, jak je to s marginálními efekty u modelu tobit.

Dospějeme k

$$\frac{\partial \mathsf{E}(y_t)}{\partial x_{ti}} = \beta_i \Phi_t.$$

Odhad parametrů σ a $\boldsymbol{\beta}$ se provede metodou maximální věrohodnosti. Provádí se tedy maximalizací logaritmické věrohodnostní funkce tvaru,

$$\begin{aligned} l(\boldsymbol{\beta}, \sigma) \;=\; & \sum_{i=1}^{T} \left\{ I_{(-\infty, d_t\rangle}(y_t) \cdot \ln F\left(\frac{d_t - \boldsymbol{x}_{t\cdot}\boldsymbol{\beta}}{\sigma}\right) + \right. \\ & I_{(d_t, h_t)}(y_t) \cdot \ln f\left(\frac{y_t - \boldsymbol{x}_{t\cdot}\boldsymbol{\beta}}{\sigma}\right) - I_{(d_t, h_t)}(y_t) \cdot \ln(\sigma) \\ & \left. I_{\langle h_t, \infty)}(y_t) \cdot \ln\left(1 - F\left(\frac{h_t - \boldsymbol{x}_{t\cdot}\boldsymbol{\beta}}{\sigma}\right)\right) \right\}, \end{aligned}$$

kde f a F jsou jako obvykle hustota a distribuční funkce daného rozdělení.
Také uvedeme logaritmickou věrohodnostní funkci pro model tobit

$$\begin{aligned} l(\boldsymbol{\beta},\sigma) &= \sum_{i=1}^{T}\left\{ I_{(-\infty,0\rangle}(y_t)\cdot\ln\left(\Phi\left(\frac{\boldsymbol{x}_{t\cdot}\boldsymbol{\beta}}{\sigma}\right)\right)+\right. \\ &\left. I_{[0,\infty)}(y_t)\cdot\ln\left(-\frac{1}{2}\ln(2\pi)-2\ln(\sigma)-\frac{1}{2\sigma^2}(y_t-\boldsymbol{x}_{t\cdot}\boldsymbol{\beta})^2\right)\right\}. \end{aligned}$$

Můžeme napsat vzorec pro odhad cenzorované veličiny
v tomto modelu

$$\hat{y_t} = \mathsf{E}(y_t|\boldsymbol{x}_{t\cdot},\hat{\boldsymbol{\beta}},\hat{\sigma}) = \Phi(\boldsymbol{x}_{t\cdot}\hat{\boldsymbol{\beta}}/\hat{\sigma})\cdot\boldsymbol{x}_{t\cdot}\hat{\boldsymbol{\beta}}+\hat{\sigma}\cdot\varphi(\boldsymbol{x}_{t\cdot}\hat{\boldsymbol{\beta}}/\hat{\sigma}),$$

kde $\hat{\boldsymbol{\beta}}$ a $\hat{\sigma}$ jsou odhady $\boldsymbol{\beta}$ a σ.

3.2 Useknuté veličiny

Nyní se stručně zabývejme useknutými veličinami.

Rozdíl mezi *useknutou* a *cenzorovanou* vysvětlovanou proměnnou jsme již vysvětlovali na začátku kapitoly (3). Co znamená useknutí zleva d_t či zprava h_t v modelu (3.1) to si nyní zase formálně zapíšeme.

Definice 3.2 (Useknutá veličina) *$y_t = y_t^*$, platí jen v případě, že $d_t < y_t^* < h_t$, jinak se y_t vůbec nebere v úvahu („usekne se“ ze vzorku).*

Znamená to tedy, že když jsou data mimo interval (d_t, h_t), zcela je vyloučíme. Taková situace může nastat, pokud některé informace nejsou například zveřejňovány nebo pokud při nějakém měření nejsme schopní určit důvody, jež vedou k tomu, že jsme na přístroji žádnou hodnotu nenaměřili (byla hodnota příliš velká, malá nebo nastala chyba měření, ale my to nejsme schopní určit).

Parametry se opět odhadnou metodou maximální věrohodnosti. Logaritmická věrohodnostní funkce bude vypadat následovně:

$$\begin{aligned} l(\boldsymbol{\beta},\sigma) &= \sum_{\substack{t=1\\ d_t<y_t^*<h_t}}^{T}\left\{\ln f\left(\frac{y_t-\boldsymbol{x}_{t\cdot}\boldsymbol{\beta}}{\sigma}\right)-\ln(\sigma)\right. \\ &\left.-\ln\left(F\left(\frac{h_t-\boldsymbol{x}_{t\cdot}\boldsymbol{\beta}}{\sigma}\right)-F\left(\frac{d_t-\boldsymbol{x}_{t\cdot}\boldsymbol{\beta}}{\sigma}\right)\right)\right\}. \end{aligned}$$

3.3 Proměnné vyjadřující dobu trvání

Vysvětlovaná proměnná teď bude představovat čas resp. dobu trvání nějaké situace. Můžeme tedy zkoumat, kdy nějaká situace nastane, často kdy daný jedinec zemře, nebo kdy se nějaký stroj rozbije a přestane fungovat. Odtud také pochází název, který je běžný pro tyto analýzy a to sice *analýza přežití (survival analysis)*.

Je zajímavé, že většina příkladů uváděných na toto téma má medicínskou tématiku. Takové příklady mohou být nalezeny v [4].

Je zřejmé, že se často setkáme s daty, u nichž ještě daný čas nenastal, ale my už je dále nepozorujeme. Např. pacient, co přestane být pozorován, ale třeba můžeme být limitování dnešním dnem, když zkoumáme, zda se stroj již rozbil či nikoli atd. Takovýto zadaný limit nám pak může vysvětlovanou proměnnou vhodně limitovat. Tedy vysvětlovanou proměnnou doby trvání můžeme mít

- cenzorovanou - neukončená pozorování do daného limitu se v původním souboru zanechají s tou hodnotou, která odpovídá tomuto limitu.

- useknutou - neukončená pozorování do daného limitu se z původního souboru vyloučí.

Pak definujeme funkci přežití

$$S(\tau) = \mathsf{P}(y_t > \tau) = 1 - F(\tau),$$

a intenzitu úmrtnosti

$$\lambda(\tau) = \lim_{\delta \to 0^+} \frac{\mathsf{P}(\tau < y_t \leq \tau + \delta | y_t > \tau)}{\delta}\,.$$

Názvy jsou dané hlavně tím, že svojí aplikaci mají převážně v analýze dat s medicínským podtextem, teorií spolehlivosti a životním pojištěním.

Protože mezi oběma nástroji existuje vzájemně jednoznačný vztah, odhadujeme právě intenzitu úmrtnosti místo hustoty pravděpodobnosti $f(\tau)$.

Platí, že

$$\lambda(\tau) = \lim_{\delta \to 0^+} \frac{F(\tau+\delta) - F(\tau)}{\delta} \frac{1}{S(\tau)} = \frac{f(\tau)}{S(\tau)} = -\frac{\partial \log S(\tau)}{\partial \tau}\,.$$

V praxi se obvykle parametrický tvar intenzity úmrtnosti volí následovně:

1. *Exponenciální model* doby trvání, který má intenzitu úmrtnosti

$$\lambda(\tau) = \gamma,$$

kde hustota je ve tvaru $f(\tau) = \gamma \exp(-\gamma\tau)$.

2. *Weibullův model* s intenzitou úmrtnosti

$$\lambda(\tau) = \alpha\gamma\tau^{\alpha-1}.$$

Hustota se zapíše ve tvaru $f(\tau) = \alpha\gamma\tau^{\alpha-1}\exp(-\gamma\tau^{\alpha})$. Když se $\alpha := 1$, přejde Weibullův model na model exponenciální.

3. *Logaritmicko-normální model*

$$\lambda(\tau) = \phi\left(\frac{\ln\tau}{\sigma}\right) \Big/ \sigma\tau\left(1 - \Phi\left(\frac{\ln\tau}{\sigma}\right)\right).$$

kde hustota je ve tvaru $\ln(y_t)$ má normální rozdělení μ a σ^2.

4. *Model s proporcionální intenzitou úmrtnosti,* neboli *Coxův model*

$$\lambda_t(\tau) = \exp(\boldsymbol{x}_{t.}\boldsymbol{\beta})\lambda_0(\tau),$$

kde $\lambda_0(\tau)$ je *bazická intenzita úmrtnosti (baseline hazard function)* a není závislá na čase t. Často je normována, aby pak vektor regresorů $\boldsymbol{x}_{t.}\boldsymbol{\beta}$ nezahrnoval intercept. Po zlogaritmování tohoto modelu dostaneme model s lineárními regresory,

$$ln\lambda_t(\tau) = \boldsymbol{x}_{t.}\boldsymbol{\beta}) + ln\lambda_0(\tau),$$

Tento model je oblíbeným modelem doby trvání v ekonometrických aplikacích.

Funkce přežití Coxova modelu má tvar

$$S_t(\tau) = S_0(\tau)^{\exp(\boldsymbol{x}_{t.}\boldsymbol{\beta})}.$$

Bazická intenzita úmrtnosti a parametry $\boldsymbol{\beta}$ se odhadnou metodou maximální věrohodnosti.

Často používaný neparametrický model je Kaplan-Meierův [6], který i my používáme v příkladu 4.1 a tak se ještě trochu budeme zabývat jím. Idea Kaplan-Meierova odhadu spočívá v podmíněné pravděpodobnosti. Na začátku předpokládáme, že jsou všechny objekty zkoumání naživu, tedy

$$\mathsf{P}(T > t_0 = 0) = 1.$$

Podmíněná pravděpodobnost pak vypadá

$$\mathsf{P}(T > t_i | T > t_{i-1}) = \frac{n_i - d_i}{n_i},$$

kde n_i je počet jedinců podléhající danému riziku v čase t_i a d_i je počet jedinců, kteří zemřou v čase t_i.

Příklad 3.1 *V následujícím příkladu se pokusíme demonstrovat využití funkce přežití na konkrétních datech a porovnáme popisované modely pro parametrický tvar intenzity úmrtnosti. Máme data od firmy SC Johnson, která se specializuje na prodej výrobků pro domácnost. Data budeme analyzovat pomocí softwaru R.*

Jak použít software R pro analýzu přežití je poměrně dobře popsáno např. v knize [5]. Také lze čerpat z článku [9].

Pozorujeme celkem 65 výrobků prodávaných za posledních 5 let (2009-2013). Budeme předpokládat, že se někdy každý výrobek přestane prodávat. Naším cílem bude právě odhadnout, za kolik měsíců tato situace nastane, při daných charakteristikách výrobku. Z našich 65 výrobků se na konci roku 2013 stále prodává 52. Tyto data budeme cenzorovat.

Ještě zmíníme, že jsme úplně na začátku ze vzorku vyloučili všechny výrobky, jež se v daných 5 letech prodávaly v méně jak v 10 měsících (z 60 celkových). Existuje rozumný důvod se domnívat, že by tyto výrobky pouze zkreslily naše výsledky. Ta pozorování, kde se výrobek i na konci našeho 5-letého cyklu prodává, budeme cenzorovat počtem měsíců v té době.

V následující tabulce ukážeme všechny charakteristiky, které budeme brát v úvahu a blíže popíšeme naše data:

Název	*Popis*	*Hodnoty*
hmot	*Hmotnost výrobku v gramech.*	$\mathbb{Z}$
holder	*Může-li být použit samostatně (holder), či jde o doplněk (refill) (1=holder).*	*0, 1*
rekl	*Má-li výrobek reklamu v TV či jinde (1=ano).*	*0, 1*
nakl	*Náklady na výrobu v korunách.*	$\mathbb{R}^+$
cena	*Prodejní cena v korunách.*	$\mathbb{R}^+$
mesicu	*Kolik měsíců se výrobek prodává, či prodával.*	*1,...,60*
cens	*Zda se výrobek stále prodává (1=ano).*	*0, 1*

Aby bylo možné v R analýzu provést, je třeba nainstalovat některé knihovny: `survival` *a* `Rms`*. Je také nutné mít nainstalovanou knihovnu* `splines`*.*

Knihovny se zavolají příkazem `library`*.*

Po načtení všech knihoven načteme data příkazem `data_SCJ=read.table ("data_SCJ.txt", header = TRUE)`*.*

Abychom mohli všechny veličiny obsažené v datech volat přímo, použijeme příkaz `attach(data_SCJ)`*.*

Ještě je třeba změnit veličinu `cens`*, neboť R přiřazuje 0, pokud je pozorování cenzorováno, ale my jsme přiřazovali 1 pokud byl výrobek stále v prodeji a je tedy pozorování cenzorováno. Proto je důležitý příkaz* `cens=abs(cens-1)`*.*

Kdybychom si chtěli ukázat exponenciální, Weibullův a Logaritmicko-normální model pro všechny proměnné pro SC Johnson výrobky, pak bychom zadali takovýto vstup

`summary(survreg( Surv(mesicu, cens) ~ obsah + reklama + cena + nakl + pack, dist= "exponential"))`*.*

Pro zbývající 2 modely bychom vždy jen v příkazu změnili "dist" na odpovídající model.

Použijeme výstup, který se shoduje s výstupem v R pro Weibullův model

```
##
## Call:
## survreg(formula = Surv(mesicu, cens) ~ obsah + reklama + cena +
##     nakl + pack, dist = "weibull")
##                Value Std. Error      z        p
## (Intercept)  4.56778    0.38532 11.855 2.04e-32
## obsah        0.00232    0.00155  1.497 1.34e-01
## reklama      0.25445    0.26405  0.964 3.35e-01
```

```
## cena         -0.01175    0.00428 -2.747 6.02e-03
## nakl          0.03097    0.02873  1.078 2.81e-01
## pack         -0.09908    0.39342 -0.252 8.01e-01
## Log(scale)  -1.03305    0.23536 -4.389 1.14e-05
##
## Scale= 0.356
##
## Weibull distribution
## Loglik(model)= -68.2   Loglik(intercept only)= -77.5
##  Chisq= 18.51 on 5 degrees of freedom, p= 0.0024
## Number of Newton-Raphson Iterations: 9
## n= 65
```

Vidíme, že P-hodnota je u modelu menší než 0.05. U zbývajících modelů je tomu také tak a tak jsou všechny modely přípustné.

Pro parametry z Weibullova modelu platí, že $\alpha = \frac{1}{Scale}$*, můžeme tedy soudit, že Weibullův model je vhodnější než exponenciální, který by vypustil parametr Scale.*

Pro Coxův model by se příkaz malinko lišil.

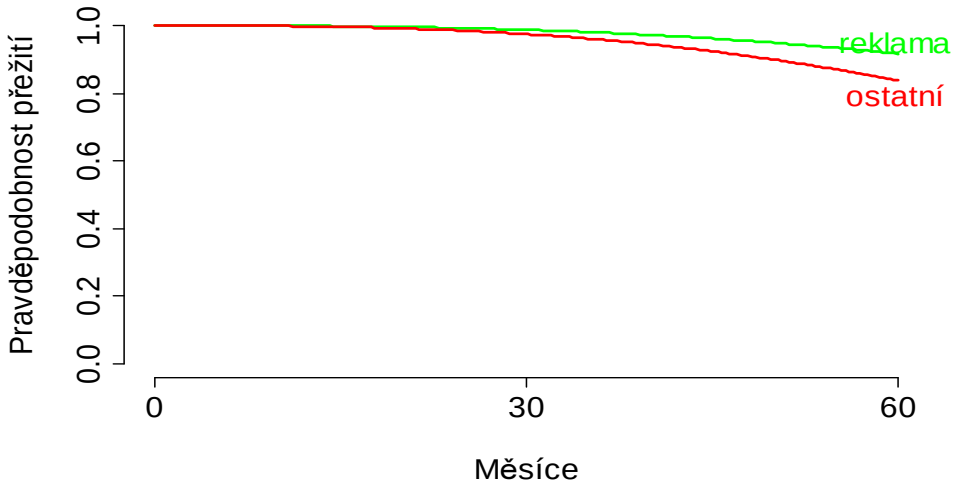
`summary(coxph(Surv(mesicu, cens) ~ obsah + reklama + cena + nakl + pack))`

Pokud bychom vzali jen charakteristiky, u kterých je p-hodnota menší než 0.05, pak bychom do všech 4 modelů zahrnuli jen "obsah" a "cenu".

Naším dalším cílem bude ukázat, jestli má reklama vliv na počet měsíců, jak dlouho se výrobek prodává. Tedy budeme zkoumat rozdíl mezi funkcí přežití u výrobku s reklamou (reklama=1) a bez ní (reklama=0). Použijeme Weibullův model se všemi veličinami, které budou nabývat svých průměrných hodnot.

Grafy funkcí přežití pro výrobky s reklamou a bez ní pak získáme v R následovně:

```
rekl=factor(reklama,levels=c(1,0),labels=c("reklama","ostatní")),
dd <- datadist(mesicu, rekl),
options(datadist='dd'),
fit=psm(Surv(mesicu, cens) ~ obsah + pack + cena + nakl + rekl,
dist="weibull"),
survplot(fit, gl=NA, obsah=mean(obsah), cena=mean(cena),nakl=mean
(nakl),pack=mean(pack), xlab="Měsíce",ylab="Pravděpodobnost přežití",
col=c("green","red"), lty=c(1,1),lwd=c(2,2)).
```



Obrázek 3.1: Funkce přežití pro reklama=1 a pro reklama=0 (Weibullův model).

Vidíme, že výrobky s reklamou mají vyšší pravděpodobnost dožití se delší doby na trhu. Z našich dat je většina pozorování v prodeji dlouhou dobu a taky jsou cenzorována (stále v prodeji), proto je pokles křivek velmi malý. Typičtější příklad, kde se v určitém časovém úseku předpokládá, že se pravděpodobnost přežití dostane až na nulu, je uveden v (4).

Kapitola 4

Porovnání softwarů pro analýzu přežití

V následující kapitole se zaměříme na porovnání 3 relativně dost odlišných a přesto poměrně hojně užívaných softwarů pro statistické analýzy, které lze použít při analýze přežití. Nejprve si řekneme výhody a nevýhody jednotlivých softwarů a pak si ukážeme jejich použití na 2 příkladech.

1. První bude typický příklad pro analýzu přežití, tedy situaci, kdy pozorujeme nějakou skupinu pacientů a zajímá nás, kolik jich v určité periodě zemřelo a kolik ne. Dále předpokládáme, že některé pacienty ztratíme ze vzorku, což jsou naše cenzorované veličiny. Použijeme neparametrickou Kaplan-Meierovu metodu.

2. Poté ještě ukážeme analýzu přežití na příkladu s našimi daty. Zde se pokusíme porovnat parametrické modely popisované v 3 kapitole.

Poznámka 4.1 *Data pro první příklad jsou převzatá z knihy [8]. Jedná se o 100 pacientů infikovaných nemocí AIDS. Známe jen délku nemoci u pacientů, jejich věk a zda jsou či již nejsou naživu. Pokud ano, pak data cenzorujeme zprava. Data pro druhý příklad jsou od firmy SC Johnson a již jsme je použili v* (3.1)*, kde jsme je také blíže popsali.*

Poznámka 4.2 *Ještě bychom upozornili na to, jak správně pracovat s proměnnou "cens". Vyskytuje se v obou případech, ale mohla by nás zmást její interpretace. Hlavně proto, že R a Mathematica odlišně nahlížejí na to, kdy proměnou považují za cenzorovanou a kdy nikoli. V R je to v případě, že cens* $= 0$*, ale*

v Mathematice tehdy, když cens = 1. Zároveň v obou příkladech má cens jiný význam. V příkladu s pacienty cens = 0 znamená, že je daný pacient naživu a pozorování jsou cenzorována, zatímco v příkladu s výrobky znamená cens = 0, že výrobek se již neprodává a cenzorovaná pozorování jsou ty, kde cens = 1. Jde jen o lehký technický problém, nad kterým je ale třeba se před analýzou zamyslet, abychom nedostali špatný výsledek.

4.1 Excel

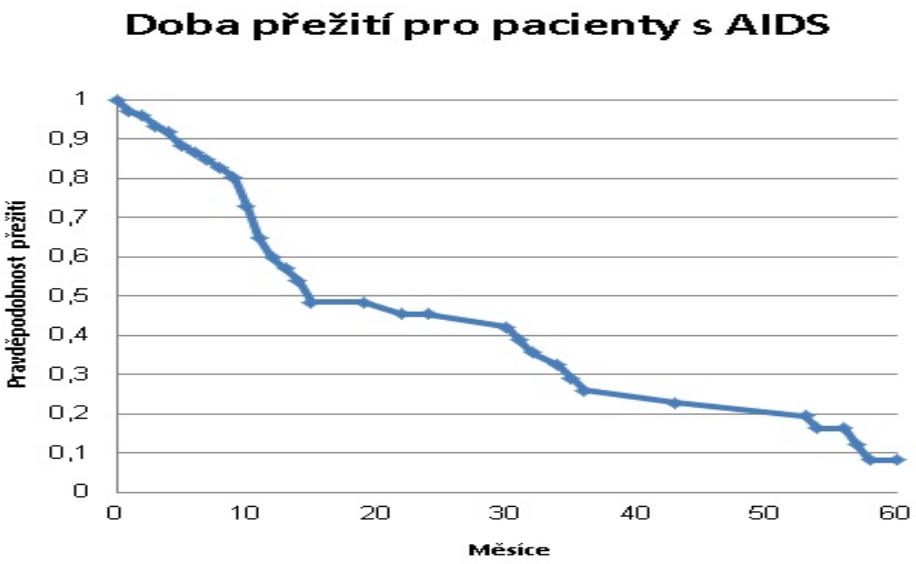
Začneme s Excelem od Microsoft Office. Ten se samozřejmě velmi nabízí pro svojí relativní jednoduchost. Vzhledem k tomu, že ve zbývajících dvou softwarech budeme řešit problém načtení dat z Excelu (protože Excel se často může volit jako hezký a přehledný zdroj pro naše data) respektive z textových souborů kam jsou data nahraná právě z Excelu, má v tomto ohledu Excel jednoznačnou výhodu. Tím ale jeho pozitiva končí.

V Excelu je možné používat jenom neparametrické modely. Nemá žádné implementované funkce a jediný způsob jak si vytvořit funkci přežití, je si celý vývoj situace namodelovat, resp. sepsat do řádku pod sebe postupně situace pro každý časový úsek a počítat průběžně pravděpodobnosti pro další a další časové periody a tím i události, které nastaly. Tady je ještě předtím třeba použít kontingenční tabulku a data si vhodně zpracovat, pro každý časový úsek si tak musíme spočítat počet jedinců, které ještě zkoumáme (v tabulce pacienti = přeživší - cenzorovaní) a ty co skutečně zemřou. Poté se již graf funkce přežití získá snadno z pravděpodobností, které s rostoucím časem jdou k nule.

Zde je ukázka, jak taková tabulka vypadá.

Čas	Pacienti	Cenzorovaná	Úmrtí	Přeživší	Pravděpodobnost
1	100	14	3	97	0,9700
2	83	9	1	82	0,9583
3	73	10	2	71	0,9321
4	61	4	1	60	0,9168
5	56	5	2	54	0,8840
6	49	2	1	48	0,8660
7	46	6	1	45	0,8472
8	39	3	1	38	0,8254
9	35	2	1	34	0,8019
10	32	1	3	29	0,7267
11	28	0	3	25	0,6488
12	25	2	2	23	0,5969
13	21	0	1	20	0,5685
14	20	0	1	19	0,5401
15	19	0	2	17	0,4832
19	17	1	0	17	0,4832
22	16	0	1	15	0,4530
24	15	1	0	15	0,4530
30	14	0	1	13	0,4207
31	13	0	1	12	0,3883
32	12	0	1	11	0,3559
34	11	0	1	10	0,3236
35	10	0	1	9	0,2912
36	9	0	1	8	0,2589
43	8	0	1	7	0,2265
53	7	0	1	6	0,1942
54	6	0	1	5	0,1618
56	5	1	0	5	0,1618
57	4	0	1	3	0,1213
58	3	0	1	2	0,0809
60	2	2	0	2	0,0809

Níže je graf funkce přežití, který se získá snadno z prvního a posledního sloupce.

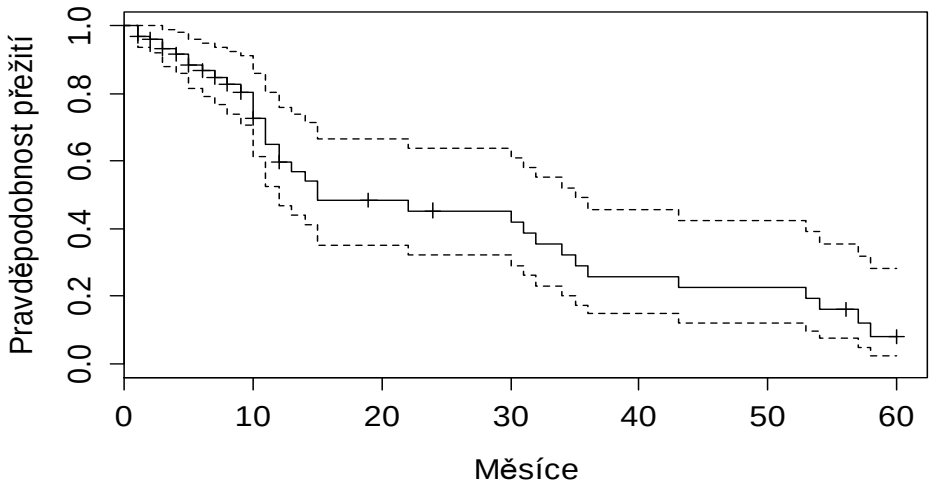


Obrázek 4.1: Funkce přežití pro pacienty s AIDS v Excelu.

Jak jsme zmínili, parametrické modely v Excelu zkoumat nelze a druhý příklad tak lze řešit jen v následujících softwarech.

4.2 R

V Rku resp. RStudiu lze poměrně dobře využívat naprogramované funkce. K tomu je třeba nahrát některé knihovny. A to konkrétně MASS, Splines a Survival. Ty se nahrají příkazem library. Pak lze využít funkce Survfit a Surv a s pomocí základní funkce plot můžeme získat graf pro funkci přežití. Jedná se o Kaplan-Meierovu metodu pro získání funkce přežití, proto naše označení "*kmsurvival*". Nejprve do Rka načteme data, `data = read.table("Pacienti.txt", header=TRUE)`, abychom mohli s daty zase pracovat přímo a volat jejich proměnné, tak použijeme příkaz `attach(data)`. Graf funkce přežití pro Kaplan-Meierovu metodu se získá z dvou následujících příkazů:
`kmsurvival = survfit(Surv(cas,cens) ~ 1)` a
`plot(kmsurvival, xlab = "Měsíce", ylab = "Pravděpodobnost přežití")`.



Obrázek 4.2: Funkce přežití pro pacienty s AIDS v Rku.

Pokud by nás zajímali pravděpodobnosti přežití v jednotlivých časech, je možné si je zobrazit následujícím příkazem `summary(kmsurvival)`.

Výstup z Rka, kde nás bude zajímat sloupec "*survival*", pak vypadá takto.

```
Call: survfit(formula = Surv(cas, cens) ~ 1)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1    100       3   0.9700  0.0171       0.9371        1.000
    2     83       1   0.9583  0.0205       0.9190        0.999
    3     73       2   0.9321  0.0270       0.8805        0.987
    4     61       1   0.9168  0.0306       0.8587        0.979
    5     56       2   0.8840  0.0373       0.8139        0.960
    6     49       1   0.8660  0.0406       0.7899        0.949
    7     46       1   0.8472  0.0439       0.7654        0.938
    8     39       1   0.8254  0.0478       0.7368        0.925
    9     35       1   0.8019  0.0520       0.7062        0.910
   10     32       3   0.7267  0.0627       0.6137        0.860
   11     28       3   0.6488  0.0702       0.5248        0.802
   12     25       2   0.5969  0.0736       0.4688        0.760
```

13	21	1	0.5685	0.0754	0.4384	0.737
14	20	1	0.5401	0.0768	0.4087	0.714
15	19	2	0.4832	0.0785	0.3514	0.664
22	16	1	0.4530	0.0792	0.3216	0.638
30	14	1	0.4207	0.0799	0.2899	0.610
31	13	1	0.3883	0.0800	0.2593	0.582
32	12	1	0.3559	0.0796	0.2296	0.552
34	11	1	0.3236	0.0787	0.2009	0.521
35	10	1	0.2912	0.0772	0.1732	0.490
36	9	1	0.2589	0.0751	0.1466	0.457
43	8	1	0.2265	0.0723	0.1211	0.424
53	7	1	0.1942	0.0689	0.0969	0.389
54	6	1	0.1618	0.0645	0.0740	0.354
57	4	1	0.1213	0.0598	0.0462	0.319
58	3	1	0.0809	0.0517	0.0231	0.283

Zde si všimněme, že hodnoty survival odpovídají spočítaným hodnotám z Excelu, tedy v R si můžeme zkontrolovat, že jsme prvně postupovali správně.

Co se týká druhého příkladu a porovnání parametrických modelů, tuto analýzu jsme již Rkem provedli v příkladu (3.1).

4.3 Mathematica

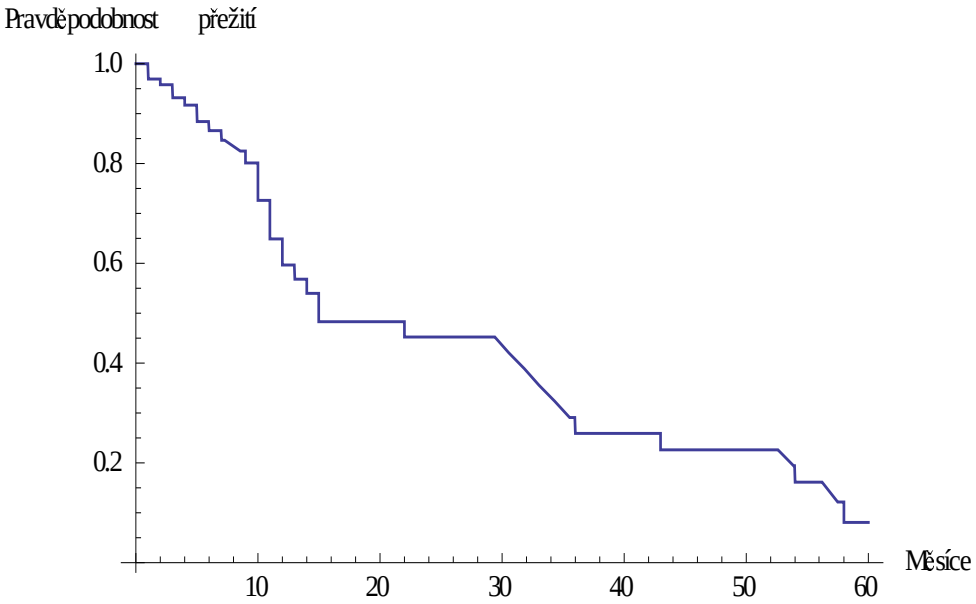
V Mathematice podobně jako v Rku jsou již implementovány funkce pro analýzu přežití.

Bohužel načtení dat se jeví jako méně pohodlné. Po zadání příkazu `Import["data.txt"]` *získáme sice data, ale nelze s nimi pracovat tak pohodlně jako v Rku. Jako nejlepší způsob se zdá načíst zvlášť proměnnou čas a cens.*

Poté použijeme funkci `SurvivalModelFit`*, která automaticky vybere nejlepší vhodný model na zadaná data. Pro cenzorování zprava je to Kaplan-Meierův odhad. Pro jiné typy cenzorování je odhad zkonstruován pomocí jiných přístupů. Příkazy do Mathematicy by pro vykreslení funkce přežití vypadaly takto*

```
S=SurvivalModelFit[EventData[cas,cens]],
Plot[S[x],{x,0,60}],Exclusions -> None, AxesLabel -> {Měsíce,
Pravděpodobnost přežití}.
```

Zde je graf funkce pravděpodobnosti přežití pro pacienty s AIDS.



Obrázek 4.3: Funkce přežití pro pacienty s AIDS v Mathematice.

Pro druhý příklad, kde chceme porovnat parametrické modely, bychom postupovali zdánlivě stejně, tedy začali funkcí `d = EventData[cas,cens]`*, ale potom bychom odhadli parametr pro naše parametrické rozdělení, pomocí funkce* `EstimatedDistribution`*, a to sice takto*

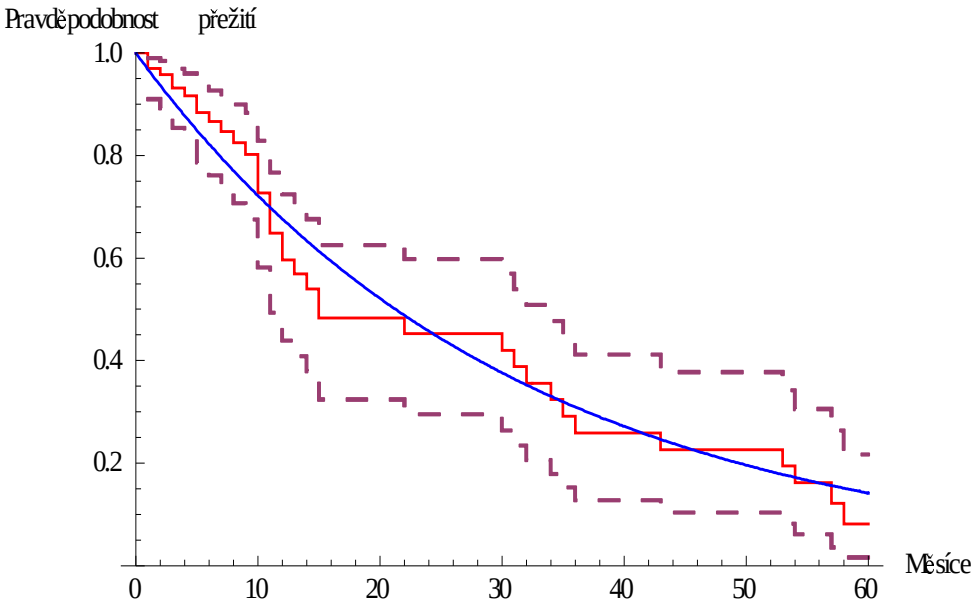
`aa = EstimatedDistribution[d,ExponentialDistribution[a]]`.

Pak by už jen stačilo použít funkci `SurvivalFunction` *a vykreslit graf.*

`Plot[SurvivalFunction[aa, t], {t, 0, 60}, PlotRange -> {0, 1}]`.

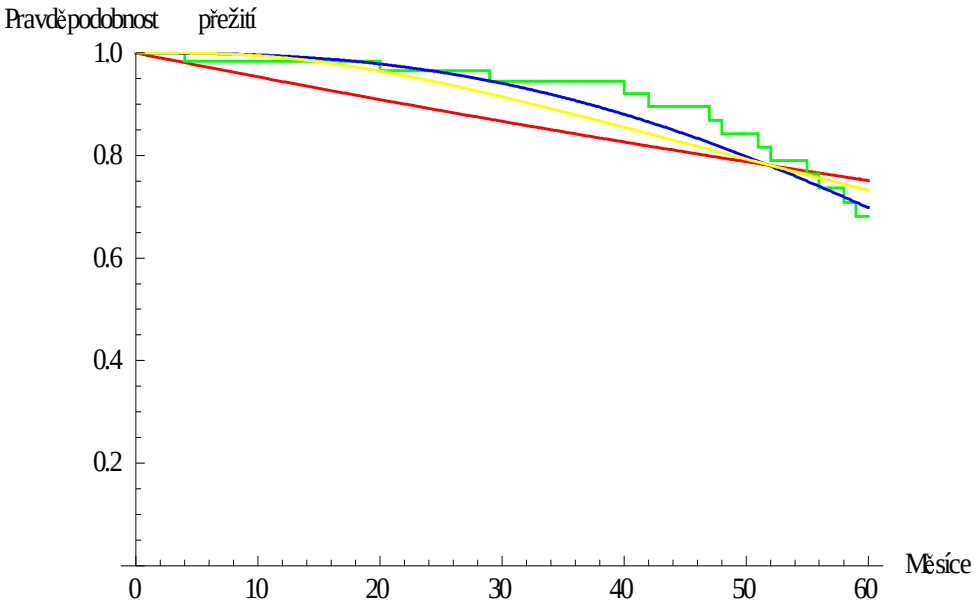
Pro ostatní modely by se postupovalo analogicky.

Teď můžeme pro zajímavost ukázat graf, kde porovnáme funkce přežití sestrojené neparametrickým modelem pomocí Kaplan-Meierova odhadu i s 95 procentním horním a dolním odhadnutým intervalem spolehlivosti a parametrickým modelem pomocí odhadnutého exponenciálního modelu.



Obrázek 4.4: Porovnání funkcí přežití pro pacienty s AIDS.

Dále je porovnání všech tří parametrických a Kaplan-Meireova neparametrického modelu pro druhý příklad s výrobky od firmy SC Johnson.



Obrázek 4.5: Porovnání funkcí přežití na trhu pro výrobky od firmy SC Johnson.

Všechny naprogramované procedury, stejně jako použitá data a některé další analýzy, které zde nebyly ukázány, jsou k nalezení v příloze na CD.

Závěr

V této práci jsme blíže ukázali omezené a cenzorované vysvětlované proměnné. Některé metody jak s nimi zacházet a jak postupovat v praxi.

Nejprve jsme popsali teorii a naší práci si částečně rozdělili, podle veličin, které se v ní zkoumali na diskrétní a spojité. Po popsání různých metod a možných modelů k použití, jsme již ukazovali, jak se postupuje v praxi.

Hodně jsme se zabývali analýzou přežití, která je velmi zajímavá a určitě přínosná. Jak jsme už zmínili, zabývá se převážně medicínskou tématikou, ale může být velmi dobře použitá i pro ekonomické analýzy, jak se nám povedlo demonstrovat v našem příkladu s našimi daty.

Také jsme porovnávali softwary. Nelze jednoznačně preferovat jeden software před druhým. V Mathematice bylo provedení analýz přeci jen elegantnější a nejjednodušší, zcela určitě i kvůli velmi dobré nápovědě, kterou Mathematica obsahuje.

V Rku se také daly provést požadované analýzy. Načtení dat a následný přístup k nim se jevil jako nejlepší.

Pro jednodušší a neparametrické modely lze velmi dobře použít i Excel, který může být mnohými preferován pro svojí jednoduchost.

Literatura

[1] J. Anděl. Základy matematické statistiky. *Matfyzpress, Praha , 2005.*

[2] T. Cipra. Finanční ekonometrie. *Ekopress, 2008.*

[3] David Collet. Modelling binary data. *Chapman and Hall/CRC, 2003.*

[4] D.R.COX and E.J.Snell. Applied Statistics - Principles and Examples. *Chapman and Hall/CRC, 2000.*

[5] Julian J.Faraway. Linear Models with R. *Chapman and Hall/CRC, 2005.*

[6] E. L. Kaplan and P. Meier. Non parametric estimation from incomplete observations. Journal of the American statistical association, *53(June):457–481, 1958.*

[7] Doug Martin Ricardo Maronna and Victor Yohai. Robust Statistics: Theory and Methods. *John Wiley and Sons, 2006.*

[8] David W.Hosmer and Stanley Lemeshow. Applied survival analysis. *Wiley Interscience, 1999.*

[9] M. Zhou. Use software R to do survival analysis and simulation. A tutorial. *Kentucky, Free download, http://www.stat.nus.edu.sg/ stachenz/Rsurv.pdf, 2008.*

Seznam obrázků