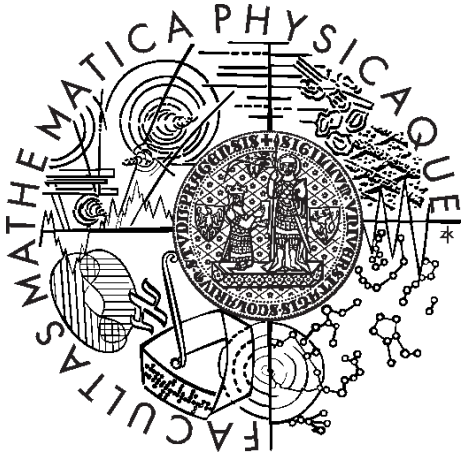
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Samuel Říha

Maximalizace Giniho koeficientu v binární logistické regresi

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: Mgr. Tomáš Hanzák
Studijní program: Matematika
Studijní obor: obecná matematika

Praha 2012

Rád bych vyjádřil poděkování Mgr. Tomášovi Hanzákovi, který mě trpělivě vedl při psaní této bakalářské práce, poskytl mi cenné rady a připomínky a také sadu reálných dat, bez nichž by má práce nemohla vzniknout. V neposlední řadě bych chtěl ocenit jeho velmi přátelský přístup.

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 ..... dne ..... Podpis autora

Název práce: Maximalizace Giniho koeficientu v binární logistické regresi

Autor: Samuel Říha

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: Mgr. Tomáš Hanzák, KPMS MFF UK

Abstrakt: V bakalářké práci je popsán model binární logistické regrese. Pomocí pojmu ztrátové funkce jsou odvozeny metody odhadu parametrů modelu. Je definována „bohatá“ množina „hezkých“ ztrátových funkcí - beta rodina Fisher-konzistentních ztrátových funkcí. V druhé části práce jsou definované základní ukazatele těsnosti modelu - Giniho koeficient, C-statistika, Kolmogorov-Smirnov statistika a koeficient determinace R^2. Dále je rozebrána možnost odhadovat parametry modelu maximalizací Giniho koeficientu. K tomuto účelu je navrženo několik algoritmů, které jsou porovnány s již existujícími metodami na jedné sadě simulovaných a třech sadách reálných dat.

Klíčová slova: Binární logistická regrese, Giniho koeficient, maximalizace Giniho koeficientu, ztrátová funkce, metoda maximální věrohodnosti.

Title: Gini coefficient maximization in binary logistic regression

Author: Samuel Říha

Department: Department of Probability and Mathematical Statistics

Supervisor: Mgr. Tomáš Hanzák, KPMS MFF UK

Abstract: This Bachelor thesis describes a binary logistic regression model. By means of the term loss function a parameter estimation for the model is derived. A „rich“ set of „proper“ loss functions – beta family of Fisher-consistent loss functions – is defined. In the second part of the thesis, four basic goodness-of-fit criteria - Gini coefficient, C-statistics, Kolmogorov-Smirnov statistics and coefficient of determination R^2 are defined. Further on, a possibility of parameter estimation by maximizing the Gini coefficient is analysed. Several algorithms are designed for this purpose. They are compared with so far existing methods in one simulated data set and three real ones.

Keywords: Binary logistic regression, Gini coefficient, Gini coefficient maximization, loss function, maximum likelihood.

Obsah

Úvod **1**

1 Binární logistická regrese **2**
1.1 Popis modelu 2
1.2 Odhad parametrů modelu 2
1.2.1 Metoda maximální věrohodnosti 2
1.2.2 Metoda nejmenších čtverců 3

2 Ztrátové funkce **4**
2.1 Definice ztrátové funkce 4
2.2 Fisher-konzistentní ztrátové funkce 5
2.3 Příklady Fisher-konzistentních ztrátových funkcí 6
2.4 Mocninné ztrátové funkce 11
2.5 Vlastnosti Fisher-konzistentních ztrátových funkcí 11
2.6 Beta rodina Fisher-konzistentních ztrátových funkcí 12
2.7 Metoda převážených nejmenších čtverců (IRLS) 14

3 Ukazatele těsnosti modelu **16**
3.1 Giniho koeficient 16
3.2 Kolmogorov-Smirnov statistika 20
3.3 Koeficient determinace R^2 21

4 Maximalizace Giniho koeficientu **22**
4.1 Maximalizace Giniho koeficientu pro normálně rozdělená skóre . . 22
4.2 Maximalizace Giniho koeficientu pro obecně rozdělená skóre . . . 27
4.3 Maximalizace Giniho koeficientu na beta rodině Fisher-konzistentních ztrátových funkcí 28

5 Aplikace na data **29**
5.1 Dvojrozměrný případ 29
5.2 Reálná data 33

Závěr **35**

Seznam použité literatury **36**

Seznam tabulek **37**

Úvod

Banky, pojišťovny a další finanční instituce mají u sebe data o statisících zákazníků. Aby tyto společnosti mohly dobře fungovat, potřebují umět vyhodnotit na základě informací obdržených od potenciálního zákazníka, jestli daný zákazník bude schopen splatit půjčku, či jaká je pravděpodobnost, že u klienta dojde během jednoho roku k pojistné události. Snaží se zjistit vliv dat (regresorů) na binární vysvětlovanou proměnnou (splatí/nesplatí úvěr). Touto problematikou se zabývá model binární logistické regrese, jenž je definován na začátku 1. kapitoly této práce.

K určení parametrů modelu, tj. k určení přesného vlivu jednotlivých regresorů na vysvětlovanou proměnnou, slouží různé metody odhadu parametrů. V naprosté většině případů se k tomu účelu používá metoda maximální věrohodnosti. V této práci si tuto metodu odvodíme a podíváme se na ni podrobněji. Zároveň si ukážeme, že existuje spousta jiných dobrých odhadových metod.

Odhadneme-li parametry modelu, chtěli bychom vědět, jak dobře jsme schopni předpovědět vysvětlovanou proměnnou. Mluvíme o ukazatelích těsnosti modelu a určujeme tzv. diverzifikační sílu modelu. Nejčastěji používáme jako ukazatel těsnosti modelu Giniho koeficient.

V praxi chceme dobře odhadnout parametry modelu, abychom spolehlivě předpověděli vysvětlovanou proměnnou, a Giniho koeficient pouze ukazuje, jak dobrá naše předpověď byla. Podívejme se ale na problém z role zaměstnance, který parametry odhaduje. Jeho ohodnocení nebude záviset na tom, jak dobře parametry odhadl, neboť to nebude schopen nadřízený posoudit. Často si nadřízený pouze spočítá Giniho koeficient a podle jeho velikosti usoudí, jak dobře zaměstnanec svou práci provedl. V některých případech je proto zaměstnancův cíl dosáhnout co největšího Giniho koeficientu, bez ohledu na přesnost odhadu parametrů modelu.

V lineární regresi se k určení diverzifikační síly modelu používá koeficient determinace R^2. Pokud odhadujeme parametry modelu metodou nejmenších čtverců (tuto odhadovou techniku lze též použít pro model binární logistické regrese), tak tím explicitně maximalizujeme koeficient determinace R^2. V binární logistické regresi žádná taková metoda, která by explicitně maximalizovala Giniho koeficient, neexistuje. Je domněnka, že takovou metodu ani není možné navrhnout. Přesto se v práci tímto tématem budeme zabývat a navrhneme několik algoritmů, které by pro data určitého charakteru měly vést k odhadu parametrů dávajícího téměř maximální Giniho koeficient.

V závěru práce tyto metody porovnáme s klasickými metodami odhadu parametrů.

1. Binární logistická regrese

1.1 Popis modelu

Nechť $Y_1, Y_2, \ldots, Y_n$ jsou náhodné veličiny s alternativním rozdělením, tedy rozdělením, jež nabývá hodnot 0 a 1. $\forall i \in \{1, 2, \ldots, n\}$ je Y_i závislá na vektoru čísel (vysvětlujících proměnných) $\boldsymbol{x_i} = (1, x_{i1}, x_{i2}, \ldots, x_{im})$. Model binární logistické regrese předpokládá

$$q(\boldsymbol{x_i}) = \frac{\exp\{\boldsymbol{\beta' x_i}\}}{1 + \exp\{\boldsymbol{\beta' x_i}\}},$$

kde $q(\boldsymbol{x_i}) = \boldsymbol{P}(Y_i = 1)$ a $\boldsymbol{\beta'} = (\beta_0, \beta_1, \ldots, \beta_m)$ je vektor parametrů modelu. Název logistická regrese je odvozen od vyjádření $q(\boldsymbol{x_i})$ pomocí funkce $\frac{e^z}{1+e^z}$, jíž říkáme logistická funkce.

Pozn: Jelikož v praxi neznáme skutečný parametr $\boldsymbol{\beta}$, ale známe jen jeho odhad, tak $q(\boldsymbol{x_i})$ není přímo $\boldsymbol{P}(Y_i = 1)$, ale pouze odhad $\eta(\boldsymbol{x_i}) = \boldsymbol{P}(Y_i = 1)$. V tomto duchu bude psaný zbytek této práce.

Příklad: Mějme hypoteční banku, která má archiv n bývalých klientů. $\forall i \in \{1, 2, \ldots, n\}$ náhodná veličina Y_i nabývá hodnoty 1, pokud i-tý klient splatil hypotéku, a hodnoty 0, pokud ji nesplatil. Vysvětlující proměnné můžou být: věk, pohlaví, plat, vzdělání. Vektor $\boldsymbol{x_i}$ = (1, věk i-tého klienta, pohlaví i-tého klienta, plat i-tého klienta, vzdělání i-tého klienta). U vysvětlujících proměnných je třeba určit jednotky tak, abychom i pohlaví klienta mohli vyjádřit číselně. Například 1 = muž, 2 = žena. Model binární logistické regrese slouží k určení pravděpodobnosti splacení hypotéky (určení $q(\boldsymbol{x_i})$) na základě znalostí věku, pohlaví, platu a vzdělání klienta (na základě $\boldsymbol{x_i}$).

1.2 Odhad parametrů modelu

Budeme se snažit zvolit vektor parametrů $\boldsymbol{\beta}$ tak, aby náš model co nejpřesněji určil $q(\boldsymbol{x_i})$, pro daný vektor vysvětlujících proměnných $\boldsymbol{x_i}$. K nalezení vhodného parametru $\boldsymbol{\beta}$ se používají například následující metody:

1.2.1 Metoda maximální věrohodnosti

Tato metoda je jednoznačně nejpoužívanější metodou odhadu parametru $\boldsymbol{\beta}$ v binární logistické regresi. Je založena na hledání takového parametru $\boldsymbol{\beta}$, při kterém je největší pravděpodobnost, že náhodným generováním dostaneme právě ta data, která máme.

Nechť $f^i_{\boldsymbol{\beta}}$ je zobecněná hustota náhodné veličiny Y_i pro $i = 1, 2, \ldots, n$ pro dané $\boldsymbol{\beta}$. Platí

$$f^i_{\boldsymbol{\beta}}(z) = \begin{cases} 1 - q(\boldsymbol{x_i}) & z = 0 \\ q(\boldsymbol{x_i}) & z = 1 \\ 0 & z \neq 0, z \neq 1\,. \end{cases}$$

Metoda maximální věrohodnosti je založena na hledání maxima tzv věrohodnostní funkce $l(\boldsymbol{\beta}), \boldsymbol{\beta} \in \boldsymbol{B}$, kde $\boldsymbol{B}$ je množina všech možných $\boldsymbol{\beta}$. Věrohodnostní funkce je za předpokladu nezávislosti $Y_i, i = 1, 2, \ldots, n$ definována takto

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} f_{\boldsymbol{\beta}}^{i}\,.$$

Označme $y_i \in \{0, 1\}$ již známé naměřené hodnoty Y_i, pro $i = 1, 2, \ldots, n$. Nyní platí

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} f_{\boldsymbol{\beta}}^{i} = \prod_{i=1}^{n} (1 - q(\boldsymbol{x_i}))^{1-y_i} q(\boldsymbol{x_i})^{y_i}\,.$$

Jelikož $0 < l(\boldsymbol{\beta}) < \infty$ a logaritmus je rostoucí na $(0, \infty)$, tak místo maxima $l(\boldsymbol{\beta})$ můžeme hledat maximum logaritmické věrohodnostní funkce

$$L(\boldsymbol{\beta}) = \ln(l(\boldsymbol{\beta})) = \ln\left(\prod_{i=1}^{n} (1 - q(\boldsymbol{x_i}))^{1-y_i} q(\boldsymbol{x_i})^{y_i}\right)$$

$$= \sum_{i=1}^{n} [(1 - y_i)\ln(1 - q(\boldsymbol{x_i})) + y_i \ln(q(\boldsymbol{x_i}))]\,.$$

1.2.2 Metoda nejmenších čtverců

Odhadová technika velmi často používaná zejména v lineární regresi. V binární logistické regresi je používána spíše zřídka.

Metoda hledá takové $\boldsymbol{\beta} \in \boldsymbol{B}$, pro které je výraz

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - q(\boldsymbol{x_i}))^2$$

nejmenší.

2. Ztrátové funkce

V první kapitole jsme se seznámili s metodami odhadu parametru $\boldsymbol{\beta}$ modelu binární logistické regrese. V této kapitole si rozšíříme množinu odhadových technik. Každá tato technika bude reprezentovaná tzv. ztrátovou funkcí. Začneme proto její definicí.

2.1 Definice ztrátové funkce

Ztrátová funkce v logistické regresi je reálná funkce závislá na proměnných y_i a $q(\boldsymbol{x_i})$. Značíme ji $\boldsymbol{L}(y_i|q(\boldsymbol{x_i}))$. Budeme často používat zkrácený zápis $\boldsymbol{L}(y|q)$. Jelikož y nabývá pouze hodnot 0 a 1, funkce $\boldsymbol{L}(y|q)$ se skládá pouze z dvou funkcí $\boldsymbol{L}(0|q)$ a $\boldsymbol{L}(1|q)$, které nazýváme částečné ztrátové funkce. Ztrátová funkce slouží jako penalizace pro nepřesné výsledky y_i v závislosti na $q(\boldsymbol{x_i})$. Velké hodnoty q by měly vést k $y = 1$, proto předpokládáme, že $\boldsymbol{L}(0|q)$ bude neklesající (velká penalizace pro $y = 0$ a q blízko 1) a $\boldsymbol{L}(1|q)$ ze stejného důvodu očekáváme nerostoucí. Značíme

$$L_0(q) = \boldsymbol{L}(0|q), \qquad L_1(1-q) = \boldsymbol{L}(1|q)\,.$$

Obě funkce $L_0(q)$, $L_1(1-q)$ jsou neklesající. Dále vzhledem k definici $L_0(q)$ a $L_1(1-q)$ platí

$$\boldsymbol{L}(y|q) = (1-y)L_0(q) + yL_1(1-q)\,.$$

Za předpokladu, že $\eta(\boldsymbol{x_i}) = P(Y_i = 1)$, si spočítáme očekávanou ztrátu

$$\boldsymbol{R}(\eta|q) = \boldsymbol{E}_Y \boldsymbol{L}(Y|q) = (1-\eta)L_0(q) + \eta L_1(1-q)\,.$$

Pozn: Víme, že $q \in (0,1)$ a $y \in \{0,1\}$. To nám umožňuje zapsat $\boldsymbol{L}(y|q)$ jako reálnou funkci jedné proměnné $r = y - q$. $\boldsymbol{L}(r)$ je nerostoucí na $[-1,0]$ a neklesající na $[0,1]$.

Definujme $\mathcal{L}(\boldsymbol{\beta})$ jako aritmetický průměr ztrát jednotlivých pozorování.

$$\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{L}(y_i|q(\boldsymbol{x_i}))$$

Minimalizace $\mathcal{L}(\boldsymbol{\beta})$ vede k některé metodě odhadu parametru $\boldsymbol{\beta}$. V předchozí kapitole jsme se seznámili se dvěma odhadovými technikami, metodou maximální věrohodnosti a metodou nejmenších čtverců. Metoda maximální věrohodnosti je založena na minimalizaci výrazu

$$-\sum_{i=1}^{n}[(1-y_i)\ln(1-q(\boldsymbol{x_i})) + y_i \ln(q(\boldsymbol{x_i}))]\,.$$

Ztrátová funkce metody maximální věrohodnosti má proto tvar

$$\boldsymbol{L}(y|q) = -(1-y)\ln(1-q) - y\ln(q))\,.$$

Obdobně metoda nejmenších čtverců je založena na minimalizaci výrazu

$$\sum_{i=1}^{n}(y_i - q(\boldsymbol{x_i}))^2,$$

což odpovídá ztrátové funkci

$$\boldsymbol{L}(y|q) = (y-q)^2 .$$

2.2 Fisher-konzistentní ztrátové funkce

Definice: Nechť $\{x_i, i = 1, 2, \ldots, n\}$ jsou data generovaná náhodnou veličinou X^θ s distribuční funkcí F^θ. Odhad $\hat{\theta}$ parametru θ je Fisher-konzistentní, pokud existuje na množině jednorozměrných distribučních funkcí funkcionál T splňující

$$\hat{\theta} = T(\hat{F_n}), \qquad \theta = T(F^\theta) ,$$

kde $\hat{F_n}$ je empirická distribuční funkce definovaná

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{[x_i \le t]} .$$

Pozn: Požadavek Fisher-konzistence nám říká, že pro dostatečné množství pozorování lze získat chybu odhadu libovolně malou. Má tedy podobné vlastnosti jako požadavek konzistence.

Věta 2.1: Nechť pro $\forall \eta \in (0,1)$ platí

$$\eta = \text{argmin}_{\pi\in(0,1)} \boldsymbol{E}_Y \boldsymbol{L}(Y|\pi) ,$$

kde Y je alternativně rozdělená náhodná veličina s parametrem η. Pokud je $q(\boldsymbol{x})$ dostatečně bohatá množina funkcí (tj. pro $\forall \boldsymbol{q} \in [0,1]^n$ lze najít parametr $\boldsymbol{\beta}$ splňující $\boldsymbol{q} = q(\boldsymbol{x})$), potom minimalizace $\mathcal{L}(\boldsymbol{\beta})$ vede na vektor $q(\boldsymbol{x})$ odhadů vektoru parametrů $\eta(\boldsymbol{x})$, jež je Fisher-konzistentní (po složkách).
Důkaz: Definujme na množině všech distribučních funkcí odhad q parametru η

$$q(F) = \text{argmin}_{\pi\in(0,1)}(F(0.5) - F(-0.5))L_0(\pi) + (F(1.5) - F(0.5))L_1(1-\pi) .$$

Pro F distribuční funkci náhodné veličiny Y s alternativním rozdělením o parametru η platí

$$\begin{aligned} q(F) &= \text{argmin}_{\pi\in(0,1)}(F(0.5) - F(-0.5))L_0(\pi) + (F(1.5) - F(0.5))L_1(1-\pi) \\ &= \text{argmin}_{\pi\in(0,1)}(1-\eta)L_0(\pi) + \eta L_1(1-\pi) \\ &= \text{argmin}_{\pi\in(0,1)} \boldsymbol{E}_Y \boldsymbol{L}(Y|\pi) . \end{aligned}$$

Dle předpokladu věty platí

$$\eta = \text{argmin}_{\pi\in(0,1)} \boldsymbol{E}_Y \boldsymbol{L}(Y|\pi) .$$

Odtud dostáváme rovnost

$$q(F) = \eta .$$

Odhad $q(\hat{F_i})$ je tedy Fisher-konzistentní, kde $\hat{F_i}$ je empirická distribuční funkce z náhodného výběru Y_i. Tato distribuční funkce je definovaná následovně

$$\hat{F_i}(t) = \begin{cases} 0 & t < y_i \\ 1 & t \ge y_i . \end{cases}$$

Odhad q parametru η za základě našich dat získáme jako

$$\begin{aligned} q(\hat{F}_i) &= \operatorname{argmin}_{\pi\in(0,1)}(\hat{F}_i(0.5) - \hat{F}_i(-0.5))L_0(\pi) + (\hat{F}_i(1.5) - \hat{F}_i(0.5))L_1(1-\pi) \\ &= \operatorname{argmin}_{\pi\in(0,1)}(1-y_i)L_0(\pi) + y_i L_1(1-\pi) \\ &= \operatorname{argmin}_{\pi\in(0,1)}\boldsymbol{L}(y_i|\pi)\,. \end{aligned}$$

Odhad $q(\boldsymbol{x})$ vektoru $\eta(\boldsymbol{x})$ pomocí minimalizace $\mathcal{L}$ je

$$\begin{aligned} q(\boldsymbol{x}) &= \operatorname{argmin}_{\boldsymbol{\pi}\in(0,1)^n}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{L}(y_i|\pi_i) \\ &= (\operatorname{argmin}_{\pi_1\in(0,1)}\boldsymbol{L}(y_1|\pi_1), \ldots, \operatorname{argmin}_{\pi_n\in(0,1)}\boldsymbol{L}(y_n|\pi_n)) \\ &= (q(\hat{F}_1), \ldots, q(\hat{F}_n))\,. \end{aligned}$$

Odtud již vidíme, že minimalizace $\mathcal{L}$ skutečně vede na Fisher-konzistentní odhad $q(\boldsymbol{x})$ vektoru parametrů $\eta(\boldsymbol{x})$. □

Pozn: Věta (2.1) *nám umožňuje definovat třídu „hezkých“ ztrátových funkcí.*

Definice: Fisher-konzistentní ztrátovou funkcí rozumíme ztrátovou funkci splňující

$$\operatorname{argmin}_{\pi\in(0,1)}\boldsymbol{R}(\eta|\pi) = \eta\,,$$

pro $\forall\eta \in (0,1)$. Pokud je minimum určeno jednoznačně, pak mluvíme o striktně Fisher-konzistentní ztrátové funkci, jinak o nestriktně Fisher-konzistentní ztrátové funkci.

Věta 2.2: Každá „hladká“ Fisher-konzistentní ztrátová funkce musí splňovat podmínku

$$(1-\eta)L_0'(\eta) = \eta L_1'(1-\eta), \qquad \forall\eta \in (0,1)\,.$$

Důkaz: Víme, že Fisher-konzistentní funkce nabývá svého globálního minima v bodě $q = \eta$. Je-li tato funkce „hladká“, musí nutně platit

$$\left.\frac{\partial}{\partial q}\right|_{q=\eta}\boldsymbol{R}(\eta|q) = 0\,.$$

Úpravami získáme

$$0 = \left.\frac{\partial}{\partial q}\right|_{q=\eta}\boldsymbol{R}(\eta|q) = \left.\frac{\partial}{\partial q}\right|_{q=\eta}[(1-\eta)L_0(q)+\eta L_1(1-q)] = (1-\eta)L_0'(\eta)-\eta L_1'(1-\eta)\,,$$

a tedy platí

$$(1-\eta)L_0'(\eta) = \eta L_1'(1-\eta)\,.$$

□

2.3 Příklady Fisher-konzistentních ztrátových funkcí

Uvedeme si čtyři používané Fisher-konzistentní ztrátové funkce. Můžeme si snadno ověřit, že jejich částečné ztrátové funkce L_0 a L_1, jsou neklesající, a tedy definované funkce jsou skutečně ztrátové funkce. Důkaz o tom, že jsou ztrátové funkce Fisher-konzistentní, odložíme na později.

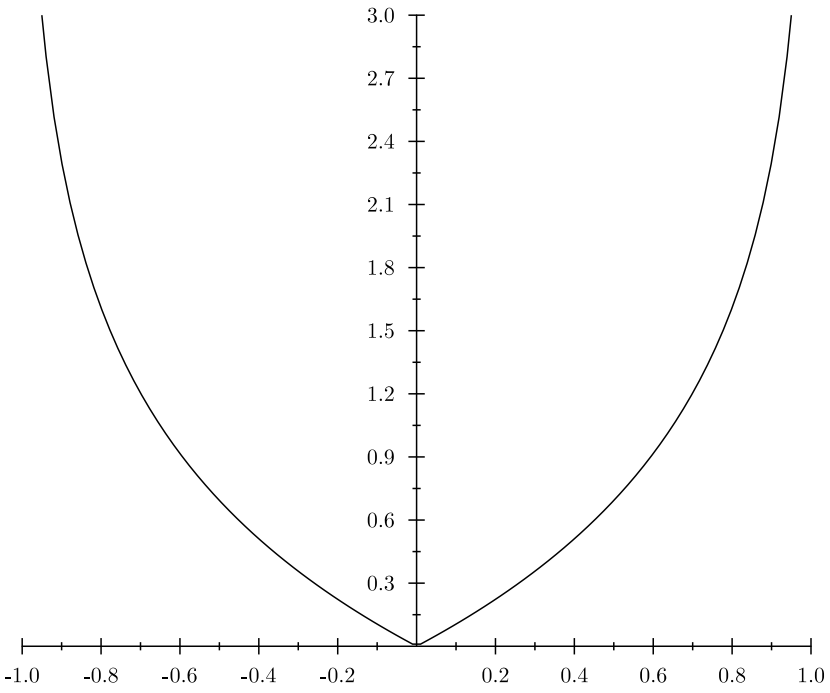
Věrohodnostní ztrátová funkce

Je to ztrátová funkce odvozená od metody maximální věrohodnosti. Jedná se o nejpoužívanější ztrátovou funkci v binární logistické regresi.

$$\boldsymbol{L}(y|q) = -(1-y)\ln(1-q) - y\ln(q)$$

Odvodíme si aritmetický průměr jednotlivých pozorování $\mathcal{L}$, částečné ztrátové funkce L_0, L_1, očekávanou ztrátu R a vyjádříme si $\boldsymbol{L}$ v závislosti na reziduu $r = y - q$.

$$\begin{aligned}
\boldsymbol{\mathcal{L}(\beta)} &= \frac{1}{n}\sum_{i=1}^{n}[-(1-y_i)\ln(1-q(\boldsymbol{x_i}) - y_i\ln(q(\boldsymbol{x_i}))] \\
L_0(q) &= -\ln(1-q) \\
L_1(1-q) &= -\ln(q) \\
\boldsymbol{L}(r) &= -\ln(1-|r|) \\
\boldsymbol{R}(\eta|q) &= -(1-\eta)\ln(1-q) - \eta\ln(q)
\end{aligned}$$



Obrázek 2.1: Graf věrohodnostní ztrátové funkce $\boldsymbol{L}(r)$.

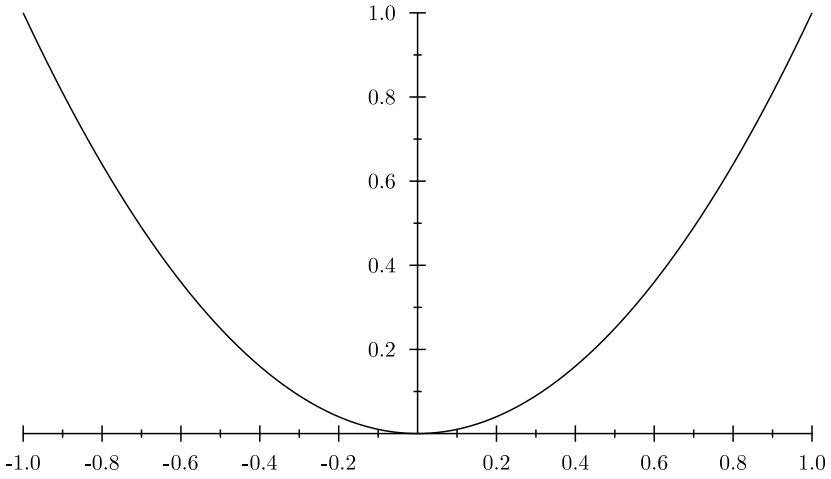
Čtvercová ztrátová funkce

Další často používaná Fisher-konzistentní ztrátová funkce odvozená z metody nejmenších čtverců.

$$\boldsymbol{L}(y|q) = (y-q)^2$$

Pomocné funkce pro tuto ztrátovou funkci jsou

$$\begin{aligned}
\boldsymbol{\mathcal{L}}(\boldsymbol{\beta}) &= \frac{1}{n}\sum_{i=1}^{n}(y_i - q(\boldsymbol{x_i}))^2 \\
L_0(q) &= q^2 \\
L_1(1-q) &= (1-q)^2 \\
\boldsymbol{L}(r) &= r^2 \\
\boldsymbol{R}(\eta|q) &= (1-\eta)q^2 + \eta(1-q)^2\,.
\end{aligned}$$



Obrázek 2.2: Graf čtvercové ztrátové funkce $\boldsymbol{L}(r)$.

Exponenciální ztrátová funkce

Jedná se o ztrátovou funkci odvozenou od metody hledání parametru $\boldsymbol{\beta}$ minimalizováním výrazu

$$\frac{1}{n}\sum_{i=1}^{n} e^{-(2y_i-1)F(\boldsymbol{x_i})} = \frac{1}{n}\sum_{i=1}^{n}[y_i e^{-F(\boldsymbol{x_i})} + (1-y_i)e^{F(\boldsymbol{x_i})}]\,, \tag{2.1}$$

kde

$$F(\boldsymbol{x_i}) = \frac{1}{2}\sum_{j=0}^{m}\beta_j x_{ij}\,.$$

V modelu binární logistické regrese má $q(\boldsymbol{x_i})$ tvar

$$q(\boldsymbol{x_i}) = \frac{1}{1+e^{-2F(\boldsymbol{x_i})}}\,.$$

Odtud vyjádříme $F(\boldsymbol{x_i})$ jako

$$F(\boldsymbol{x_i}) = \frac{1}{2}\ln\left(\frac{q(\boldsymbol{x_i})}{1-q(\boldsymbol{x_i})}\right)\,.$$

Toto vyjádření $F(\boldsymbol{x_i})$ dosadíme do (2.1). Minimalizujeme tedy výraz

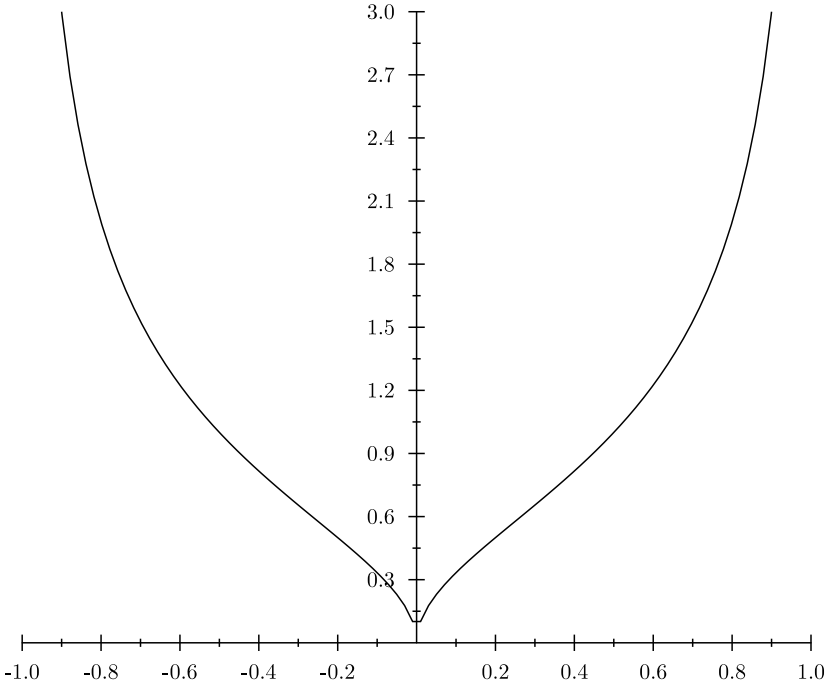
$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^{n}\left[y_i e^{-\frac{1}{2}\ln\left(\frac{q(\boldsymbol{x_i})}{1-q(\boldsymbol{x_i})}\right)} + (1-y_i)e^{\frac{1}{2}\ln\left(\frac{q(\boldsymbol{x_i})}{1-q(\boldsymbol{x_i})}\right)}\right] \\ = \quad & \frac{1}{n}\sum_{i=1}^{n}\left[y_i\left(\frac{1-q(\boldsymbol{x_i})}{q(\boldsymbol{x_i})}\right)^{1/2} + (1-y_i)\left(\frac{q(\boldsymbol{x_i})}{1-q(\boldsymbol{x_i})}\right)^{1/2}\right]. \end{aligned}$$

Tomu odpovídající ztrátová funkce nazývaná exponenciální ztrátová funkce je

$$\boldsymbol{L}(y|q) = y\left(\frac{1-q}{q}\right)^{1/2} + (1-y)\left(\frac{q}{1-q}\right)^{1/2}.$$

Pomocné funkce pro exponenciální ztrátovou funkci jsou

$$\begin{aligned} \mathcal{L}(\boldsymbol{\beta}) &= \frac{1}{n}\sum_{i=1}^{n}\left[y_i\left(\frac{1-q(\boldsymbol{x_i})}{q(\boldsymbol{x_i})}\right)^{1/2} + (1-y_i)\left(\frac{q(\boldsymbol{x_i})}{1-q(\boldsymbol{x_i})}\right)^{1/2}\right] \\ L_0(q) &= \left(\frac{q}{1-q}\right)^{1/2} \\ L_1(1-q) &= \left(\frac{1-q}{q}\right)^{1/2} \\ \boldsymbol{L}(r) &= \left(\frac{|r|}{1-|r|}\right)^{\frac{1}{2}} \\ \boldsymbol{R}(\eta|q) &= \eta\left(\frac{1-q}{q}\right)^{1/2} + (1-\eta)\left(\frac{q}{1-q}\right)^{1/2}. \end{aligned}$$



Obrázek 2.3: Graf exponenciální ztrátové funkce $\boldsymbol{L}(r)$.

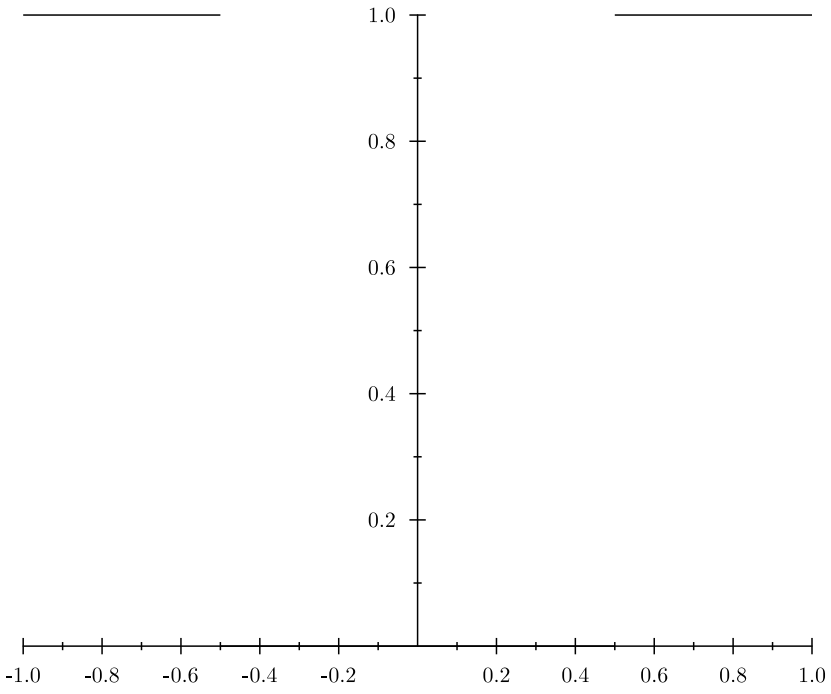
Binární ztrátová funkce

Tato ztrátová funkce, na rozdíl od předchozích tří, není striktně Fisher-konzistentní. Penalizuje špatnou klasifikaci $q(\boldsymbol{x_i})$. Nerozlišuje ale, jak moc je tato klasifikace špatná. Pro konstantu $c \in (0,1)$ definujeme binární ztrátovou funkci

$$\boldsymbol{L}(y|\pi) = y\mathbb{I}_{[q\leq c]} + (1-y)\mathbb{I}_{[q>c]}\,.$$

Zjistíme, pro jaké parametry c, se jedná o Fisher-konzistentní ztrátovou funkci. K tomu budeme nejdříve potřebovat spočítat pomocné funkce.

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\beta}) &= \frac{1}{n}\sum_{i=1}^{n}\left[y_i\mathbb{I}_{[q(\boldsymbol{x_i})\leq c]} + (1-y_i)\mathbb{I}_{[q(\boldsymbol{x_i})>c]}\right] \\
L_0(q) &= \mathbb{I}_{[q>c]} \\
L_1(1-q) &= \mathbb{I}_{[q\leq c]} \\
\boldsymbol{L}(r) &= \mathbb{I}_{[r\leq -c]} + \mathbb{I}_{[r>1-c]} \\
\boldsymbol{R}(\eta|q) &= (1-\eta)\mathbb{I}_{[q>c]} + \eta\mathbb{I}_{[q\leq c]}
\end{aligned}$$



Obrázek 2.4: Graf binární ztrátové funkce $\boldsymbol{L}(r)$ pro $c = 0.5$.

Potřebujeme ověřit z definice podmínku Fisher-konzistentních ztrátových funkcí

$$\text{argmin}_{\pi\in(0,1)}\boldsymbol{R}(\eta|\pi) = \eta, \qquad \forall\eta \in (0,1)\,.$$

Tato podmínka pro binární ztrátovou funkci má tvar

$$\text{argmin}_{\pi\in(0,1)}[(1-\eta)\mathbb{I}_{[\pi>c]} + \eta\mathbb{I}_{[\pi\leq c]}] = \eta, \qquad \forall\eta \in (0,1)\,.$$

Pro $\eta \leq 0.5$ nastává minimum pro všechna $\pi \in (0, c]$. Pro $\eta \geq 0.5$ nastává minimum pro všechna $\pi \in [c, 1)$. Platí tedy, že minimum vždy nastává pro $\pi = \eta$ pouze pro případ $c = 0.5$. Binární ztrátová funkce je Fisher-konzistentní právě tehdy, když $c = 0.5$.

2.4 Mocninné ztrátové funkce

Metoda nejmenších čtverců je používaná metoda odhadu parametru a jí příslušná čtvercová ztrátová funkce je Fisher-konzistentní. Mocninná ztrátová funkce je zobecněním funkce a je definovaná

$$\boldsymbol{L}(y|\pi) = |y - \pi|^k ,$$

kde $k \in \mathbb{R}^+$ je parametr. Pro $k = 2$ se jedná o čtvercovou ztrátovou funkci.

Věta 2.3: Mocninná ztrátová funkce $\boldsymbol{L}(y|\pi) = |y - \pi|^k$ je Fisher-konzistentní právě tehdy, když $k = 2$. Tedy jediná mocninná Fisher-konzistentní ztrátová funkce je čtvercová ztrátová funkce.
Důkaz: Již víme, že čtvercová ztrátová funkce je Fisher-konzistentní. Stačí nám dokázat, že pokud je mocninná ztrátová funkce Fisher-konzistentní, pak už nutně $k = 2$. Předpokládejme tedy, že mocninná ztrátová funkce s parametrem k je Fisher-konzistentní. Spočteme částečné ztrátové funkce L_0 a L_1.

$$\begin{aligned} L_0(q) &= q^k \\ L_1(1-q) &= (1-q)^k \end{aligned}$$

Dopočítáme derivace L_0' a L_1'.

$$\begin{aligned} L_0'(q) &= kq^{k-1} \\ L_1'(1-q) &= k(1-q)^{k-1} \end{aligned}$$

Podle věty 2.2 musí pro $\forall \eta \in (0,1)$ platit

$$(1-\eta)L_0'(\eta) = \eta L_1'(1-\eta)$$

Přepíšeme tuto podmínku pro mocninnou ztrátovou funkci a následně upravíme.

$$\begin{aligned} (1-\eta)k\eta^{k-1} &= \eta k(1-\eta)^{k-1} \\ \eta^{k-2} &= (1-\eta)^{k-2} \end{aligned}$$

Tato podmínka je splněna pouze pro $k = 2$, a proto jediná mocninná Fisher-konzistentní ztrátová funkce je čtvercová ztrátová funkce. □

2.5 Vlastnosti Fisher-konzistentních ztrátových funkcí

Uvažujme pouze „hladké“ ztrátové funkce.

Věta 2.4: $\boldsymbol{L}(y|\pi)$ je Fisher-konzistentní ztrátová funkce právě tehdy, když platí

$$L_0'(q) = \omega(q)q\,, \qquad L_1'(1-q) = \omega(q)(1-q)$$

pro nějakou váhovou funkci $\omega(q) \geq 0$ na $(0,1)$, splňující $\int_\varepsilon^{1-\varepsilon} \omega(t)\mathrm{d}t < \infty, \forall \varepsilon > 0$. Fisher-konzistentní ztrátová funkce je striktní právě tehdy, když $\omega(q) > 0$ skoro všude na $(0,1)$.

Důkaz:
"⇐": Rozepíšeme si, jak vypadá derivace očekávané ztráty $\boldsymbol{R}(\eta|q)$.

$$\begin{aligned}\frac{\partial}{\partial q}\boldsymbol{R}(\eta|q) &= \frac{\partial}{\partial q}[(1-\eta)L_0(q)+\eta L_1(1-q)] = (1-\eta)L_0'(q)-\eta L_1'(1-q)\\ &= (1-\eta)\omega(q)q-\eta\omega(q)(1-q) = (q-\eta)\omega(q)\end{aligned}$$

Derivace $\boldsymbol{R}'(\eta|q)$ je nekladná na $(0,\eta)$ a nezáporná na $(\eta,1)$. Proto je $\boldsymbol{R}(\eta|q)$ nerostoucí na $(0,\eta]$ a nekelsající na $[\eta,1)$, a tudíž očekávaná ztráta má globální minimum v bodě $q=\eta$. Pokud je navíc $\omega(q)>0$ skoro všude na $(0,1)$, pak $\boldsymbol{R}'(\eta|q)$ je nenulová na každém okolí η, a tedy $\boldsymbol{R}(\eta|q)$ není konstantní na žádném okolí η. Odtud plyne, že je globální minimum jediné. Naopak, pokud není $\omega(q)>0$ skoro všude na $(0,1)$, pak vzhledem ke spojitosti ω existuje netriviální interval, na kterém je $\omega(q)=0$. Předpokládáme $\boldsymbol{R}'(\eta|q)$ spojitou, proto η musí ležet v tomto intervalu, a tedy na celém tomto intervalu má $\boldsymbol{R}(\eta|q)$ globální minimum, a ztrátová funkce $\boldsymbol{L}(y|q)$ tudíž není striktní.
"⇒": Z věty 2.2 plyne $\frac{L_0'(\eta)}{\eta}=\frac{L_1'(1-\eta)}{1-\eta}$. Proto je následující definice korektní.

$$\omega(\eta)=\frac{L_0'(\eta)}{\eta}=\frac{L_1'(1-\eta)}{1-\eta}$$

Váhová funkce $\omega(\eta)$ tedy splňuje pro $\forall q\in(0,1)$ vztahy

$$L_0'(q)=\omega(q)q\,,\qquad L_1'(1-q)=\omega(q)(1-q)\,.$$

Jelikož je L_0 neklesající, tak $L_0'\geq 0$, a tedy i $\omega(q)\geq 0, \forall q\in(0,1)$. Zbývá ukázat $\int_\varepsilon^{1-\varepsilon}\omega(t)\mathrm{d}t<\infty, \forall\varepsilon>0$.

$$\int_\varepsilon^{1-\varepsilon}\omega(t)\mathrm{d}t=\int_\varepsilon^{1-\varepsilon}\frac{L_0'(t)}{t}\mathrm{d}t\leq\frac{1}{\varepsilon}\int_\varepsilon^{1-\varepsilon}L_0'(t)\mathrm{d}t=\frac{1}{\epsilon}\left[L_0(1-\varepsilon)+L_0(\varepsilon)\right]<\infty$$

□

2.6 Beta rodina Fisher-konzistentních ztrátových funkcí

Následující věta nám umožní definovat zajímavou podmnožinu Fisher-konzistentních ztrátových funkcí.

Věta 2.5: Ztrátová funkce definovaná vztahem

$$\boldsymbol{L}(y|q)=(1-y)\int_0^q t\omega(t)\mathrm{d}t+y\int_q^1(1-t)\omega(t)\mathrm{d}t$$

pro $\omega(q)=q^{\alpha-1}(1-q)^{\beta-1}$ s parametry $\alpha,\beta>-1$ je Fisher-konzistentní.
Důkaz: Nejdříve si uvědomme, že integrály

$$\int_0^q t\omega(t)\mathrm{d}t\qquad\text{a}\qquad\int_q^1(1-t)\omega(t)\mathrm{d}t$$

jsou konečné pro $\forall q \in (0,1)$ a $\alpha, \beta > -1$. Definice ztrátové funkce je tudíž korektní. Nyní ověřme podmínky z věty 2.4 pro váhovou funkci $\omega(q)$.

$$\begin{aligned} L_0'(q) &= \frac{\partial}{\partial q}\int_0^q t\omega(t)\mathrm{d}t = q\omega(q) \\ L_1'(1-q) &= \frac{\partial}{\partial(1-q)}\int_q^1 (1-t)\omega(t)\mathrm{d}t = \frac{\partial}{\partial(1-q)}\int_{1-q}^0 -s\omega(1-s)\mathrm{d}s \\ &= \frac{\partial}{\partial(1-q)}\int_0^{1-q} s\omega(1-s)\mathrm{d}s = (1-q)\omega(q) \end{aligned}$$

Váhová funkce tedy má patřičné vztahy s derivacemi částečných ztrátových funkcí L_0' a L_1'. Zřejmě platí $\omega(q) \geq 0$ na $(0,1)$. Pro $\forall \varepsilon > 0$ je $\omega(q)$ spojitá na intervalu $[\varepsilon, 1-\varepsilon]$. Odtud vyplývá

$$\int_\varepsilon^{1-\varepsilon} \omega(t)\mathrm{d}t < \infty\,.$$

Tím jsme ověřili všechny podmínky věty 2.4 a ztrátová funkce $\boldsymbol{L}(y|q)$ je tudíž Fisher-konzistentní. □

Množině Fisher-konzistentních ztrátových funkcí, odvozených podle váhové funkce $\omega(q) = q^{\alpha-1}(1-q)^{\beta-1}$ s parametry $\alpha, \beta > -1$, říkáme beta rodina Fisher-konzistentních ztrátových funkcí. Podle definice ve větě 2.4 si spočítáme $\omega(q)$ pro často používané ztrátové funkce a zjistíme, že patří do beta rodiny konzistentních ztrátových funkcí. Mimo jiné tím dokážeme, že jsou dané ztrátové funkce Fisher-konzistentní.

Exponenciální ztrátová funkce

$$\omega(q) = \frac{L_0'(q)}{q} \asymp \frac{1}{q^{3/2}(1-q)^{3/2}}$$

Odpovídá beta Fisher-konzistentní ztrátové funkci s parametry $\alpha = \beta = -\frac{1}{2}$.

Věrohodnostní ztrátová funkce

$$\omega(q) = \frac{L_0'(q)}{q} = \frac{1}{q(1-q)}$$

Odpovídá beta Fisher-konzistentní ztrátové funkci s parametry $\alpha = \beta = 0$.

Čtvercová ztrátová funkce

$$\omega(q) = \frac{L_0'(q)}{q} \asymp 1$$

Odpovídá beta Fisher-konzistentní ztrátové funkci s parametry $\alpha = \beta = 1$.

Binární ztrátová funkce

Lze spočítat, že binární ztrátová funkce pro $c = 0.5$ je limitním případem beta Fisher-konzistentní ztrátové funkce pro $\alpha, \beta \to \infty$.

2.7 Metoda převážených nejmenších čtverců (IRLS)

Popíšeme, jak se v praxi numericky odhadují parametry $\boldsymbol{\beta}$ modelu binární logistické regrese. Chceme minimalizovat výraz

$$\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{L}(y_i|q(\boldsymbol{x_i}))\,.$$

Pokud je funkce $\mathcal{L}$ konvexní, k tomuto účelu výborně poslouží Newtonova metoda tečen. Ta spočívá v iterativním hledání vektoru $\boldsymbol{\beta}$. Iterační krok je definován následovně

$$\boldsymbol{\beta}_{new} = \boldsymbol{\beta}_{old} - (\partial^2_{\boldsymbol{\beta}}\mathcal{L}(\boldsymbol{\beta}_{old}))^{-1}(\partial_{\boldsymbol{\beta}}\mathcal{L}(\boldsymbol{\beta}_{old}))\,.$$

Abychom tuto metodu mohli aplikovat, musíme spočítat gradient $\partial_{\boldsymbol{\beta}}\mathcal{L}(\boldsymbol{\beta})$.

$$\begin{aligned}\frac{\partial}{\partial\beta_j}\mathcal{L}(\boldsymbol{\beta}) &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\beta_j}\boldsymbol{L}(y_i|q(\boldsymbol{x_i})) \\ &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{L'}(y_i|q(\boldsymbol{x_i}))q(\boldsymbol{x_i})(1-q(\boldsymbol{x_i}))x_{ij}\end{aligned}$$

a Jacobiho matici $\partial^2_{\boldsymbol{\beta}}\mathcal{L}(\boldsymbol{\beta})$.

$$\begin{aligned}\frac{\partial^2}{\partial\beta_j\beta_k}\mathcal{L}(\boldsymbol{\beta}) &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\beta_j\beta_k}\boldsymbol{L}(y_i|q(\boldsymbol{x_i})) \\ &= \frac{1}{n}\sum_{i=1}^{n}[\boldsymbol{L''}(y_i|q(\boldsymbol{x_i}))q(\boldsymbol{x_i})^2(1-q(\boldsymbol{x_i}))^2 \\ &\qquad +\boldsymbol{L'}(y_i|q(\boldsymbol{x_i}))q(\boldsymbol{x_i})(1-q(\boldsymbol{x_i}))(1-2q(\boldsymbol{x_i}))]x_{ij}x_{ik}\end{aligned}$$

V předchozích výrazech $\boldsymbol{L'}$ značí derivaci $\boldsymbol{L}$ podle q. Dále definujme váhovou matici

$$\boldsymbol{W} = \mathrm{diag}(w_n)$$

a pomocný vektor

$$\boldsymbol{z} = (z_1, \ldots, z_n)$$

tak, aby platilo

$$\boldsymbol{X'WX} = n(\partial^2_{\boldsymbol{\beta}}\mathcal{L}(\boldsymbol{\beta})) \qquad \text{a} \qquad \boldsymbol{X'Wz} = n\left(-\partial_{\boldsymbol{\beta}}\mathcal{L}(\boldsymbol{\beta})\right)\,.$$

Vyhovuje

$$w_i = \boldsymbol{L''}(y_i|q(\boldsymbol{x_i}))q(\boldsymbol{x_i})^2(1-q(\boldsymbol{x_i}))^2 + \boldsymbol{L'}(y_i|q(\boldsymbol{x_i}))q(\boldsymbol{x_i})(1-q(\boldsymbol{x_i}))(1-2q(\boldsymbol{x_i}))$$

$$z_i = -\frac{\boldsymbol{L'}(y_i|q(\boldsymbol{x_i}))q(\boldsymbol{x_i})(1-q(\boldsymbol{x_i}))}{w_i}$$

Iterativní hledání parametru $\boldsymbol{\beta}$ tedy vypadá

$$\boldsymbol{\beta}_{new} = \boldsymbol{\beta}_{old} + (\boldsymbol{X'WX})^{-1}(\boldsymbol{X'Wz})\,,$$

což odpovídá iteračnímu kroku v iterační metodě převážených nejmenších čtverců (IRLS).

Dosazením

$$\begin{aligned} \boldsymbol{L}'(y_i|q(\boldsymbol{x_i})(\boldsymbol{\beta}'\boldsymbol{x_i})) &= -\omega(q(\boldsymbol{x_i}))(y_i - q(\boldsymbol{x_i})) \\ \boldsymbol{L}''(y_i|q(\boldsymbol{x_i})(\boldsymbol{\beta}'\boldsymbol{x_i})) &= -\omega'(q(\boldsymbol{x_i}))(y_i - q(\boldsymbol{x_i})) + \omega(q(\boldsymbol{x_i})) \end{aligned}$$

spočteme $\boldsymbol{W}$ a $\boldsymbol{z}$ pro Fisher-konzistentní ztrátové funkce. Budeme psát zkráceně q_i místo $q(\boldsymbol{x_i})$.

$$\begin{aligned} w_i &= \omega(q_i)(q_i(1-q_i))^2 - (y_i - q_i)[q_i(1-q_i)(1-2q_i)\omega(q_i) + \omega'(q_i)(q_i(1-q_i))^2] \\ z_i &= \frac{\omega(q_i)(y_i - q_i)q_i(1-q_i)}{w_i} \end{aligned}$$

Speciálně pro metodu maximální věrohodnosti, tj. $\omega(q) = q^{-1}(1-q)^{-1}$, platí

$$\begin{aligned} w_i &= q_i(1-q_i) \\ z_i &= \frac{y_i - q_i}{w_i}\,. \end{aligned}$$

Pro ostatní odhadové metody je možné porovnat derivaci jejich ztrátových funkcí s derivací věrohodnostní ztrátové funkce. Tímto podílem vynásobit váhovou matici $\boldsymbol{W}$ pro metodu maximální věrohodnosti, a získat tak váhovou matici $\boldsymbol{W}$ pro libovolnou ztrátovou funkci. Derivace věrohodnostní ztrátové funkce je

$$\frac{\partial}{\partial r}\boldsymbol{L} = -\frac{\partial}{\partial r}\ln(1-|r|) = \frac{\operatorname{sign}(r)}{1-|r|} = \frac{r}{q(1-q)}\,,$$

kde $r = y - q$. Pro obecnou ztrátovou funkci $\boldsymbol{L}$ vypadá váhová matice $\boldsymbol{W}$ a pomocný vektor $\boldsymbol{z}$ takto

$$\begin{aligned} w_i &= \frac{\boldsymbol{L}'(r_i)}{r_i} q_i^2(1-q_i)^2 \\ z_i &= \frac{r_i}{w_i}\,. \end{aligned}$$

Použitím této zjednodušené váhové matice $\boldsymbol{W}$ a pomocného vektoru $\boldsymbol{z}$ sice přicházíme o přesnost spočítaného odhadu $\boldsymbol{\beta}$, ale díky jednoduchosti algoritmus běží rychleji a jsme schopni jej implementovat i pro složité ztrátové funkce.

3. Ukazatele těsnosti modelu

Mějme data $(\boldsymbol{x_i}, y_i), i = 1, 2, \ldots, n$. V předchozí části jsme ze zabývali metodami hledání parametru $\boldsymbol{\beta}$ tak, aby model co nejlépe předpovídal výsledky y_i. Chtěli bychom mít nástroje, které nám umožní říct, jak dobře lze tyto výsledky y_i předpovědět. K tomu nám slouží ukazatele těsnosti modelu. V Evropě se v binární logistické regresi nejčastěji používá Giniho koeficient, v Severní Americe Kolmogorov-Smirnov statistika.

3.1 Giniho koeficient

Každému prvku $(\boldsymbol{x_i}, y_i)$ přiřadíme skóre, tj. číslo $s_i \in \mathbb{R}$. Čím větší očekáváme pravděpodobnost $\boldsymbol{P}(Y_i = 1)$, tím větší skóre pro prvek $(\boldsymbol{x_i}, y_i)$ volíme. V binární logistické regresi je zvyklostí volit jako skóre $q(\boldsymbol{x_i})$.

Seřadíme prvky vzestupně podle jejich skóre. Sestavíme dvě distribuční funkce pro tyto prvky. Jednu pro prvky s $y_i = 0$, druhou pro prvky s $y_i = 0$.

$$\begin{aligned} F_0(s) &= \frac{1}{n_0} \sum_{i=1}^{n} \mathbb{I}_{(-\infty < s_i \leq s)}(1 - y_i) \\ F_1(s) &= \frac{1}{n_1} \sum_{i=1}^{n} \mathbb{I}_{(-\infty < s_i \leq s)} y_i \,, \end{aligned}$$

kde n_1 je počet prvků, pro které $y_i = 1$ a n_0 je počet prvků, pro které $y_i = 0$.

Definujme parametricky Lorentzovu křivku.

$$\begin{aligned} x &= F_0(s) \\ y &= F_1(s)\,, \qquad s \in (-\infty, \infty) \end{aligned}$$

Lorentzova křivka je neklesající reálná funkce spojující body o souřadnicích $[0, 0]$ a $[1, 1]$. Označme A obsah orientované plochy mezi diagonálou jednotkového čtverce a Lorentzovou křivkou. Dále označme B obsah plochy pod Lorentzovou křivkou.

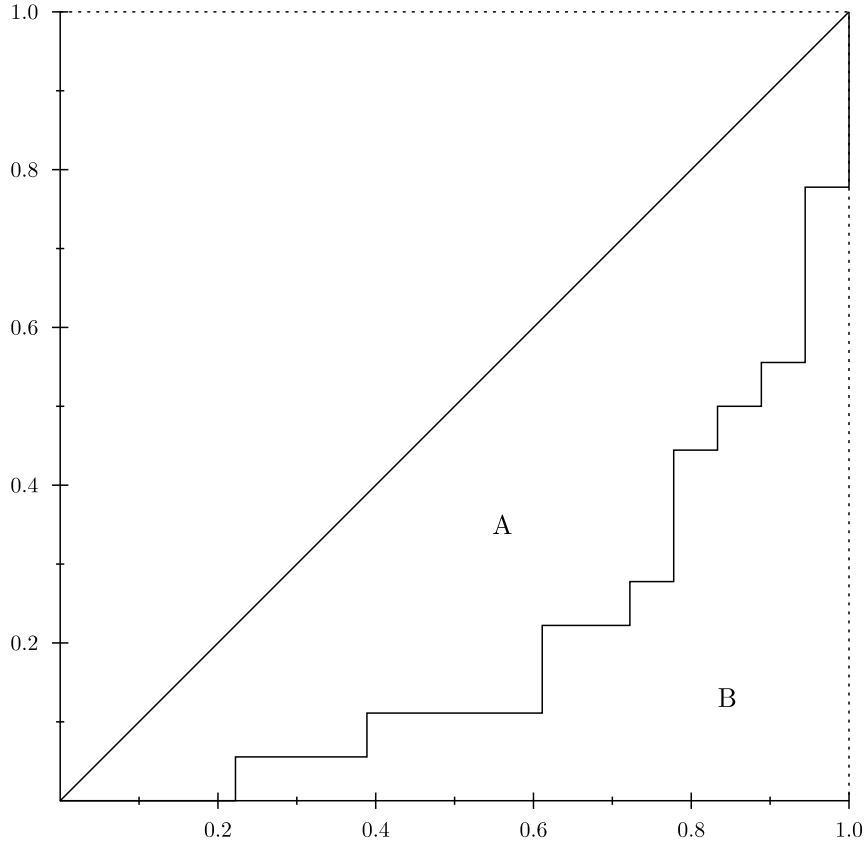
Giniho koeficient definujeme jako

$$G = \frac{A}{A + B} = 1 - 2B\,. \tag{3.1}$$

Tato definice je ekvivalentní s definicí

$$G = 1 - 2 \int_0^1 F_1(s) \mathrm{d}F_0(s)\,.$$

Vidíme, že $G \in [-1, 1]$, přičemž $G = 1$ nastane pro případ, kdy mají všechny prvky, pro které je $y_i = 0$, skóre menší než prvky s $y_i = 1$. Naopak je $G = -1$, pokud všechny prvky s $y_i = 0$ mají skóre větší, než prvky s $y_i = 1$. Zjednodušeně lze říct, že čím je větší Giniho koeficient, tím přesněji jsme zvolili skóre prvků. $G = 0$ znamená nulovou závislost skóre a hodnot y_i. Záporný Giniho koeficient znamená, že vysoké skóre indikuje spíše $y_i = 0$, tedy je model postaven obráceně.



Obrázek 3.1: Lorentzova křivka

Pro počítání Giniho koeficientu v praxi je potřeba umět jednoduše vyjádřit obsah B plochy pod Lorentzovou křivkou. Tuto plochu můžeme rozdělit na úzké obdélníky (viz obr. 3.2). Každý z těchto obdélníků má šířku $\frac{1}{n_1}$ a výšku $\frac{C_i}{n_0}$, kde C_i značí počet prvků $(\boldsymbol{x_j}, y_j)$ pro které $y_j = 0, j > i$. Odtud

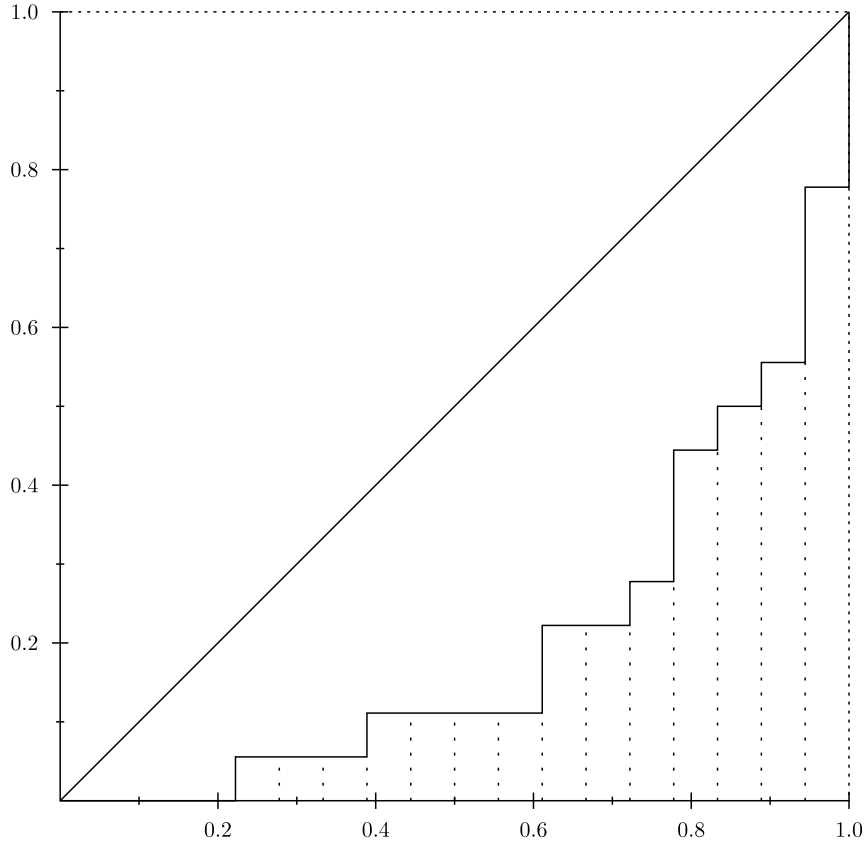
$$B = \sum_{i=1}^{n} y_i \frac{1}{n_1} \frac{C_i}{n_0} \,. \tag{3.2}$$

Všimněme si, že lze vyjádřit

$$C_i = \sum_{j=i}^{n} (1 - y_j) \,.$$

Dosaďme tuto rovnost do (3.2) a následně výraz upravme.

$$\begin{aligned} B &= \sum_{i=1}^{n} y_i \frac{1}{n_1} \frac{C_i}{n_0} \\ &= \sum_{i=1}^{n} y_i \frac{1}{n_1} \frac{\sum_{j=i}^{n} (1 - y_j)}{n_0} \\ &= \frac{1}{n_0 n_1} \sum_{i=1}^{n} \sum_{j=i}^{n} y_i (1 - y_j) \\ &= \frac{1}{n_0 n_1} \left[\sum_{i=1}^{n} \sum_{j=i}^{n} y_i - \sum_{i=1}^{n} \sum_{j=i}^{n} y_i y_j \right] \end{aligned}$$



Obrázek 3.2: Obsah pod Lorentzovou křivkou rozdělen na obdélníky.

$$
\begin{aligned}
&= \frac{1}{n_0 n_1}\left[\sum_{i=1}^{n}(n+1-i)y_i - \sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n}\sum_{j>i}^{n} y_i y_j\right] \\
&= \frac{1}{n_0 n_1}\left[\sum_{i=1}^{n}(n+1-i)y_i - n_1 - \binom{n_1}{2}\right] \\
&= \frac{1}{n_0 n_1}\left[n_1(n+1) - \sum_{i=1}^{n} i y_i - n_1 - \frac{n_1(n_1-1)}{2}\right] \qquad (3.3)
\end{aligned}
$$

Spočítejme Giniho koeficient pomocí dosazení (3.3) do (3.1).

$$
\begin{aligned}
G &= 1 - 2B \\
&= 1 - 2\frac{1}{n_0 n_1}\left[n_1(n+1) - \sum_{i=1}^{n} i y_i - n_1 - \frac{n_1(n_1-1)}{2}\right] \\
&= \frac{1}{n_0 n_1}\left[-n n_1 - n_1 + 2\sum_{i=1}^{n} i y_i\right] \\
&= \frac{2}{n_0 n_1}\sum_{i=1}^{n} i y_i - \frac{n+1}{n_0}
\end{aligned}
$$

Pokud prvky nemáme seřazené podle skóre, tak Giniho koeficient spočítáme jako

$$G = \frac{2}{n_0 n_1} \sum_{i=1}^{n} R_i y_i - \frac{n+1}{n_0}\,, \tag{3.4}$$

kde R_i je pořadí skóre prvku $(\boldsymbol{x_i}, y_i)$. Giniho koeficient je tedy závislý pouze na pořadí skóre jednotlivých prvků.

S Giniho koeficientem přímo souvisí C-statistika, která je definována jako pravděpodobnost, že skóre náhodně vybraného prvku s $y_i = 0$ je menší, než skóre náhodně vybraného prvku s $y_i = 1$.

$$C = \frac{\sum_{s_j \in S_1} \sum_{s_i \in S_0} \mathbb{I}_{(s_j > s_i)}}{n_0 n_1}\,,$$

kde S_0 je množina skóre prvků s $y_i = 0$ a S_1 je množina skóre prvků s $y_i = 1$. Elementárními úpravami dostaneme vyjádření C-statistiky pomocí Giniho koeficientu.

$$\begin{aligned} C &= \frac{\sum_{s_j \in S_1} \sum_{s_i \in S_0} \mathbb{I}_{(s_j > s_i)}}{n_0 n_1} = \frac{\sum_{s_j \in S_1} (n_0 - C_j)}{n_0 n_1} = \frac{\sum_{i=1}^{n} y_i (n_0 - C_i)}{n_0 n_1} \\ &= 1 - \frac{\sum_{i=1}^{n} y_i C_i}{n_0 n_1} = 1 - \sum_{i=1}^{n} y_i \frac{1}{n_1} \frac{C_i}{n_0} = 1 - B = \frac{1+G}{2} \end{aligned}$$

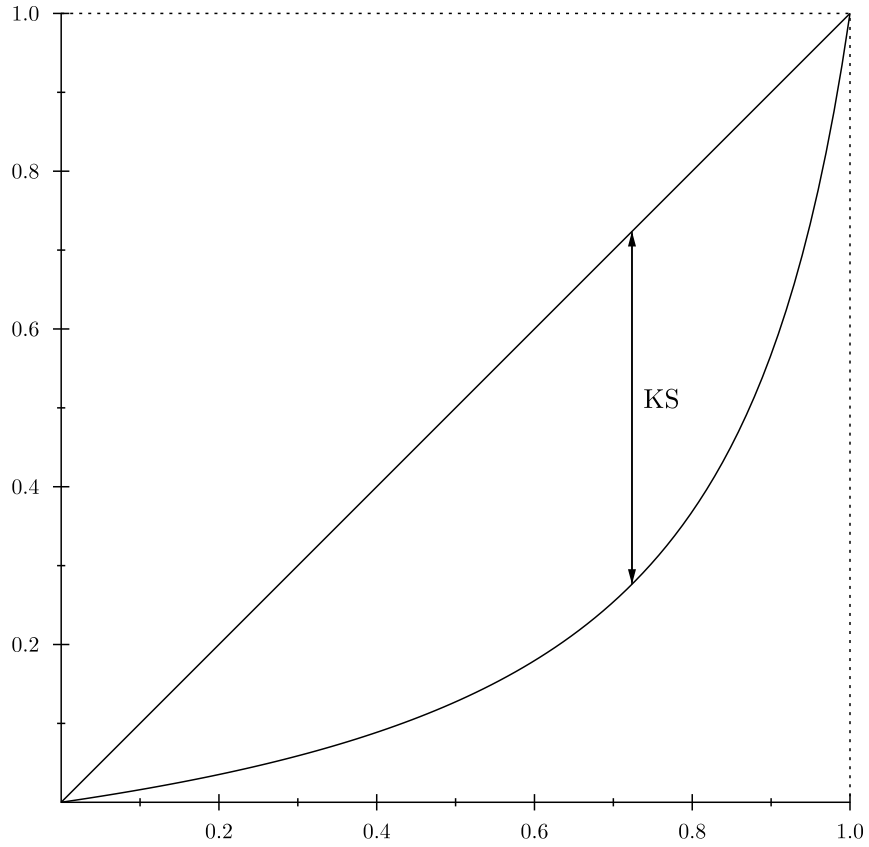
Připomeňme, že C_i značí počet prvků, pro které $y_j = 0, j > i$. Předpokládáme, že všechny prvky mají různé skóre. Pokud existují 2 prvky $s_j \in S_1, s_i \in S_0$ se stejným skóre, tak identifikátor $\mathbb{I}_{(s_j > s_i)}$ započítáme jako $\frac{1}{2}$. V případě C_i hodnotu navýšíme o $\frac{1}{2}$.

3.2 Kolmogorov-Smirnov statistika

Kolmogorov-Smirnov statistika je definována pomocí již zavedených distribučních funkcí F_0, F_1 následovně:

$$KS = \max_{s \in \mathbb{R}} \{F_0(s) - F_1(s)\} .$$

Kolmogorov-Smirnov statistika běžně nabývá hodnot z intervalu $[0, 1]$, kde hodnota $KS = 1$ značí dokonalou diverzifikační schopnost modelu (všechny prvky, pro které je $y_i = 0$ mají skóre menší než všechny prvky s $y_i = 1$). $KS = 0$ značí nulovou diverzifikační schopnost. Z definice plyne, že Kolmogorov-Smirnov statistika lze spočítat jako maximální vzdálenost bodů na Lorentzově křivce od diagonály jednotkového čtverce vynásobenou $\sqrt{2}$.



Obrázek 3.3: Kolmogorov-Smirnov statistika

Neexistuje žádný pevný vztah mezi Kolmogorov-Smirnov statistikou a Giniho koeficientem. Lze dokázat, že za předpokladu, že skóre prvků jsou normálně rozdělená (tvoří náhodný výběr z normálního rozdělení), platí mezi Kolmogorov-Smirnov statistikou a Giniho koeficientem téměř lineární závislost.

$$KS \approx \frac{\sqrt{2}}{2} G$$

3.3 Koeficient determinace R^2

V lineární regresi se používá k určení diverzifikační schopnosti modelu koeficient determinace R^2. Můžeme jej ekvivalentně definovat i pro model binární logistické regrese vztahem

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - q(\boldsymbol{x_i}))^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2},$$

kde $\overline{y}$ je aritmetický průměr $y_i, i = 1, 2, \ldots, n$. V binární logistické regresi se častěji než normální R^2 používají tzv. pseudo R kvadráty. Jedním z nich je pseudo R^2 McFadden. Vyjadřuje míru věrohodnosti odhadu $\boldsymbol{\beta}$. Definujeme jej jako

$$R^2_{\text{McFadden}} = 1 - \frac{L(q)}{L(\overline{y})} = 1 - \frac{\sum_{i=1}^{n}[(1-y_i)\ln(1-q(\boldsymbol{x_i})) + y_i \ln(q(\boldsymbol{x_i}))]}{\sum_{i=1}^{n}[(1-y_i)\ln(1-\overline{y}) + y_i \ln(\overline{y})]}.$$

$L(q)$ zde představuje zlogaritmovanou věrohodnost, $L(\overline{y})$ zlogaritmovanou věrohodnost nevyužívající vysvětlující proměnné $\boldsymbol{x_i}$.

$R^2_{\text{McFadden}} = 1$ implikuje dokonalou diverzifikační schopnost modelu. V praxi je diverzifikační schopnost modelu výborná, pokud $R^2_{\text{McFadden}} > 0.3$.

4. Maximalizace Giniho koeficientu

Pokusíme se na základě různých teoretických úvah vymyslet metodu odhadu parametru $\boldsymbol{\beta}$ modelu, která bude dávat co největší Giniho koeficient. Vzhledem k charakteru výpočtu Giniho koeficientu je tato úloha velmi obtížná a není v našiách silách najít parametr $\boldsymbol{\beta}$, který by Giniho koeficient maximalizoval úplně.

Ve větě 2.1 jsme dokázali, že v současné době používané odhadové techniky - metoda maximální věrohodnosti a metoda nejmenších čtverců - v jistém smyslu vedou na Fisher-konzistentní odhad parametru $\boldsymbol{\beta}$. Ukážeme, že metoda maximalizují Giniho koeficient taktéž asymptoticky vede na skutečnou hodnotu parametru $\boldsymbol{\beta}$ ve smyslu

$$\boldsymbol{\beta}_{\mathrm{R}} \in \mathrm{argmax}_{\boldsymbol{\beta}} \boldsymbol{E}G\,,$$

kde $\boldsymbol{\beta}_{\mathrm{R}}$ značí skutečnou hodnotu parametru $\boldsymbol{\beta}$ a $\boldsymbol{E}G$ je střední hodnota Giniho koeficientu spočítaného pomocí odhadu $\boldsymbol{\beta}$ a y_i vygenerovaných z $\mathrm{Alt}(q(\boldsymbol{x_i}))$.

Pro pořadí skóre R_i vypočítaného pomocí $\boldsymbol{\beta}$, podle vyjádření Giniho koeficientu z (3.4) platí

$$\begin{aligned}
& \mathrm{argmax}_{\boldsymbol{\beta}} \boldsymbol{E}G \\
= \; & \mathrm{argmax}_{\boldsymbol{\beta}} \boldsymbol{E} \left(\frac{2}{n_0 n_1} \sum_{i=1}^{n} R_i y_i - \frac{n+1}{n_0} \right) \\
= \; & \mathrm{argmax}_{\boldsymbol{\beta}} \left(\frac{2}{n_0 n_1} \sum_{i=1}^{n} R_i \boldsymbol{E} y_i - \frac{n+1}{n_0} \right) \\
= \; & \mathrm{argmax}_{\boldsymbol{\beta}} \sum_{i=1}^{n} R_i q(\boldsymbol{x_i})\,.
\end{aligned}$$

Z lineární algebry víme, že hodnota skalárního součinu

$$\sum_{i=1}^{n} R_i q(\boldsymbol{x_i})$$

je největší, pokud $q(\boldsymbol{x_i})$ a R_i mají stejné seřazení dle velikosti. Tuto podmínku splňuje $\boldsymbol{\beta}_{\mathrm{R}}$, a proto platí $\boldsymbol{\beta}_{\mathrm{R}} \in \mathrm{argmax}_{\boldsymbol{\beta}} \boldsymbol{E}G$.

Z toho tedy plyne, že na rozumných a dostatečně velkých datech budou všechny uvažované odhadové techniky dávat velice podobné výsledky jak odhadu parametru, tak jednotlivých ukazatelů těsnosti modelu. Toto se nám potvrdilo v kapitole 5.

4.1 Maximalizace Giniho koeficientu pro normálně rozdělená skóre

Abychom se mohli pokusit o maximalizaci Giniho koeficientu, nejdříve se musíme podívat, na čem tento koeficient závisí. Připomeňme si definici Giniho koeficientu.

$$G = \frac{2}{n_0 n_1} \sum_{i=1}^{n} R_i y_i - \frac{n+1}{n_0}\,,$$

kde R_i je pořadí skóre i-tého prvku. Jelikož n_0, n_1 a n jsou konstanty, maximalizace Giniho koeficientu je ekvivalentní maximalizaci skalárního součinu

$$\sum_{i=1}^{n} R_i y_i \,.$$

Zkusme předpokládat, že jsou skóre prvků normálně rozdělená. Tedy pro

$$sk_i := \sum_{k=0}^{m} \beta_k x_{ik}$$

představuje $\{sk_i, j = 1, 2, \ldots, n\}$ náhodný výběr z $\mathcal{N}(\mu, \sigma^2)$. Pro případ, kdy odhadujeme větší množství parametrů $\beta_0, \beta_1, \ldots, \beta_m$, (většinou stačí $m > 10$) a pro málo vzájemně závislé vysvětlující proměnné $x_{i1}, x_{i2}, \ldots, x_{im}$ je předpoklad reálný díky CLV. Odhadněme μ výběrovým průměrem

$$\overline{sk_n} = \frac{1}{n} \sum_{i=1}^{n} sk_i \,,$$

σ^2 pomocí výběrového rozptylu

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(sk_i - \overline{sk_n} \right)^2 \,.$$

Pro dostatečně velké n můžeme předpokládat, že jsou integrály

$$\int_{-\infty}^{sk_i} \frac{1}{\sqrt{2\pi} S_n} \exp\left\{ -\frac{(t - \overline{sk_n})^2}{2S_n^2} \right\} \mathrm{d}t \qquad i = 1, 2, \ldots, n$$

rovnoměrně rozdělené na intervalu $[0, 1]$. To nám umožní odhadnout pořadí skóre i-tého prvku

$$R_i \sim n \int_{-\infty}^{sk_i} \frac{1}{\sqrt{2\pi} S_n} \exp\left\{ -\frac{(t - \overline{sk_n})^2}{2S_n^2} \right\} \mathrm{d}t \,.$$

Giniho koeficient je tedy největší, když následující výraz nabývá svého maxima

$$V(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \int_{-\infty}^{sk_i} \frac{1}{\sqrt{2\pi} S_n} \exp\left\{ -\frac{(t - \overline{sk_n})^2}{2S_n^2} \right\} \mathrm{d}t \,. \tag{4.1}$$

Předpokládejme, že je V konvexní. Nabízí se numericky odhadnout parametr $\boldsymbol{\beta}$ Newtonovou metodou tečen. To ovšem není přímo možné, protože Jacobiho matice druhých parciálních derivací V není regulární. Víme, že Giniho koeficient nezáleží na β_0 a je invariantní na přenásobení $\boldsymbol{\beta}$ kladnou reálnou konstantou. Odhadneme parametr $\boldsymbol{\beta}$ metodou maximální věrohodnosti. Zafixujeme β_0, β_1. Dále budeme odhadovat pouze parametry $\beta_2, \beta_3, \ldots, \beta_m$. Označme

$$\hat{\boldsymbol{\beta}} = (\beta_2, \beta_3, \ldots, \beta_m) \,.$$

Tyto parametry odhadneme Newtonovou metodou tečen. Připomeňme si, že iterační krok v této metodě je

$$\hat{\boldsymbol{\beta}}_{new} = \hat{\boldsymbol{\beta}}_{old} - (\partial^2_{\hat{\boldsymbol{\beta}}} V(\hat{\boldsymbol{\beta}}_{old}))^{-1} (\partial_{\hat{\boldsymbol{\beta}}} V(\hat{\boldsymbol{\beta}}_{old})) \,.$$

Spočtěme $\partial_{\hat{\boldsymbol{\beta}}} V(\hat{\boldsymbol{\beta}})$.

$$\begin{aligned}\frac{\partial}{\partial\beta_j}V(\hat{\boldsymbol{\beta}}) &= \frac{\partial}{\partial\beta_j}\sum_{i=1}^{n} y_i \int_{-\infty}^{sk_i} \frac{1}{\sqrt{2\pi}S_n}\exp\left\{-\frac{(t-\overline{sk_n})^2}{2S_n^2}\right\}\mathrm{d}t \\ &= \frac{\partial}{\partial\beta_j}\sum_{i=1}^{n} y_i \int_{-\infty}^{sk_i-\overline{sk_n}} \frac{1}{\sqrt{2\pi}S_n}\exp\left\{-\frac{t^2}{2S_n^2}\right\}\mathrm{d}t \qquad (4.2)\end{aligned}$$

Pro jednoduchodst počítejme s tím, že pro všechny indexy $i \in 1, 2, \ldots, n$ je $sk_i - \overline{sk_n} > 0$. Pro indexy i takové, že $sk_i - \overline{sk_n} < 0$, vyjde integrál z výrazu (4.2) pouze o konstantu jinak, než pro $sk_i - \overline{sk_n} > 0$, a proto je derivace stejná. V případě $sk_i - \overline{sk_n} = 0$ vyjde integrál jako konstanta. Derivace z něj je nulová, tedy stejná jako pro případ $sk_i - \overline{sk_n} > 0$. Pokračujem v úpravě výrazu (4.2).

$$\frac{\partial}{\partial\beta_j}V(\hat{\boldsymbol{\beta}}) = \frac{\partial}{\partial\beta_j}\sum_{i=1}^{n} y_i \int_{-\infty}^{1} \frac{1}{\sqrt{2\pi}S_n}\exp\left\{-\frac{(t(sk_i-\overline{sk_n}))^2}{2S_n^2}\right\}(sk_i-\overline{sk_n})\mathrm{d}t$$

Tento výraz dále upravíme použitím Fubiniho věty.

$$= \sum_{i=1}^{n} y_i \int_{-\infty}^{1} \frac{\partial}{\partial\beta_j}\left[\frac{1}{\sqrt{2\pi}S_n}\exp\left\{-\frac{t^2(sk_i-\overline{sk_n})^2}{2S_n^2}\right\}(sk_i-\overline{sk_n})\right]\mathrm{d}t \qquad (4.3)$$

Použijme substituci $A_i = \frac{sk_i-\overline{sk_n}}{S_n}$ a dále upravme výraz (4.3).

$$\begin{aligned}&= \sum_{i=1}^{n} y_i \int_{-\infty}^{1} \frac{1}{\sqrt{2\pi}}\frac{\partial}{\partial\beta_j}\left[A_i\exp\left\{-\frac{t^2}{2}A_i^2\right\}\right]\mathrm{d}t \\ &= \sum_{i=1}^{n} y_i \int_{-\infty}^{1} \frac{1}{\sqrt{2\pi}}\left[A_i\exp\left\{-\frac{t^2}{2}A_i^2\right\}(-t^2A_i)A_i' + A_i'\exp\left\{-\frac{t^2}{2}A_i^2\right\}\right]\mathrm{d}t \\ &= \sum_{i=1}^{n} y_i \int_{-\infty}^{1} \frac{1}{\sqrt{2\pi}}\left[A_i'\exp\left\{-\frac{t^2}{2}A_i^2\right\}\left[A_i(-t^2A_i)+1\right]\right]\mathrm{d}t \\ &= \sum_{i=1}^{n} y_iA_i' \int_{-\infty}^{1} \frac{1}{\sqrt{2\pi}}\left[\exp\left\{-\frac{t^2}{2}A_i^2\right\}\left[A_i(-t^2A_i)+1\right]\right]\mathrm{d}t \\ &= \sum_{i=1}^{n} y_iA_i'\frac{1}{\sqrt{2\pi}}\left[\int_{-\infty}^{1}\exp\left\{-\frac{t^2}{2}A_i^2\right\}\mathrm{d}t - \int_{-\infty}^{1}\exp\left\{-\frac{t^2}{2}A_i^2\right\}A_i^2t^2\mathrm{d}t\right], \qquad (4.4)\end{aligned}$$

kde $A_i' = \frac{\partial}{\partial\beta_j}A_i$. Pomocí metody *per-partes* upravme integrál

$$\begin{aligned}\int_{-\infty}^{1}\exp\left\{-\frac{t^2}{2}A_i^2\right\}A_i^2t^2\mathrm{d}t &= \left[\exp\left\{-\frac{t^2}{2}A_i^2\right\}(-t)\right]_{-\infty}^{1} + \int_{-\infty}^{1}\exp\left\{-\frac{t^2}{2}A_i^2\right\}\mathrm{d}t \\ &= -\exp\left\{-\frac{1}{2}A_i^2\right\} + \int_{-\infty}^{1}\exp\left\{-\frac{t^2}{2}A_i^2\right\}\mathrm{d}t\,.\end{aligned}$$

Tuto rovnost dosadíme zpět do (4.4) a získáme tak

$$\frac{\partial}{\partial\beta_j}V(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} y_i\frac{\partial A_i}{\partial\beta_j}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}A_i^2\right\}. \qquad (4.5)$$

Spočtěme $\partial^2_{\hat{\boldsymbol{\beta}}} V(\hat{\boldsymbol{\beta}})$.

$$\frac{\partial^2}{\partial\beta_j \partial\beta_k} V(\hat{\boldsymbol{\beta}}) = \frac{\partial}{\partial\beta_k}\left(\frac{\partial}{\partial\beta_j} V(\hat{\boldsymbol{\beta}})\right)$$

Dosadíme z (4.5) a výraz dále upravíme.

$$\begin{aligned}
&= \frac{\partial}{\partial\beta_k} \sum_{i=1}^{n} y_i \frac{\partial A_i}{\partial\beta_j} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}A_i^2\right\} \\
&= \sum_{i=1}^{n} y_i \frac{\partial}{\partial\beta_k}\left[\frac{\partial A_i}{\partial\beta_j} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}A_i^2\right\}\right] \\
&= \sum_{i=1}^{n} y_i \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}A_i^2\right\}\left[\frac{\partial^2 A_i}{\partial\beta_j\partial\beta_k} - A_i \frac{\partial A}{\partial\beta_j}\frac{\partial A_i}{\partial\beta_k}\right] \qquad (4.6)
\end{aligned}$$

Ve výrazech (4.5) a (4.6) jsme spočítali $\partial_{\hat{\boldsymbol{\beta}}} V(\hat{\boldsymbol{\beta}})$ a $\partial^2_{\hat{\boldsymbol{\beta}}} V(\hat{\boldsymbol{\beta}})$. Jsme tedy schopni provést iterační krok v Newtonově metodě tečen a najít $\hat{\boldsymbol{\beta}}_{new}$. Po dostatečném počtu iterací získáme odhad parametru $\hat{\boldsymbol{\beta}}$ teoreticky dávající maximální Giniho koeficient. Této odhadové metodě říkejme maxGini.

Pro lepší představu o tom, jak tato metoda odhadu parametru funguje, zkusme najít ztrátovou funkci $\boldsymbol{L}(y|q)$, aby $\boldsymbol{\mathcal{L}}(\boldsymbol{\beta})$ nabýval svého minima pro stejný parametr $\boldsymbol{\beta}$, pro který by výraz (4.1) nabýval svého maxima. Potom tato ztrátová funkce teoreticky povede k parametru $\boldsymbol{\beta}$ dávajícímu maximální Giniho koeficient. Tuto podmínku splňuje „ztrátová“ funkce definovaná následovně

$$\boldsymbol{L}(y|q(\boldsymbol{x_i})) = 1 - y_i \int_{-\infty}^{sk_i} \frac{1}{\sqrt{2\pi}S_n} \exp\left\{-\frac{(t-\overline{sk_n})^2}{2S_n^2}\right\} \mathrm{d}t\,.$$

Chtěli bychom mít tuto funkci vyjádřenou pouze pomocí proměnných y a q, aby se skutečně jednalo o ztrátovou funkci. Ze vztahu

$$q(\boldsymbol{x_i}) = \frac{\exp\{\boldsymbol{\beta' x_i}\}}{1+\exp\{\boldsymbol{\beta' x_i}\}}$$

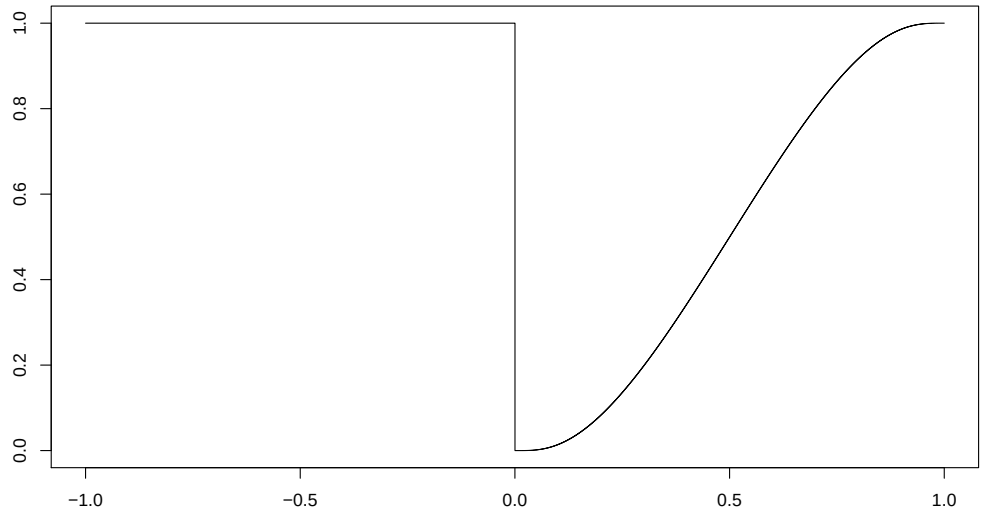
vyjádříme sk_i.

$$sk_i = \sum_{k=0}^{m} \beta_k x_{ik} = \boldsymbol{\beta' x_i} = \ln\left(\frac{q(\boldsymbol{x_i})}{1-q(\boldsymbol{x_i})}\right)$$

Nejsme ovšem schopni vyjádřit $\overline{sk_n}$ a S_n pomocí $q(\boldsymbol{x_i})$. Aby $\boldsymbol{L}(y|q(\boldsymbol{x_i}))$ byla ztrátová funkce, zvolíme $\overline{sk_n} = K_1$ a $S_n = K_2$ jako konstanty. Ztrátová funkce má poté tvar

$$\boldsymbol{L}(y|q) = 1 - y \int_{-\infty}^{\ln\left(\frac{q}{1-q}\right)} \frac{1}{\sqrt{2\pi}K_2} \exp\left\{-\frac{(t-K_1)^2}{2K_2^2}\right\} \mathrm{d}t\,. \qquad (4.7)$$

Pokusme se určit konstanty K_1, K_2 a získat tak novou metodu odhadu parametru $\boldsymbol{\beta}$ založenou na ztrátové funkci (4.7). Aby ztrátová funkce spolehlivěji vedla k dobrému odhadu, tak je dobré, aby byla symetrická. Proto ji pro $r < 0$ předefinujeme $\boldsymbol{L}(r) = \boldsymbol{L}(|r|)$.



Obrázek 4.1: Graf „ztrátové funkce $\boldsymbol{L}(r)$ metody maxGini“ pro $K_1 = 0, K_2 = 1$.

Odhadneme parametr $\boldsymbol{\beta}$ metodou maximální věrohodnosti a dopočítáme $\overline{sk_n}$ a S_n. Tvar ztrátové funkce této nové odhadové metody je

$$\boldsymbol{L}(r) = 1 - \int_{-\infty}^{\ln\left(\frac{1-|r|}{|r|}\right)} \frac{1}{\sqrt{2\pi}S_n} \exp\left\{-\frac{(t-\overline{sk_n})^2}{2S_n^2}\right\} \mathrm{d}t\,. \tag{4.8}$$

Pro numerické počítání odhadu $\tilde{\boldsymbol{\beta}}$ parametru $\boldsymbol{\beta}$ použijeme IRLS. Váhovou matici $\boldsymbol{W}$ spočítáme jako

$$\begin{aligned} w_i &= \frac{\boldsymbol{L}'(r_i)}{r_i} q(\boldsymbol{x_i})^2 (1 - q(\boldsymbol{x_i}))^2 \\ &= \frac{q(\boldsymbol{x_i})^2(1-q(\boldsymbol{x_i}))^2}{r_i} \frac{\partial}{\partial r_i} \left(1 - y_i \int_{-\infty}^{\ln\left(\frac{1-r_i}{r_i}\right)} \frac{1}{\sqrt{2\pi}S_n} \exp\left\{-\frac{(t-\overline{sk_n})^2}{2S_n^2}\right\} \mathrm{d}t\right) \\ &= y_i \frac{1-r_i}{\sqrt{2\pi}S_n} \exp\left\{-\frac{\left(\ln\left(\frac{1-r_i}{r_i}\right) - \overline{sk_n}\right)^2}{2S_n^2}\right\}. \end{aligned}$$

Dále spočtěme pomocný vektor $\boldsymbol{z}$.

$$z_i = \frac{r_i}{w_i}$$

Nakonec se pokusíme odhad $\tilde{\boldsymbol{\beta}}$ vylepšit, aby dával pořád stejný Giniho koeficient, ale aby byl více věrohodný. Víme, že Giniho koeficient nezávisí na velikosti parametru $\tilde{\beta}_0$ a je invariantní na přenásobení $\tilde{\boldsymbol{\beta}}$ kladnou reálnou konstantou. Označme

$$\tilde{x}_{i0} = 1\,, \qquad \tilde{x}_{i1} = \sum_{j=1}^{m} \tilde{\beta}_j x_{ij}$$

pro $\forall i = 1, 2, \ldots, n$. Platí

$$q(\boldsymbol{x_i}) = \frac{1}{1 + e^{-\boldsymbol{\beta}'\boldsymbol{x_i}}} = \frac{1}{1 + e^{-\gamma_0 \tilde{x}_{i0} - \gamma_1 \tilde{x}_{i1}}}\,.$$

Metodou maximální věrohodnosti odhadneme parametry γ_0, γ_1. Získáme tak odhad

$$(\gamma_0, \gamma_1\tilde{\beta}_1, \gamma_1\tilde{\beta}_2, \ldots, \gamma_1\tilde{\beta_m})$$

parametru $\boldsymbol{\beta}$ teoreticky dávající maximální Giniho koeficient, jenž je na množině parametrů, dávajících tento Giniho koeficient, nejvíce věrohodný. Této metodě odhadu parametru $\boldsymbol{\beta}$ říkejme maxGini2.

4.2 Maximalizace Giniho koeficientu pro obecně rozdělená skóre

Idea bude podobná, jako v případě maximalizace Giniho koeficientu pro normálně rozdělená skóre. Zde ovšem budeme počítat s tím, že distribuční funkce rozdělení skóre je nějaká obecná distribuční funkce F (nezávislá na parametru $\boldsymbol{\beta}$) a hustota tohoto rozdělení je f. Největší Giniho koeficient získáme maximalizací výrazu

$$V(\boldsymbol{\beta}) = \sum_{i=1}^{n} F(sk_i)y_i\,.$$

Toto maximum se pokusíme najít Newtonovou metodou tečen. Iterační krok je

$$\hat{\boldsymbol{\beta}}_{new} = \hat{\boldsymbol{\beta}}_{old} - (\partial^2_{\hat{\boldsymbol{\beta}}} V(\hat{\boldsymbol{\beta}}_{old}))^{-1}(\partial_{\hat{\boldsymbol{\beta}}} V(\hat{\boldsymbol{\beta}}_{old}))\,.$$

Je třeba spočítat $\partial_{\hat{\boldsymbol{\beta}}} V(\hat{\boldsymbol{\beta}})$ a $\partial^2_{\hat{\boldsymbol{\beta}}} V(\hat{\boldsymbol{\beta}})$.

$$\begin{aligned}
\frac{\partial}{\partial\beta_j}V(\hat{\boldsymbol{\beta}}) &= \frac{\partial}{\partial\beta_j}\sum_{i=1}^{n} F(sk_i)y_i \\
&= \frac{\partial}{\partial\beta_j}\sum_{i=1}^{n} y_i \int_{-\infty}^{sk_i} f(t)\mathrm{d}t \\
&= \frac{\partial}{\partial\beta_j}\sum_{i=1}^{n} y_i \int_{-\infty}^{\sum_{j=1}^{m}\beta_j x_{ij}} f(t)\mathrm{d}t \\
&= \frac{\partial}{\partial\beta_j}\sum_{i=1}^{n} y_i \int_{-\infty}^{\sum_{k=1}^{m}\frac{\beta_k x_{ik}}{x_{ij}}} f(tx_{ij})x_{ij}\mathrm{d}t \\
&= \sum_{i=1}^{n} y_i f(\beta_j x_{ij})x_{ij}
\end{aligned}$$

Pro $k = j$ je

$$\frac{\partial^2}{\partial\beta_j\partial\beta_k}V(\hat{\boldsymbol{\beta}}) = \frac{\partial}{\partial\beta_k}\left(\frac{\partial}{\partial\beta_j}V(\hat{\boldsymbol{\beta}})\right) = \frac{\partial}{\partial\beta_j}\sum_{i=1}^{n} y_i f(\beta_j x_{ij})x_{ij} = \sum_{i=1}^{n} y_i f'(\beta_j x_{ij})x_{ij}^2\,.$$

Pro $k \neq j$ je

$$\frac{\partial^2}{\partial\beta_j\partial\beta_k}V(\hat{\boldsymbol{\beta}}) = \frac{\partial}{\partial\beta_k}\left(\frac{\partial}{\partial\beta_j}V(\hat{\boldsymbol{\beta}})\right) = \frac{\partial}{\partial\beta_k}\sum_{i=1}^{n} y_i f(\beta_j x_{ij})x_{ij} = 0\,.$$

4.3 Maximalizace Giniho koeficientu na beta rodině Fisher-konzistentních ztrátových funkcí

Na základě charakteru dat budeme hledat takové parametry α a β, které určují ztrátovou funkci, z beta rodiny Fisher-konzistentních ztrátových funkcí, dávající větší Giniho koeficient, než ztrátové funkce určené jinými parametry α, β. Připomeňme si tvar ztrátové funkce z beta rodiny Fisher-konzistentních ztrátových funkcí.

$$L_1(1-q) = \int_{1-q}^{1} t^{\alpha-1}(1-t)^{\beta}\mathrm{d}t \qquad L_0(q) = \int_0^q t^{\alpha}(1-t)^{\beta-1}\mathrm{d}t$$

Pokusíme se maximalizovat Giniho koeficient tím, že najdeme ztrátovou funkci, určenou parametry α, β, co nejpodobnější ztrátové funkci metody maxGini2. Podle (4.8) je ztrátová funkce $\boldsymbol{L^1}(r)$ maxGini2

$$\boldsymbol{L^1}(r) = 1 - \int_{-\infty}^{\ln\left(\frac{1-|r|}{|r|}\right)} \frac{1}{\sqrt{2\pi}S_n} \exp\left\{-\frac{(t-\overline{sk_n})^2}{2S_n^2}\right\} \mathrm{d}t\,,$$

kde $\overline{sk_n}$ a S_n jsou spočítány pomocí parametru $\boldsymbol{\beta}$ odhadnutého metodou maximální věrohodnosti. Ztrátová funkce z beta rodiny Fisher-konzistentních ztrátových funkcí je pro $r \in [-1, 0]$

$$\boldsymbol{L^2}(r) = \int_0^{-r} t^{\alpha}(1-t)^{\beta-1}\mathrm{d}t\,.$$

Na intervalu $[0, 1]$ vypadá ztrátová funkce takto

$$\boldsymbol{L^2}(r) = \int_{1-r}^{1} t^{\alpha-1}(1-t)^{\beta}\mathrm{d}t\,.$$

Metodou max. věrohodnosti odhadneme parametr $\boldsymbol{\beta}$. Následně spočteme hodnoty $r_i = y_i - q(\boldsymbol{x_i})$. Máme tak rozdělení r_i. Rozdíl, mezi ztrátovou funkcí maxGini2 a ztrátovou funkcí z beta rodiny Fisher-konzistentních funkcí s parametry α, β, definujme jako

$$R = \sum_{i=1}^{n} y_i|\boldsymbol{L^1}(r_i) - p_1\boldsymbol{L^2}(r_i) + p_2| + (1-y_i)|\boldsymbol{L^1}(r_i) - p_1\boldsymbol{L^2}(r_i) + p_3|\,.$$

Jedná se o součet rozdílů ztrátových funkcí $\boldsymbol{L^1}$ a $\boldsymbol{L^2}$ v bodech r_i. Když ztrátovou funkci přenásobíme kladnou reálnou konstantou nebo když ji posuneme o konstantu, tak bude vést pořád ke stejnému odhadu parametru $\boldsymbol{\beta}$. Proto jsme do definice rozdílu ztrátových funkcí přidali parametry p_1, p_2, p_3.

Chceme-li nalézt ztrátovou funkci nejpodobnější ztrátové funkci maxGini2, zabýváme se úlohou minimalizace výrazu R v závislosti na proměnných $\alpha, \beta, p_1, p_2, p_3$.

5. Aplikace na data

V předchozích kapitolách jsme definovali model binární logistické regrese. Uvedli jsme tři používané metody odhadu parametru $\boldsymbol{\beta}$ - metodu maximální věrohodnosti (ML) a metodu nejmenších čtverců (OLS), metodu odvozenou z exponenciální ztrátové funkce (EXP). Ve čtvrté kapitole jsme odvodili dvě nové metody - maxGini a maxGini2. V této kapitole tyto metody (doplněné o metodu maxGiniAB - odvozenou viz níže) aplikujeme na různé sady reálných nebo simulovaných dat. Pro jednotlivé metody spočteme odhad parametru $\boldsymbol{\beta}$, Giniho koeficient (Gini), Kolmogorov-Smirnov statistiku (KS), koeficient determinace R^2 a pseudo R^2_{McFadden}. Poté jednotlivé metody navzájem srovnáme, a to i z hlediska časové náročnosti.

5.1 Dvojrozměrný případ

Uvažme situaci, kdy $m = 2$. Tedy $\boldsymbol{\beta}' = (\beta_0, \beta_1, \beta_2)$. Víme, že Giniho koeficient nezávisí na β_0 a je invariantní na přenásobení $\boldsymbol{\beta}$ kladnou reálnou konstantou. Proto pro případ $m = 2$ Giniho koeficient závisí pouze na poměru $\frac{\beta_1}{\beta_2}$.

Nasimulujeme si data „DvojrozmernyPripad“. Vysvětlující proměnné x_{i1} a x_{i2} vybereme náhodně nezávisle z normovaného normálního rozdělení. Zvolíme $\boldsymbol{\gamma} = (2, 1, 3)$. Spočítáme $\mu_i, i = 1, 2, \ldots, 200$ jako

$$\mu_i = \frac{1}{1 + \exp^{-\boldsymbol{\gamma}' \boldsymbol{x_i}}} .$$

Na závěr vygenerujeme vysvětlované proměnné $y_i, i = 1, 2, \ldots, 200$.

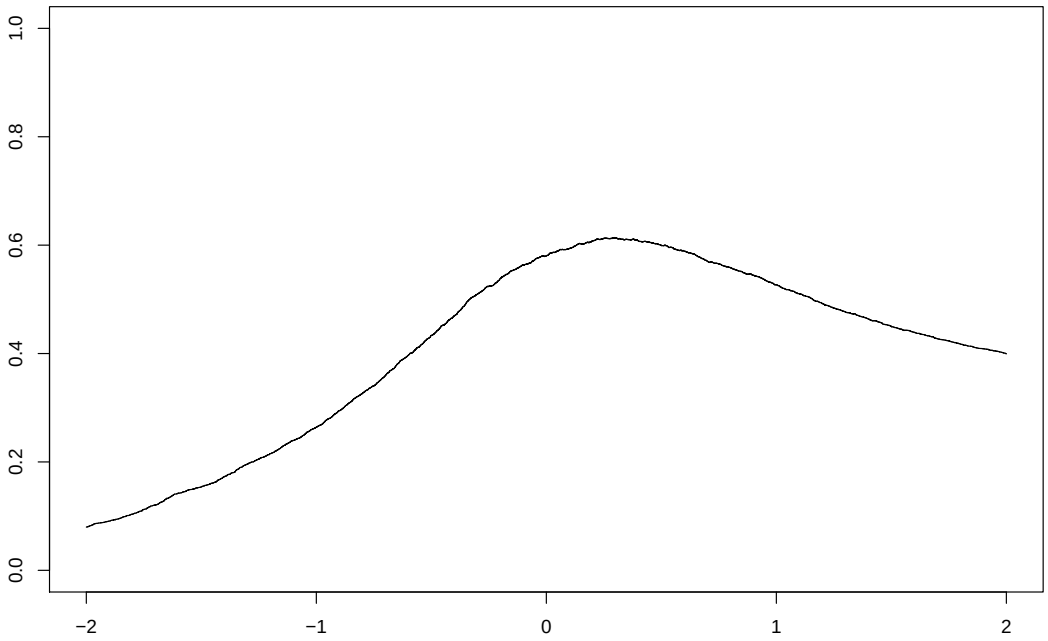
$$y_i = \begin{cases} 0 & q_i < 0.5 \\ 1 & q_i \geq 0.5 \end{cases},$$

kde q_i je náhodně vybráno z $\mathcal{N}(\mu_i, 0.4)$. Poměr hodnot $y_i = 1$ ku celkovému počtu hodnot je 0.705.

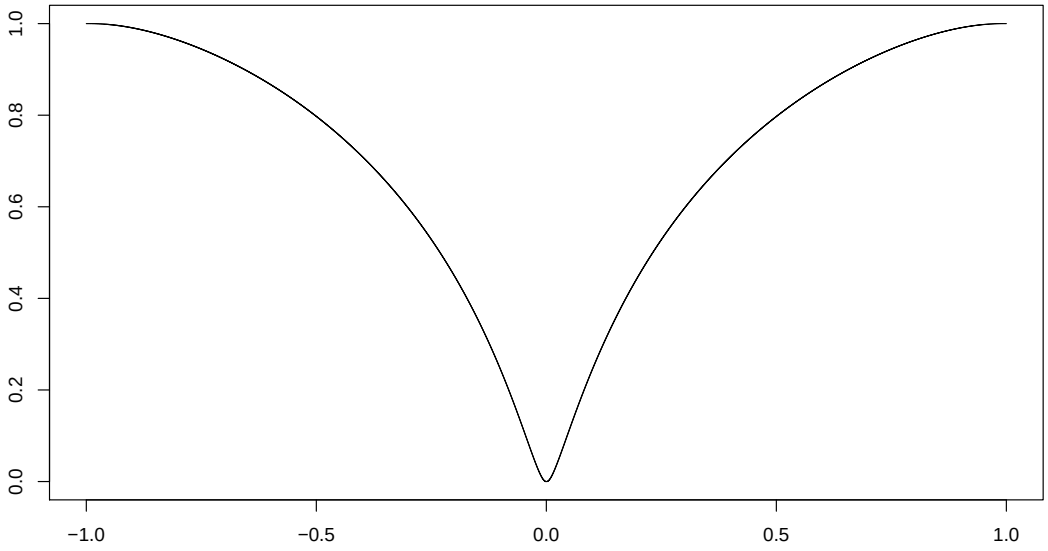
Pozn: Úmyslně nevolíme jednodušší způsob simulace $Y_i = \text{Alt}(\mu_i)$. Model by při takovém generování přesně odpovídal modelu binární logistické regrese. V praxi je to velmi nepravděpodobné a výsledky analýz by mohly být zkreslené, mohly by zvýhodňovat nějaké metody před ostatními.

Vykreslíme si graf závislosti Giniho koeficientu na $\frac{\beta_1}{\beta_2}$ (viz obr. 5.1). Giniho koeficient nabývá maximální hodnoty 0.614 pro $\frac{\beta_1}{\beta_2} = 0.296$. V tabulce 1 je spočítaný odhad parametru $\boldsymbol{\beta}$ a základní ukazatele těsnosti modelu pro různé odhadové metody. Pro lepší představu o tom, jak konkrétně na těchto datech funguje metoda maxGini2, se podívejme na její ztrátovou funkci na obrázku 5.2.

Všechny metody kromě metody nejmenších čtverců (OLS) odhadují parametr $\boldsymbol{\beta}$ velmi podobně, a proto jednotlivé ukazatele těsnosti modelu jsou srovnatelné. Metoda nejmenších čtverců se liší v odhadu parametru $\boldsymbol{\beta}$, nicméně jednotlivé ukazatele těsnosti modelu má přibližně stejně dobré jako ostatní metody. Pouze věrohodnostní R^2_{McFadden} je horší. Z toho můžeme usoudit, že teoretická skutečná hodnota parametru $\boldsymbol{\beta}$ je blíže odhadům od ostatních metod.

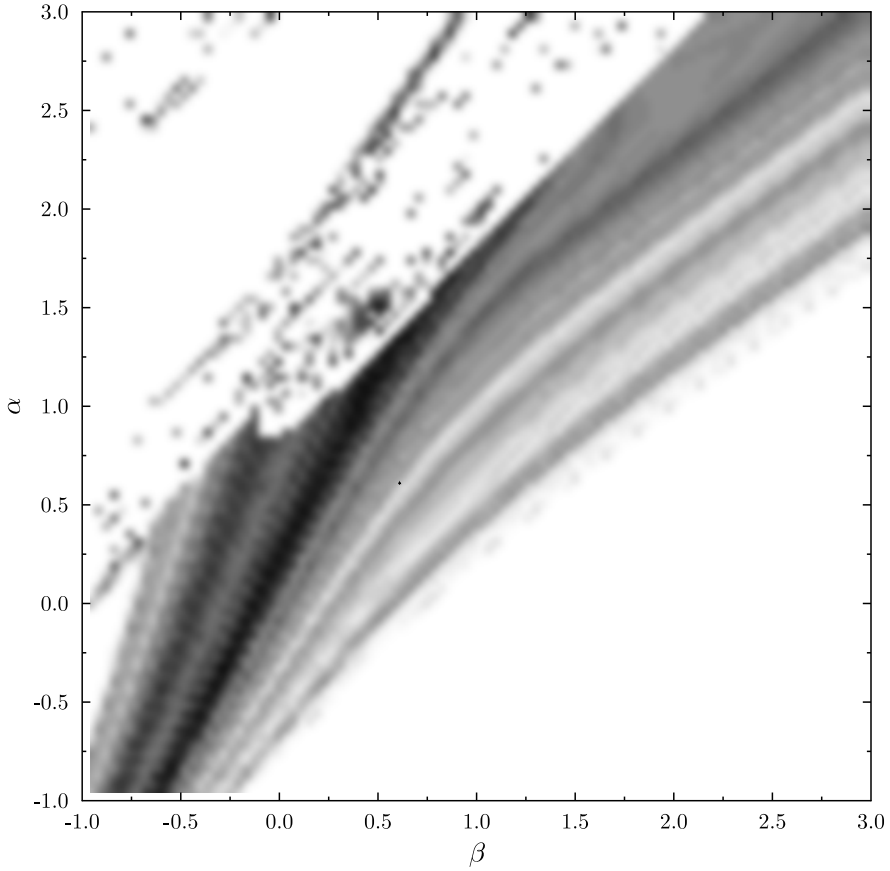


Obrázek 5.1: Graf závislosti Giniho koeficientu na poměru $\frac{\beta_1}{\beta_2}$.

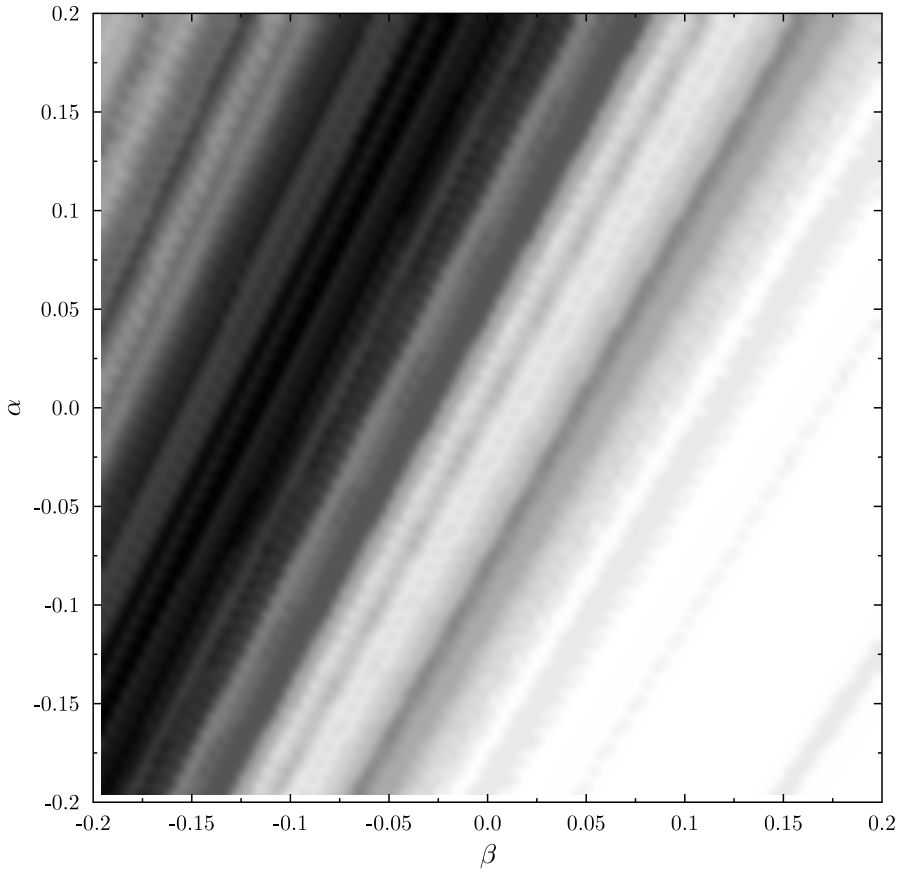


Obrázek 5.2: Ztrátová funkce metody maxGini2.

Dále se podívejme na analýzu toho, jaký Giniho koeficient dávají metody určené ztrátovými funkcemi z beta rodiny Fisher-konzistentních ztrátových funkcí. Tyto Giniho koeficienty jsou znázorněné na obrázcích 5.3 a 5.4. Čím tmavší místo v grafech je, tím větší Giniho koeficient dává metoda určená příslušnými parametry α, β.



Obrázek 5.3: Graf závislosti Giniho koeficientu na koeficientech α, β.



Obrázek 5.4: Graf závislosti Giniho koeficientu na koeficientech α, β. Výřez na okolí bodu o souřadnicích $\alpha = \beta = 0$.

Zajímavé je, že koeficienty α, β určující největší Giniho koeficienty leží přibližně na přímce procházející poblíž bodu $\alpha = \beta = 0$. Je tedy pravděpodobné, že když zvolíme parametry $\alpha = 0$ a vhodný $\beta \in [-0.3, 0.3]$, dostaneme metodu určující téměř maximální Giniho koeficient.

Na malých datech jsme si mohli dovolit odhadnout parametr $\boldsymbol{\beta}$ pro velké množství různých koeficientů α, β. Na větších datech nám čas dovolí udělat jen velmi omezený počet odhadů. Vzhledem k analýze z obrázků 5.1 a 5.2 definujme novou odhadovou metodu maxGiniAB. Pro sedm různých metod odvozených od ztrátových funkcí z beta rodiny Fisher-konzistentních ztrátových funkcích s parametry $\alpha = 0, \beta \in \{-0.3, -0.2, -0.1, 0, 0.1, 0.2, 0.3\}$ odhadneme parametr $\boldsymbol{\beta}$ modelu a pomocí něj spočítáme Giniho koeficient. Metoda maxGiniAB použije z těchto 7 odhadů $\boldsymbol{\beta}$ ten, který dává největší Giniho koeficient.

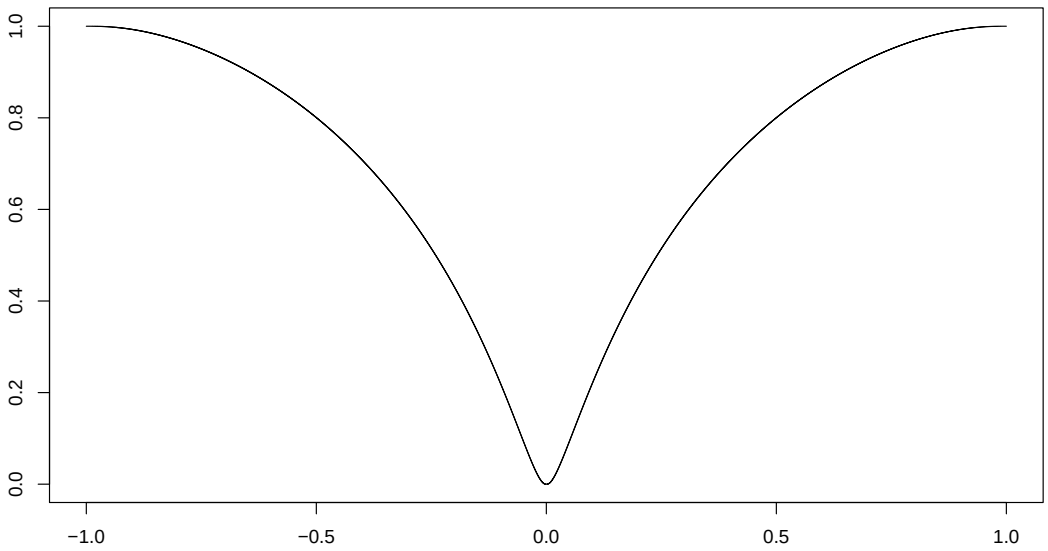
5.2 Reálná data

Analyzujeme jednotlivé odhadové metody na třech různých sadách reálných dat. První data „Kredit“, obsahující informace o 1000 klientech banky, pocházejí z webových stránek http://www.stat.uni-muenchen.de. Vysvětlovaná proměnná Y se jmenuje *kredit*. Hodnota $y_i = 1$ znamená, že i-tý klient splatil úvěr. Hodnota $y_i = 0$, že nesplatil. Poměr „dobrých“ klientů (tzn. podíl $y_i = 1$) je 0.7. V tabulce 2 jsou zaznamenány výsledky testování odhadových metod na těchto datech.

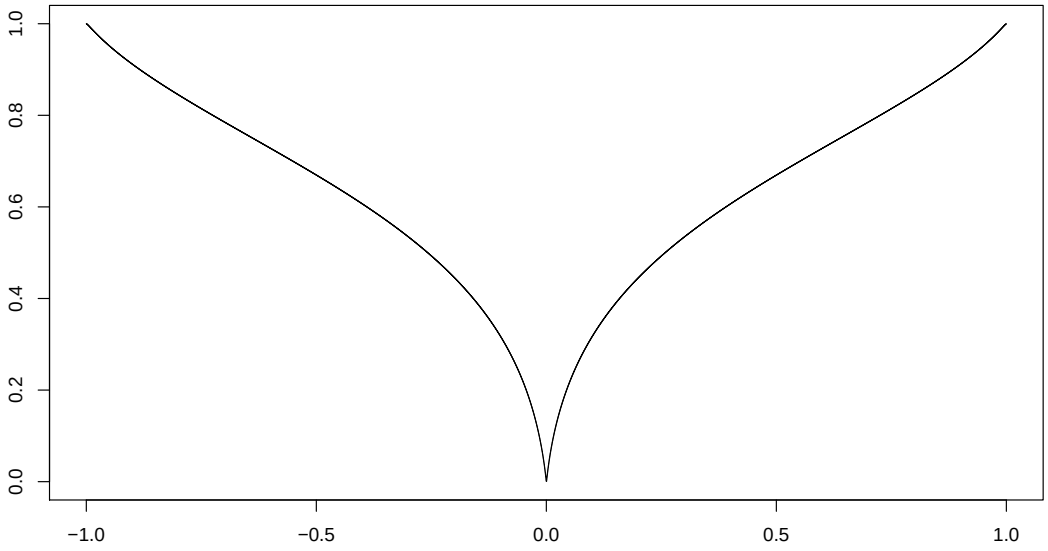
Druhá data „Bankloan“ opět obsahují informace o klientech banky. Jedná se o ukázková data na logistickou regresi, která používá na školení společnost ACREA ČR, spol. s.r.o. Vysvětlovaná proměnná Y značící, jestli daný klient splatil úvěr ($Y = 1$) nebo nesplatil ($Y = 0$), se v těchto datech jmenuje *default*. Poměr „dobrých“ klientů je 0.635 na vzorku 1500 pozorování. Aplikace odhadových metod na tato data jsme zaznamenali do tabulky 3.

Poslední data „Churn“ obsahují informace o 4431 zákaznících týkající se jejich vztahu k jistému poskytovateli služeb. Jsou to další data využívaná na školeních společnosti ACREA ČR, spol. s.r.o. Vysvětlovanou proměnnou je zde *Churn_ 2*. Hodnota *Churn_ 2* = 0 znamená, že zákazník u poskytovatele služeb zůstal. Hodnota *Churn_ 2* = 1, že odešel. Poměr „dobrých“ zákazníků (tzn. těch, kteří u poskytovatele zůstali) je 0.437. Před analyzováním dat musíme vyřadit proměnnou *Churn_ 3*, která podrobněji rozebírá, kam zákazníci od firmy odešli a *Churn_ 2* je na ní přímo závislá. Experimentováním se ukázalo, že proměnná *internat* (udávající počet minut mezinárodních hovorů) přináší do modelu singularitu a je potřeba ji taktéž vyřadit. Do tabulky 4 jsme zaznamenali aplikace odhadových metod na tato data. Všechny tři sady dat jsou k dispozici na přiloženém CD ve složce data.

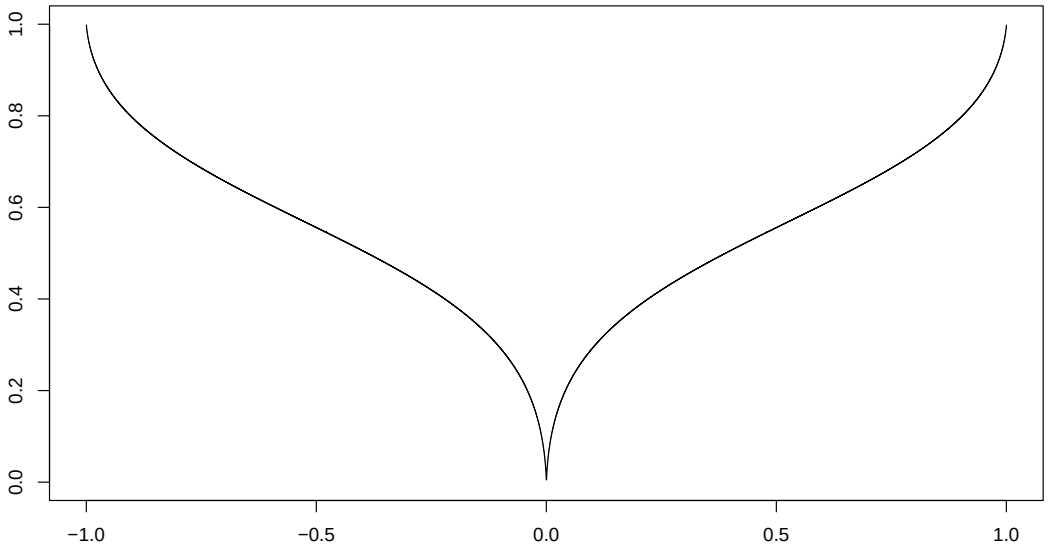
Na obrázcích 5.5, 5.6 a 5.7 jsou znázorněny ztrátové funkce metody maxGini2 pro jednotlivá data. Z tabulek 2, 3 a 4 vyčteme, že nevyšších hodnot ukazatelů těsnosti modelu dosahuje na všech třech datových sadách metoda maximální věrohodnosti (ML). Metody OLS, EXP a maxGiniAB mají srovnatelné výsledky. Metoda maxGini2 na datech „Kredit“ dosáhla dobrých výsledků. Na sadě dat „Bankloan“ byl odhad touto metodou nepřesný. Na datech „Churn“ metoda úplně selhala. Tento neúspěch byl pravděpodobně způsoben tím, že ztrátová funkce maxGini2 na těchto datech zdaleka není konvexní, a proto nefunguje hledání minima aritmetického průměru pozorování $\mathcal{L}(\boldsymbol{\beta})$ pomocí Newtonovy metody tečen. Metoda maxGini se na všech datech ukazuje jako nedostatečná a to přesto, že její běh zabral počítači řádově minuty, oproti ostatním metodám, které byly schopny odhadnout parametr modelu v řádu sekund. Giniho koeficient spočítaný pomocí metody maxGini sice není úplně špatný, ovšem z nízkých hodnot ukazatelů R^2 a R^2_{Mcfadden} můžeme usoudit, že odhad parametru $\boldsymbol{\beta}$ touto metodou bude velmi nepřesný. Vzhledem k tomu, že metoda nedává dle očekávání vyšší Giniho koeficient než ostatní metody, v praxi nemá smysl tuto metodu používat.



Obrázek 5.5: Graf ztrátové funkce $\boldsymbol{L}(r)$ pro data „Kredit“.



Obrázek 5.6: Graf ztrátové funkce $\boldsymbol{L}(r)$ pro data „Bankloan“.



Obrázek 5.7: Graf ztrátové funkce $\boldsymbol{L}(r)$ pro data „Churn“.

Závěr

Představili jsme model binární logistické regrese a odvodili jsme nejpoužívanější metodu odhadu jeho parametrů - metodu maximální věrohodnosti. Ve druhé kapitole jsme definovali ztrátovou funkci a ukázali jsme, jak s její pomocí odhadovat parametry modelu. Zaměřili jsme se na množinu ztrátových funkcí splňující podmínku Fisher-konzistence, kterou jsme nakonec zredukovali na stále velmi bohatou množinu - beta rodinu Fisher-konzistentních ztrátových funkcí. Tato množina obsahuje všechny běžně používané metody odhadu parametrů modelu binární logistické regrese, včetně metody maximální věrohodnosti.

K numerickým výpočtům odhadu parametrů modelu nám slouží metoda převážených nejmenších čtverců (IRLS). Pro ztrátové funkce z beta rodiny Fisher-konzistentních ztrátových funkcí jsme přímo odvodili tvar iterační rovnice v tomto algoritmu.

Ve čtvrté kapitole jsme se pokusili najít metodu, která by odhadla parametry modelu tak, aby maximalizovala Giniho koeficient. Za předpokladu normality rozdělení skóre jednotlivých prvků jsme navrhli teoreticky velmi dobře fungující metodu odhadu parametrů - maxGini. Z její ztrátové funkce jsme odvodili další, mnohem rychlejší, metodu odhadu parametrů - maxGini2. Při testování na reálných datech se ukázalo, že obě metody maxGini a maxGini2 jsou pro praktické použití nevyhovující.

Nejlepší metodou odhadu parametrů se jeví metoda maximální věrohodnosti. Také beta rodina Fisher-konzistentních ztrátových funkcí v sobě má velký potenciál. Ukázalo se, že často existují takové parametry α a β definující ztrátovou funkci z beta rodiny Fisher-konzistentních ztrátových funkcí, jež vedou k odhadu parametrů dávajícího větší Giniho koeficient než metoda maximální věrohodnosti. Předmětem dalšího zkoumání může být určení těchto parametrů α, β v závislosti na charakteru dat.

Seznam použité literatury

[1] Ondrušková, Markéta. *Odhadování a kritéria těsnosti modelu logistické regrese*. Bakalářská práce, MFF UK, 2011.

[2] Dupač, Václav a Hušková, Marie. *Pravděpodobnost a matematická statistika*. 1. vydání. Praha: Nakladatelství Karolinum 2009.
ISBN 978-80-246-0009-3.

[3] Andreas, Buja, Werner, Stuetzle a Yi, Shen. *Loss Functions for Binary Class Probability Estimation and Classification: Structure and Applications*. 2005.
http://www-stat.wharton.upenn.edu/∼buja/PAPERS/paper-proper-scoring.pdf [25.7.2012].

[4] Řezáč, Martin a Řezáč, František. *How to Mesure Quality of Credit Scoring Models*. Příspěvek na Konferenci na počest docenta Osvalda Vašíčka, Brno 22. 10. 2010.
http://is.muni.cz/do/econ/soubory/konference/vasicek/20667044/Rezac.pdf [25.7.2012].

[5] UCLA Academic Technology Services: FAQ: *What are pseudo R-squareds?*
http://www.ats.ucla.edu/stat/mult_pkg/faq/general/psuedo_rsquareds.htm [25.7.2012].

[6] Nagy, Ivan. *Teorie hromadné obsluhy*. Učební text, FD ČVUT, Praha. http://staff.utia.cas.cz/nagy/skola/Tho/Tho_lectures.pdf [25.7.2012].

Seznam tabulek

Název	$\boldsymbol{\beta}$	$\frac{\beta_1}{\beta_2}$	Gini	KS	R^2	R^2_{McFadden}
ML	(1.25, 0.43, 1.38)	0.312	0.612	0.569	0.293	0.224
OLS	(1.97, 0.76, 2.24)	0.338	0.609	0.577	0.308	0.169
EXP	(1.12, 0.35, 1.25)	0.277	0.612	0.566	0.284	0.221
maxGini	(1.25, 0.43, 1.89)	0.229	0.611	0.549	0.281	0.202
maxGini2	(1.25, 0.43, 1.38)	0.312	0.612	0.569	0.293	0.224
maxGiniAB	(1.21, 0.40, 1.38)	0.300	0.614	0.559	0.290	0.223

Tabulka 1: Analýza dat „DvojrozmernyPripad“.

Název	Gini	KS	R^2	R^2_{McFadden}
ML	0.607	0.489	0.257	0.217
OLS	0.605	0.487	0.258	0.215
EXP	0.606	0.483	0.254	0.216
maxGini	0.603	0.483	0.091	0.078
maxGini2	0.607	0.489	0.257	0.217
maxGiniAB	0.607	0.491	0.255	0.217

Tabulka 2: Analýza dat „Kredit“.

Název	Gini	KS	R^2	R^2_{McFadden}
ML	0.692	0.523	0.346	0.306
OLS	0.692	0.524	0.347	0.305
EXP	0.688	0.519	0.340	0.304
maxGini	0.666	0.483	0.199	0.151
maxGini2	0.657	0.495	0.303	0.269
maxGiniAB	0.692	0.523	0.346	0.306

Tabulka 3: Analýza dat „Bankloan“.

Název	Gini	KS	R^2	R^2_{McFadden}
ML	0.883	0.743	0.604	0.547
OLS	0.881	0.736	0.608	0.536
EXP	0.882	0.743	0.596	0.543
maxGini	0.873	0.733	0.564	0.504
maxGini2	-0.107	0.203	0.011	0.007
maxGiniAB	0.883	0.741	0.604	0.546

Tabulka 4: Analýza dat „Churn“.