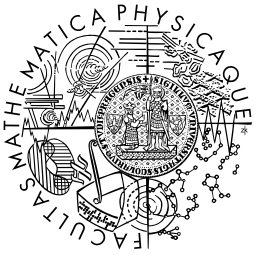


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

DIPLOMOVÁ PRÁCE

Bc. Patrik Janáček

Robustní regrese a robustní neuronové sítě

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: RNDr. Jan Kalina, Ph.D.

Studijní program: Pravděpodobnost, matematická statistika a ekonometrie

Studijní obor: Pravděpodobnost, matematická statistika a ekonometrie

Praha 2023

Prohlášení

Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Tato práce nebyla využita k získání jiného nebo stejného titulu.

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 má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V Praze dne 5. ledna 2023 .....

Poděkování

Na tomto místě bych rád poděkoval svému vedoucímu, RNDr. Janu Kalinovi, Ph.D., za obětavou pomoc a připomínky, které mi během psaní této práce poskytl.

Název práce: Robustní regrese a robustní neuronové sítě

Autor: Bc. Patrik Janáček

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: RNDr. Jan Kalina, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Klasická metoda nejmenších čtverců v lineární regresi je náchylná na přítomnost odlehlých hodnot v datech. Cílem této práce je představit několik robustních alternativ metody nejmenších čtverců v rámci lineární regrese a diskutovat jejich vlastnosti. Následuje představení robustních neuronových sítí inspirovaných těmito odhady, které jsou porovnány v rámci simulační studie. Slibnou se jeví zejména metoda nejmenších vážených čtverců v kombinaci s adaptivními váhami, která je schopna kombinovat vysokou robustnost s efektivitou při absenci kontaminace v datech.

Klíčová slova: Robustnost, regrese, strojové učení, neuronové sítě

Title: Robust regression and robust neural networks

Author: Bc. Patrik Janáček

Department: Department of Probability and Mathematical Statistics

Supervisor: RNDr. Jan Kalina, Ph.D., Department of Probability and Mathematical Statistics

Abstract: The classical least squares approach in linear regression is prone to the presence of outliers in the data. The aim of this thesis is to present several robust alternatives to the least squares method in the linear regression framework and discuss their properties. Then robust neural networks based on these estimators are introduced and compared in a simulation study. In particular, the least weighted squares method with adaptive weights seems promising, as it is able to combine high robustness with efficiency in the absence of contamination in the data.

Keywords: Robustness, regression, machine learning, neural networks

Obsah

Značení

Tato sekce shrnuje značení používané v této diplomové práci.

Čísla, vektory a matice

a	Skalár
$\boldsymbol{a}$	Sloupcový vektor
$\mathbb{A}$	Matice
$\mathsf{diag}(\boldsymbol{a})$	Diagonální matice obsahující složky vektoru $\boldsymbol{a}$ na diagonále
$\mathsf{trace}\,\mathbb{A}$	Stopa matice $\mathbb{A}$
$\mathsf{det}\,\mathbb{A}$	Determinant matice $\mathbb{A}$
$\mathsf{rank}\,\mathbb{A}$	Hodnost matice $\mathbb{A}$
$\mathsf{vec}\,\mathbb{A}$	Po sloupcích vektorizovaná matice $\mathbb{A}$
$\boldsymbol{a}^\top, \mathbb{A}^\top$	Transpozice vektoru $\boldsymbol{a}$ a matice $\mathbb{A}$
$\boldsymbol{a}^{\otimes 2}$	$\boldsymbol{a}\boldsymbol{a}^\top$
$\mathbb{A}^{-1}$	Inverzní matice k matici $\mathbb{A}$
$\mathbf{0}_n$	n-složkový nulový vektor
$\mathbb{O}_{n\times m}$	Nulová matice o rozměrech $n \times m$
$\mathbf{1}_n$	n-složkový vektor jedniček
$\mathbb{I}_n$	Jednotková matice řádu n
$\|\boldsymbol{a}\|, \|\boldsymbol{a}\|_2$	ℓ_2 norma vektoru $\boldsymbol{a}$
$\|\boldsymbol{a}\|_p$	ℓ_p norma vektoru $\boldsymbol{a}$
$\|\boldsymbol{a}\|_\infty$	Maximová norma vektoru $\boldsymbol{a}$

Množiny a grafy

A	Množina
$\mathbb{N}$	Množina přirozených čísel
$\mathbb{Z}$	Množina celých čísel
$\mathbb{R}$	Množina reálných čísel
$\mathbb{R}_{\geq 0}$	Nezáporná reálná čísla
$\mathsf{card}\,A, \lvert A\rvert$	Mohutnost množiny A
$\mathcal{G} = (V, E)$	Graf s množinou vrcholů V a hran E

Funkce

$f : \mathbb{R} \to \mathbb{R}$	Skalární funkce
$f : \mathbb{R}^n \to \mathbb{R}$	Funkce více proměnných
$\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$	Vektorová funkce více proměnných
$\mathcal{C}(X, Y)$	Množina spojitých zobrazení z X do Y
$f \circ g$	Složení funkcí f a g
$\nabla f(\boldsymbol{x})$	Gradient f v bodě $\boldsymbol{x}$
$\mathbb{J}_{\boldsymbol{f}}(\boldsymbol{x})$	Jakobiho matice $\boldsymbol{f}$ v bodě $\boldsymbol{x}$
$\mathbb{1}\{\dots\}$	Indikátorová funkce
$\log(x)$	Přirozený logaritmus x
$\operatorname{sgn}(x)$	Signum x

Aplikaci skalární funkce f po složkách budeme značit $\boldsymbol{f}(\boldsymbol{x})$. Tedy zápis $\boldsymbol{a} = \boldsymbol{f}(\boldsymbol{x})$ budeme chápat jako $a_i = f(x_i)$ pro všechna i.

Pravděpodobnost a statistika

$(\Omega, \mathcal{A}, \mathsf{P})$	Pravděpodobnostní prostor s množinou jevů Ω, σ-algebrou $\mathcal{A}$ a pravděpodobnostní mírou P
$\mathcal{B}_0, \mathcal{B}_0^n$	Borelovská σ-algebra na $\mathbb{R}$, resp. $\mathbb{R}^n$
X	Náhodná veličina
$\boldsymbol{X}$	Náhodný vektor
$\xrightarrow{\mathsf{P}}$	Konvergence v pravděpodobnosti
$\xrightarrow{\mathsf{D}}$	Konvergence v distribuci
$X \sim \mathsf{L}$	X má přesné rozdělení L
$X \overset{\text{as.}}{\sim} \mathsf{L}$	X má asymptoticky rozdělení L
$\mathcal{H}_0, \mathcal{H}_1$	Nulová hypotéza, alternativa
$\mathsf{E}\, X$	Střední hodnota náhodné veličiny X
$\mathsf{med}\, X$	Medián náhodné veličiny X
$\mathsf{var}\, X$	Rozptyl náhodné veličiny X
N	Normální rozdělení
$\mathsf{Laplace}$	Laplaceovo rozdělení
Unif	Spojité rovnoměrné rozdělení
χ^2_k	χ^2 rozdělení s k stupni volnosti

Úvod

Neuronovým sítím se v poslední době dostává čím dál větší pozornosti díky jejich nejmodernějším výsledkům napříč mnoha různorodými obory. Jejich aplikace zahrnují regresní problémy včetně predikcí časových řad, rozpoznání vzorů a ručně psaného textu, detekce objektů, strojový překlad nebo například řešení parciálních diferenciálních rovnic ve fyzice.

Tato práce se zabývá regresními úlohami, ve kterých zkoumáme vztahy mezi závisle proměnnou a jednou či více nezávisle proměnnými. Klasickým přístupem je odhadovat regresní koeficienty metodou nejmenších čtverců, která je však citlivá na odlehlá, resp. vzdálená pozorování. Proto si robustní regrese klade za cíl navrhnout regresní odhady, které nejsou tak silně ovlivněny kontaminací v datech. Robustní metody jsou však často neefektivní v případě, kdy jsou všechny předpoklady splněny. Cílem práce je představit několik robustních odhadů v kontextu lineární regrese a diskutovat jejich základní vlastnosti, včetně adaptivních metod, které umí kombinovat vysokou robustnost (bod selhání) spolu s vysokou relativní eficiencí při splnění všech předpokladů.

Robustní regresní odhady jsou oblíbené téma ve statistické literatuře, ale nejsou tolik populární ve spojitosti s neuronovými sítěmi, které jsou však často trénovány na velkém množství automatizovaně získaných dat. Druhým cílem této práce je tedy ukázat, že i klasické neuronové sítě jsou náchylné na kontaminaci v trénovacích datech a při vyšší kontaminaci mohou benefitovat z představených robustních odhadů.

V první kapitole formulujeme úlohu statistického učení a představíme princip minimalizace rizika, přičemž se zaměříme na regresní problémy. Ve druhé kapitole připomeneme lineární model a metodu nejmenších čtverců spolu s kvantilovou regresí. Třetí kapitola je věnována robustním regresním odhadům a jejich teoretickým vlastnostem, zejména konzistenci, asymptotické normalitě a robustnosti ve smyslu bodu selhání. Čtvrtá kapitola představí neuronové sítě a běžně používaný heuristický přístup jejich trénování. Také představíme několik robustních neuronových sítí a upozorníme na některé praktické problémy. V poslední kapitole prezentujeme výsledky simulační studie a diskutujeme vhodnost představených robustních neuronových sítí pro různé druhy kontaminace v trénovacích datech.

1. Statistické učení

V této kapitole formulujeme úlohu statistického učení a představíme princip minimalizace rizika. Jedná se o obecný princip, který zahrnuje klasické statistické metody, jako metodu nejmenších čtverců nebo metodu maximální věrohodnosti. To je užitečné, neboť jak na regresní metody, které průběžně představíme, tak na neuronové sítě se můžeme dívat jako na minimalizaci (empirického) rizika. V našem výkladu se zaměříme na regresní problémy, kterým se budeme dále věnovat. Nastíníme, jaká úskalí s sebou nese minimalizace empirického rizika a diskutujeme, jak pomocí křížové validace zvolit optimální hyperparametry. Nakonec představíme metodu stochastického gradientu, jejíž varianty patří v současnosti mezi nejpoužívanější optimalizační algoritmy ve strojovém učení.

Problému minimalizace rizika se věnují knihy Vapnik (2000), Shalev-Shwartz a Ben-David (2014) a článek Vapnik (1991). Stručně je diskutován také v článku Bottou a kol. (2018), který se zabývá především metodou stochastického gradientu a ze kterého vychází i práce Janáček (2020). Náš výklad stochastického gradientu je doplněn knihami Goodfellow a kol. (2016) a Shalev-Shwartz a Ben-David (2014).

1.1 Model

Uvažujme posloupnost n nezávislých náhodných vektorů

$$\mathcal{D} = \left\{[Y_1, \boldsymbol{X}_1^\top]^\top, \ldots, [Y_n, \boldsymbol{X}_n^\top]^\top\right\},$$

definovaných na pravděpodobnostním prostoru $(\Omega, \mathcal{A}, \mathsf{P})$ s hodnotami v měřitelném prostoru $(\mathcal{Y} \times \mathcal{X}, \mathcal{A}_\mathcal{Y} \otimes \mathcal{A}_\mathcal{X})$, rozdělených jako obecný náhodný vektor $[Y, \boldsymbol{X}^\top]^\top$ s distribuční funkcí $F(y, \boldsymbol{x})$. Náhodnému výběru $\mathcal{D}$ říkáme **trénovací data**. Pokud je veličina Y diskrétní, mluvíme o klasifikačním problému. V této práci se budeme věnovat regresním problémům, ve kterých je veličina Y spojitá.

Cílem učení je zvolit funkci z předem stanovené rodiny měřitelných funkcí $\mathcal{F} = \{f(\boldsymbol{x}; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$, která nejlépe modeluje závislost veličiny Y na veličinách $\boldsymbol{X}$. Rodinu $\mathcal{F}$ nazýváme **modelem** (prostorem hypotéz). Funkci $f \in \mathcal{F}$ se říká **hypotéza** (predikční funkce), $\boldsymbol{\theta} \in \Theta$ je **parametr modelu**. O procesu volby parametru $\boldsymbol{\theta} \in \Theta$ mluvíme jako o **trénování modelu**.

Terminologie. V regresních úlohách nazýváme veličinu Y **odezvou** (závisle proměnnou). Předpokládáme, že náhodný vektor $\boldsymbol{X}$ má $k < n$ složek, kterým říkáme **regresory** (prediktory, vysvětlující proměnné, nezávisle proměnné). V tomto kontextu říkáme funkci $f \in \mathcal{F}$ také **regresní funkce**. Vektor odezev budeme značit $\boldsymbol{Y} = [Y_1, \ldots, Y_n]^\top$ a **regresní maticí** budeme rozumět $\mathbb{X} = [\boldsymbol{X}_1^\top, \ldots, \boldsymbol{X}_n^\top]^\top$.

Příklad. V lineární regresi se typicky snažíme na základě trénovacích dat $\mathcal{D}$ modelovat podmíněnou střední hodnotu $\mathsf{E}[Y|\boldsymbol{X}]$ pomocí lineárního modelu

$$\mathcal{F} = \{f(\boldsymbol{x}; \boldsymbol{\beta}) = \boldsymbol{x}^\top \boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^k\}.$$

Příklad. Příkladem nelineárního modelu je neuronová síť. Pro pevnou architekturu definuje model $\mathcal{F} = \{f(\boldsymbol{x}; \boldsymbol{w}) = f_K \circ \boldsymbol{f}_{K-1} \circ \cdots \circ \boldsymbol{f}_1(\boldsymbol{x}) : \boldsymbol{w} \in W\}$, kde $\boldsymbol{f}_k$ jsou nelineární funkce. V regresních úlohách $f_K(\boldsymbol{x}; \boldsymbol{w}_K) = \boldsymbol{x}^\top \boldsymbol{w}_K$.

Volba modelu

Volba modelu $\mathcal{F}$ je motivována povahou problému, který se chystáme řešit. Pokud je naším cílem predikce odezvy pro nové hodnoty regresorů, je užitečné uvažovat bohaté a flexibilní modely, neboť interpretace parametrů pro nás většinou není tak důležitá. Důležitou roli hraje validace modelu.

Často je ale naším cílem rozhodnout, které nezávisle proměnné jsou asociované s odezvou a odhadnout efekt daného prediktoru na odezvu. Potom je pro nás konkrétní tvar regresní funkce $f \in \mathcal{F}$ podstatný a parametry zájmu by měly mít přímočarou interpretaci.

1.2 Problém minimalizace rizika

Definice 1. Ztrátovou funkcí *rozumíme funkci* $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$*, která pomocí ztráty* $\ell(y, f(\boldsymbol{x}; \boldsymbol{\theta}))$ *měří, jak se predikce* $f(\boldsymbol{x}; \boldsymbol{\theta})$ *liší od skutečné odezvy* y*. Derivaci ztrátové funkce podle parametru* $\boldsymbol{\theta}$ *značíme* ψ *a nazýváme* **skórová funkce**.

Příklad. Nejpoužívanější ztrátovou funkcí v regresních úlohách je ℓ_2 ztráta

$$\ell_2(y, f(\boldsymbol{x}; \boldsymbol{\theta})) = [y - f(\boldsymbol{x}; \boldsymbol{\theta})]^2.$$

Populární je také ℓ_1 ztráta, definována vztahem

$$\ell_1(y, f(\boldsymbol{x}; \boldsymbol{\theta})) = |y - f(\boldsymbol{x}; \boldsymbol{\theta})|.$$

Poznámka. Podle kontextu budeme někdy také uvažovat ztrátovou funkci jako funkci reziduí $e(\boldsymbol{\theta}) = y - f(\boldsymbol{x}; \boldsymbol{\theta})$.

Ztrátová funkce je náhodná veličina, neboť je funkcí trénovacích dat. Předpokládáme, že je integrovatelná pro každou $f \in \mathcal{F}$.

Definice 2. Očekávané riziko $\mathsf{R} : \Theta \to \mathbb{R}_{\geq 0}$ *definujeme jako*

$$\mathsf{R}(\boldsymbol{\theta}) = \int_{\mathcal{Y} \times \mathcal{X}} \ell(y, f(\boldsymbol{x}; \boldsymbol{\theta}))\, \mathrm{d}F(y, \boldsymbol{x}) = \mathsf{E}[\ell(Y, f(\boldsymbol{X}; \boldsymbol{\theta}))].$$

Očekávané riziko posuzuje vhodnost dané hypotézy $f(\boldsymbol{x}; \boldsymbol{\theta}) \in \mathcal{F}$. Cílem je tedy najít $f(\boldsymbol{x}; \boldsymbol{\theta}_0)$, kde $\boldsymbol{\theta}_0 = \operatorname{argmin}_{\boldsymbol{\theta} \in \Theta} \mathsf{R}(\boldsymbol{\theta})$. To však není v praxi možné, neboť distribuční funkce $F(y, \boldsymbol{x})$ je neznámá a všechna dostupná informace je obsažena v trénovacích datech $\mathcal{D}$.

Příklad. V regresním případě uvažujme, že množina reálných funkcí

$$\mathcal{F} = \{f(\boldsymbol{x}; \boldsymbol{\beta}) : \boldsymbol{\beta} \in \mathbb{R}^k\}$$

obsahuje skutečnou regresní funkci

$$f(\boldsymbol{x}; \boldsymbol{\beta}_0) = \int_{\mathbb{R}} y\, \mathrm{d}F(y|\boldsymbol{x}).$$

Pak je známo, že regresní funkce minimalizuje očekávané riziko s ℓ_2 ztrátou

$$\mathsf{R}(\boldsymbol{\beta}) = \mathsf{E}[Y - f(\boldsymbol{X}; \boldsymbol{\beta})]^2.$$

Tudíž problém odhadu regresních koeficientů odpovídá problému minimalizace očekávaného rizika s ℓ_2 ztrátou v situaci, kdy distribuční funkci $F(y, \boldsymbol{x})$ neznáme, ale máme k dispozici trénovací data $\mathcal{D}$.

Pokud skutečná regresní funkce $f(\boldsymbol{x})$ neleží v $\mathcal{F}$, potom $f(\boldsymbol{x}; \boldsymbol{\beta}_0)$ minimalizující očekávané riziko s ℓ_2 ztrátovou funkcí je nejbližší k $f(\boldsymbol{x})$ ve smyslu $\mathcal{L}_2(F)$ metriky, která je definována následovně

$$\rho\Big(f(\boldsymbol{x}), f(\boldsymbol{x}; \boldsymbol{\beta}_0)\Big) = \sqrt{\int_{\mathbb{R}^k} [f(\boldsymbol{x}) - f(\boldsymbol{x}; \boldsymbol{\beta}_0)]^2 \,\mathrm{d}F(\boldsymbol{x})}.$$

Příklad. V úloze mediánové regrese nechť $\mathcal{F}$ obsahuje skutečnou regresní funkci

$$f(\boldsymbol{x}; \boldsymbol{\beta}_0) = \mathsf{med}[Y|\boldsymbol{X} = \boldsymbol{x}].$$

Potom víme, že regresní funkce minimalizuje očekávané riziko s ℓ_1 ztrátou

$$\mathsf{R}(\boldsymbol{\beta}) = \mathsf{E}\,|Y - f(\boldsymbol{X}; \boldsymbol{\beta})|.$$

Minimalizace empirického rizika

Cílem učení je tedy minimalizovat očekávané riziko $\mathsf{R}(\boldsymbol{\theta})$ v situaci, kdy distribuční funkci $F(y, \boldsymbol{x}) = F(y|\boldsymbol{x})F(\boldsymbol{x})$ neznáme. Hledáme tedy řešení problému, který zahrnuje odhad očekávaného rizika.

Definice 3. Empirickým rizikem *nazveme empirický odhad očekávaného rizika založený na trénovacích datech, tedy*

$$\mathsf{R}_{\mathcal{D}}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(\boldsymbol{X}_i; \boldsymbol{\theta})).$$

Empirické riziko je nestranným odhadem očekávaného rizika, neboť

$$\mathsf{E}\,\mathsf{R}_{\mathcal{D}}(\boldsymbol{\theta}) = \mathsf{E}\left[\frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(\boldsymbol{X}_i; \boldsymbol{\theta}))\right] = \mathsf{E}\left[\ell(Y, f(\boldsymbol{X}; \boldsymbol{\theta}))\right] = \mathsf{R}(\boldsymbol{\theta}),$$

kde jsme využili, že pozorování jsou nezávislá, stejně rozdělená.

V rámci **principu minimalizace empirického rizika** (ERM) tedy nahradíme očekávané riziko $\mathsf{R}(\boldsymbol{\theta})$ empirickým rizikem $\mathsf{R}_{\mathcal{D}}(\boldsymbol{\theta})$ a funkci $f(\boldsymbol{x}; \boldsymbol{\theta}_0)$ minimalizující $\mathsf{R}(\boldsymbol{\theta})$ aproximujeme funkcí $f(\boldsymbol{x}; \widehat{\boldsymbol{\theta}})$, kde $\widehat{\boldsymbol{\theta}} = \operatorname{argmin}_{\boldsymbol{\theta} \in \Theta} \mathsf{R}_{\mathcal{D}}(\boldsymbol{\theta})$.

Terminologie. Pro danou hypotézu $f(\boldsymbol{x}; \boldsymbol{\theta}) \in \mathcal{F}$, $\boldsymbol{\theta} \in \Theta$ pevné, budeme někdy hodnotu empirického rizika $\mathsf{R}_{\mathcal{D}}(\boldsymbol{\theta})$ nazývat **trénovací chybou** a hodnotě očekávaného rizika $\mathsf{R}(\boldsymbol{\theta})$ říkat **generalizační chyba**.

ERM je obecný princip zahrnující klasické metody, jako metodu nejmenších čtverců, metodu nejmenších absolutních odchylek nebo například metodu maximální věrohodnosti.

Příklad. V regresní úloze dostaneme po substituci ℓ_2 ztráty do empirického rizika

$$\mathsf{R}_{\mathcal{D}}(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} [Y_i - f(\boldsymbol{X}_i; \boldsymbol{\beta})]^2.$$

Minimalizace potom vede k odhadu regresních koeficientů metodou nejmenších čtverců

$$\widehat{\boldsymbol{\beta}}^{\mathsf{LS}} = \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^k} \sum_{i=1}^{n} [Y_i - f(\boldsymbol{X}_i; \boldsymbol{\beta})]^2.$$

Příklad. Podobně, substitucí ℓ_1 ztráty do empirického rizika dostáváme odhad regresních koeficientů metodou nejmenších absolutních odchylek

$$\widehat{\boldsymbol{\beta}}^{\mathsf{LAD}} = \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^k} \sum_{i=1}^{n} |Y_i - f(\boldsymbol{X}_i; \boldsymbol{\beta})|.$$

Příklad. Nevýhodou předchozích ztrátových funkcí je, že jsou náchylné na odlehlá pozorování, případně vzdálená pozorování (leverage points). V této práci se budeme zabývat robustními odhady, které se snaží odhadnout regresní koeficienty i pro vysoce kontaminovaná data. Tuto schopnost můžeme kvantifikovat například pomocí bodu selhání, jak uvidíme ve třetí kapitole, kde současně podáme formální definici. Příkladem robustního odhadu s vysokým bodem selhání je metoda nejmenších ořezaných čtverců, která hledá odhad regresních koeficientů jako řešení optimalizačního problému

$$\widehat{\boldsymbol{\beta}}^{\mathsf{LTS}} = \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^k} \sum_{i=1}^{\delta} \left(Y_i - f(\boldsymbol{X}_i; \boldsymbol{\beta})\right)^2_{(i)},$$

kde $\frac{n}{2} < \delta \leq n$. Tato metoda tedy odhaduje regresní koeficienty na základě δ trénovacích dat s nejmenším součtem čtverců. Později uvidíme, že pro vhodnou volbu parametru δ dosahuje maximálního možného bodu selhání.

Konzistentnost ERM

Princip minimalizace empirického rizika předpokládá, že proces učení vedoucí k minimalizaci empirického rizika vede k malé hodnotě očekávaného rizika (říkáme, že je schopen zobecňovat). Jinými slovy potřebujeme, aby metoda minimalizace empirického rizika byla konzistentní. Nutným a postačujícím podmínkám pro (netriviální) konzistenci ERM metody se podrobně věnuje například druhá kapitola knihy Vapnik (2000).

Meze pro schopnost generalizace

S pomocí stejnoměrného zákona velkých čísel a konceptu **kapacity** $d_{\mathcal{F}}$[1] rodiny funkcí $\mathcal{F}$[2] je možné sestrojit horní mez pro generalizační chybu procesu učení, viz například třetí kapitola knihy Vapnik (2000).

[1] Oblíbená míra kapacity (složitosti) $\mathcal{F}$ je **VC dimenze** (Vapnikova-Červoněnkisova dimenze). VC dimenze rodiny indikátorových funkcí $\mathcal{F}$ je definována jako maximální počet $d_{\mathcal{F}}$ vektorů, které můžeme separovat do dvou skupin všemi $2^{d_{\mathcal{F}}}$ způsoby použitím funkcí z $\mathcal{F}$. Následně se dá zobecnit i pro množinu reálných funkcí. Například rodina lineárních funkcí $\mathcal{F} = \{\boldsymbol{x} \mapsto \boldsymbol{a}^\top \boldsymbol{x} + b : \boldsymbol{a} \in \mathbb{R}^k, b \in \mathbb{R}\}$ má VC dimenzi $d_{\mathcal{F}} = k + 1$.

[2] Technicky bychom měli uvažovat spíše množinu ztrátových funkcí

$$\mathcal{L} = \{\ell(y, f(\boldsymbol{x}; \boldsymbol{\theta})) : \boldsymbol{\theta} \in \Theta\}.$$

Tyto meze ukazují, že pro malou generalizační chybu musíme mít kapacitu $\mathcal{F}$ pod kontrolou. Pokud je podíl $\frac{n}{d_{\mathcal{F}}}$ velký, pak malá hodnota empirického rizika garantuje malou hodnotu očekávaného rizika. Pokud je však $\frac{n}{d_{\mathcal{F}}}$ malé, pak tuto záruku nemáme. Situaci, kdy je $f \in \mathcal{F}$ příliš přizpůsobena trénovacím datům, ale nemá schopnost generalizace a selhává na datech nových, říkáme **přeučení**. Z diskuse výše také vidíme, že zvýšení rozsahu výběru n obvykle sníží riziko přeučení, neboť relativní kapacita modelu se zmenší.

Minimalizace regularizované ztráty

Jedním z oblíbených přístupů, jak minimalizovat rizikový funkcionál na základě konečných trénovacích dat a současně zabránit přeučení je **minimalizace regularizované ztráty** (RLM), kdy současně minimalizujeme empirické riziko a regularizační funkci. **Regularizační funkcí** rozumíme zobrazení $\mathcal{R} : \Theta \to \mathbb{R}$ a predikční funkci hledáme na základě

$$\widehat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \mathsf{R}_{\mathcal{D}}(\boldsymbol{\theta}) + \mathcal{R}(\boldsymbol{\theta}).$$

Regularizace se snaží zabránit přeučení tím, že měří složitost predikčních funkcí pomocí hodnoty regularizační funkce.

Příklad (ℓ_2 regularizace)**.** Uvažujme regularizační funkci založenou na ℓ_2 normě, potom princip minimalizace regularizované ztráty odpovídá optimalizačnímu problému

$$\widehat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \mathsf{R}_{\mathcal{D}}(\boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|_2^2, \quad \lambda > 0.$$

V kontextu lineární regrese dostáváme **hřebenovou regresi** (ang. ridge)

$$\widehat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^k} \|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2.$$

Efekt ℓ_2 regularizace je takový, že pro rostoucí λ se začnou odhady regresních koeficientů čím dál více blížit nule, avšak na rozdíl od ℓ_1 regularizace jsou typicky nenulové.

Příklad (ℓ_1 regularizace)**.** Pokud budeme uvažovat ℓ_1 normu, dostaneme optimalizační problém

$$\operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \mathsf{R}_{\mathcal{D}}(\boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|_1, \quad \lambda > 0.$$

V případě lineární regerese dostáváme **lasso**

$$\widehat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^k} \|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1.$$

Pro dostatečně velká λ dostáváme řídké řešení, kdy některé koeficienty mají optimální hodnotu 0. Proto se používá jako variable selection metoda. Hlavní výhodou je, že lasso umožňuje odhadovat přeparametrizované modely ($k \gg n$), nicméně počet nenulových parametrů ve výsledném modelu je nejvýše n.

Poznámka. Chování regularizace je ovlivněno volbou hyperparametru λ. S rostoucí hodnotou λ se snižuje hodnota koeficientů a tím zabraňuje přeučení. Po určité hodnotě však začne model ztrácet důležité vlastnosti, což vede k nedoučení (ang. underfitting). Později uvidíme, jak volit optimální hodnotu λ pomocí k-násobné křížové validace.

1.3 Křížová validace

Horní mez pro generalizační chybu procesu učení je teoreticky zajímavá, ale jelikož platí pro všechny funkce z $\mathcal{F}$ a všechna pravděpodobnostní rozdělení současně, může být příliš volná, pesimistická. V praxi se tedy používá zřídkakdy, neboť je typicky jednodušší odhadnout očekávané riziko pomocí **validace**, kdy rozdělíme trénovací data $\mathcal{D}$ na **trénovací** a **validační** sadu. Na trénovací sadě model natrénujeme a na validační sadě odhadneme jeho generalizační chybu.

Při trénování je naším cílem najít hypotézu $f \in \mathcal{F}$ s nejlepší generalizační schopností. Běžná praxe je porovnávat funkce z dané rodiny $\mathcal{F}$ pomocí **křížové validace**, kdy jsou trénovací data rozdělena na tři disjunktní podmnožiny. Proces nalezení predikční funkce pomocí minimalizace empirického rizika je proveden na **trénovací** sadě a cílem je vybrat malou podmnožinu $\mathcal{F}$ vhodných kandidátů. Schopnost generalizace těchto funkcí je potom ověřena na **validační** sadě, přičemž nejlepší funkce je zvolena. **Testovací** sada potom slouží k odhadnutí skutečného rizika.

k-násobná křížová validace

***k*-násobná křížová validace** se snaží odhadnout skutečné riziko a současně ušetřit co nejvíce dat pro trénování modelu. Nejprve rozdělíme trénovací data na k podmnožin $\mathcal{D}_1, \ldots, \mathcal{D}_k$ o stejné velikosti, uvažujme $\mathsf{card}\, \mathcal{D}_i = \frac{n}{k} \in \mathbb{N}$[3]. Pro každé $i \in \{1, \ldots, k\}$ trénujeme model na datech

$$\bigcup_{j \in \{1,\ldots,k\} \setminus \{i\}} \mathcal{D}_j$$

a pomocí množiny $\mathcal{D}_i$ odhadneme generalizační chybu $\widehat{\mathsf{R}}_i$. Výsledný odhad skutečného rizika dostaneme zprůměrováním

$$\widehat{\mathsf{R}} = \frac{1}{k} \sum_{i=1}^{k} \widehat{\mathsf{R}}_i.$$

k-násobná křížová validace pro volbu hyperparametrů

Trénování modelu často závisí na volbě dalších **hyperparametrů**, které budeme značit $\boldsymbol{\lambda}$. Množinu všech přípustných hodnot označme jako Λ. Příkladem může být parametr λ uvažovaný při ℓ_2 regularizaci. Jedním ze způsobů, jak zvolit optimální hodnotu $\boldsymbol{\lambda}$ je ***k*-násobná křížová validace**.

Nejprve rozdělíme trénovací data $\mathcal{D}$ na k podmnožin $\mathcal{D}_1, \ldots, \mathcal{D}_k$. Následně pro každou hodnotu $\boldsymbol{\lambda} \in \Lambda$ a pro každé $i \in \{1, \ldots, k\}$ natrénujeme model na datech $\mathcal{D} \setminus \mathcal{D}_i$, což vede k predikční funkci $f_i(\boldsymbol{\lambda})$, pro kterou odhadneme její očekávané riziko na množině $\mathcal{D}_i$, označme $\widehat{\mathsf{R}}_i(\boldsymbol{\lambda})$. Výsledný odhad generalizační chyby pro volbu $\boldsymbol{\lambda}$ dostaneme zprůměrováním

$$\widehat{\mathsf{R}}(\boldsymbol{\lambda}) = \frac{1}{k} \sum_{i=1}^{k} \widehat{\mathsf{R}}_i(\boldsymbol{\lambda}).$$

[3] Speciální případ $k = n$ se nazývá **leave-one-out**.

Dostáváme optimální volbu hyperparametrů

$$\boldsymbol{\lambda}_0 = \operatorname*{argmin}_{\boldsymbol{\lambda}\in\Lambda} \widehat{\mathsf{R}}(\boldsymbol{\lambda}).$$

1.4 Optimalizační metody

Uvažujme, že minimalizujeme diferencovatelnou účelovou funkci $F(\boldsymbol{\theta})$. V úlohách strojového učení se účelová funkce většinou rozpadne na sumu přes trénovací data, $F(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} F_i(\boldsymbol{\theta})$. Dále uvažujme počáteční volbu parametrů $\widehat{\boldsymbol{\theta}}_1 \in \Theta$ a posloupnost kladných kroků $\{\alpha_k : k \in \mathbb{N}\}$.

K minimalizaci účelové funkce $F(\boldsymbol{\theta})$ můžeme použít **metodu největšího spádu** (gradientní metoda), která je založena na pozorování, že účelová funkce klesá z daného bodu nejrychleji ve směru záporně vzatého gradientu,

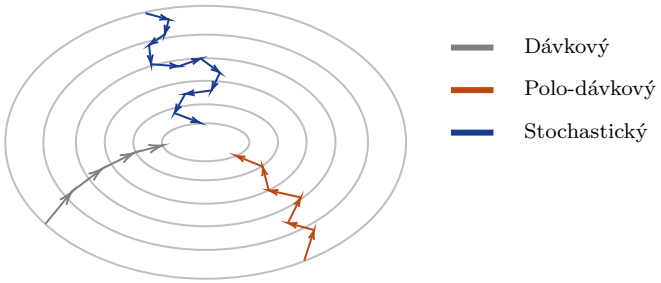
$$\widehat{\boldsymbol{\theta}}_{k+1} \leftarrow \widehat{\boldsymbol{\theta}}_k - \alpha_k \nabla F(\widehat{\boldsymbol{\theta}}_k) = \widehat{\boldsymbol{\theta}}_k - \frac{\alpha_k}{n}\sum_{i=1}^{n} \nabla F_i(\widehat{\boldsymbol{\theta}}_k), \quad k \in \mathbb{N}.$$

Optimalizačním metodám, které využívají pro výpočet všechna trénovací data, říkáme **deterministické** nebo **dávkové** (ang. batch) gradientní metody.

Optimalizační algoritmy používané ve strojovém učení většinou odhadují gradient účelové funkce pouze na základě náhodně zvolené podmnožiny m trénovacích dat,

$$\widehat{\boldsymbol{\theta}}_{k+1} \leftarrow \widehat{\boldsymbol{\theta}}_k - \frac{\alpha_k}{|I_k|}\sum_{i\in I_k} \nabla F_i(\widehat{\boldsymbol{\theta}}_k), \quad k \in \mathbb{N},$$

kde $I_k \subseteq \{1,\ldots,n\}$ jsou indexy příslušných vzorků. Parametru $m = |I_k|$, $k \in \mathbb{N}$, říkáme **batch size**. Pokud $m = 1$, mluvíme o **metodě stochastického gradientu**. V opačném případě ($1 < m < n$) se bavíme o **polo-dávkovém** (ang. mini-batch) **stochastickém gradientu**. V tomto případě je $\{\widehat{\boldsymbol{\theta}}_k\}$ stochastický proces, jehož chování je ovlivněno náhodně zvolenými vzorky v každém kroku. Každá iterace této metody je výpočetně méně náročná než v případě deterministického přístupu, avšak lze očekávat horší krok, jak ilustruje obrázek 1.1.



Obrázek 1.1. Ilustrace konvergence metody největšího spádu a metody stochastického gradientu, včetně polo-dávkové varianty.

Metoda stochastického gradientu

Metoda stochastického gradientu (SGD) a její varianty jsou nejpoužívanější optimalizační algoritmy ve strojovém učení, speciálně v hlubokém učení. Jako

účelovou funkci $F : \Theta \to \mathbb{R}$ můžeme uvažovat jak očekávané riziko, tak empirické riziko. Pokud budeme vzorky vybírat rovnoměrně z konečné trénovací sady a následně je pro každou iteraci do sady vrátíme, bude SGD minimalizovat empirické riziko $\mathsf{R}_{\mathcal{D}}(\boldsymbol{\theta})$. Pokud budeme vybírat vzorky v každé iteraci vzhledem k rozdělení $\mathsf{P}(y, \boldsymbol{x})$, budeme minimalizovat očekávané riziko $\mathsf{R}(\boldsymbol{\theta})$.

Většinou je však výhodné uvažovat více **epoch** a v každé vybírat vzorky bez vracení, dokud trénovací sadu nevyčerpáme. Potom SGD minimalizuje generalizační chybu pouze u první epochy, nicméně menší empirické riziko převáží zvýšení rozdílu mezi očekávaným a empirickým rizikem.

Algoritmus 1: Stochastický gradient.

Vstup: Posloupnost kroků $\{\alpha_k : k \in \mathbb{N}\}$.
Vstup: Počáteční volba parametru $\widehat{\boldsymbol{\theta}}_1$.
Vstup: Batch size m.
for $k = 1, 2, \ldots$ **do**
  Náhodně vyber m vzorků $[Y_i, \boldsymbol{X}_i^\top]^\top$ z trénovacích dat $\mathcal{D}$.
  Spočítej odhad gradientu $\widehat{\boldsymbol{g}}_k \leftarrow \frac{1}{m} \sum_{i=1}^{m} \nabla \ell(Y_i, f(\boldsymbol{X}_i; \widehat{\boldsymbol{\theta}}_k))$.
  Polož $\widehat{\boldsymbol{\theta}}_{k+1} \leftarrow \widehat{\boldsymbol{\theta}}_k - \alpha_k \widehat{\boldsymbol{g}}_k$.
end

Stochastický gradient vnáší do optimalizační procedury zdroj šumu (náhodný výběr m trénovacích vzorků), který nemizí ani po nalezení minima. Postačující podmínkou pro zaručení konvergence SGD je **mizející posloupnost kroků** $\{\alpha_k : k \in \mathbb{N}\}$ splňující

$$\sum_{k=1}^{\infty} \alpha_k = \infty, \quad \sum_{k=1}^{\infty} \alpha_k^2 < \infty.$$

Všimněme si, že poslední podmínka implikuje $\alpha_k \to 0$ pro $k \to \infty$. Výsledky v této oblasti shrnuje například práce Janáček (2020), která vychází z článku Bottou a kol. (2018). V praxi se posloupnost kroků často volí metodou pokus-omyl. Užitečné je při trénování sledovat křivku učení – graf účelové funkce jako funkce času.

Populární varianty stochastického gradientu, jako RMSProp nebo Adam, jsou podrobně diskutovány například v osmé kapitole knihy Goodfellow a kol. (2016).

Poznámka. Technicky vzato metoda stochastického gradientu předpokládá diferencovatelnou účelovou funkci F. SGD však můžeme aplikovat i na nediferencovatelné účelové funkce, pokud namísto gradientu budeme uvažovat subgradient funkce F v bodě $\widehat{\boldsymbol{\theta}}_k$. Připomeňme, že vektor $\boldsymbol{g}$ nazveme **subgradientem** funkce F v bodě $\widehat{\boldsymbol{\theta}}_k$, pokud pro všechny $\boldsymbol{\theta}$ platí

$$F(\boldsymbol{\theta}) \geq F(\widehat{\boldsymbol{\theta}}_k) + \boldsymbol{g}^\top (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_k).$$

Jako příklad si ukážeme, jak použít stochastický gradient pro výpočet odhadu regresních koeficientů v lineární regresi s ℓ_2 regularizací[4].

[4] Poznamenejme však, že lineární regrese s ℓ_2 regularizací má vždy explicitní řešení

$$(\mathbb{X}^\top \mathbb{X} + \lambda \mathbb{I})^{-1} \mathbb{X}^\top \boldsymbol{Y}.$$

Příklad. Účelovou funkci hřebenové regrese můžeme ekvivalentně přepsat jako

$$F(\boldsymbol{\beta}) = \frac{1}{2n}\sum_{i=1}^{n}(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta})^2 + \frac{\lambda}{2}\|\boldsymbol{\beta}\|_2^2.$$

Uvažujme batch size m a indexy $I \subseteq \{1, \ldots, n\}$, potom

$$F(\boldsymbol{\beta}) \approx \frac{1}{|I|}\sum_{i\in I}\left[\tfrac{1}{2}(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta})^2\right] + \frac{\lambda}{2}\|\boldsymbol{\beta}\|_2^2,$$

a dostáváme odhad gradientu

$$\nabla F(\boldsymbol{\beta}) \approx \frac{1}{|I|}\sum_{i\in I}\left[(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta})\boldsymbol{X}_i\right] + \lambda\boldsymbol{\beta}.$$

Algoritmus 2: Řešení hřebenové regrese pomocí SGD.

Vstup: Posloupnost kroků $\{\alpha_k : k \in \mathbb{N}\}$.
Vstup: Batch size m.
Vstup: Hyperparametr $\lambda \in \mathbb{R}$.
Inicializace: Polož $\widehat{\boldsymbol{\beta}}_1 \leftarrow \mathbf{0}$ nebo inicializuj $\widehat{\boldsymbol{\beta}}_1$ náhodně.
for *k = 1, 2, ...* **do**
  Náhodně vyber m vzorků $[Y_i, \boldsymbol{X}_i^\top]^\top$ z trénovacích dat $\mathcal{D}$.
  Spočítej odhad gradientu $\widehat{\boldsymbol{g}}_k \leftarrow \frac{1}{m}\sum_{i=1}^{m}\left[(Y_i - \boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}_k)\boldsymbol{X}_i\right] + \lambda\widehat{\boldsymbol{\beta}}_k$.
  Polož $\widehat{\boldsymbol{\beta}}_{k+1} \leftarrow \widehat{\boldsymbol{\beta}}_k - \alpha_k\widehat{\boldsymbol{g}}_k$.
end

Všimněme si, že matice $\mathbb{X}^\top\mathbb{X} + \lambda\mathbb{I}$ je vždy regulární pro $\lambda > 0$. Tedy další efekt ℓ_2 regularizace je, že inverz vždy existuje.

2. Regresní analýza

Regresní analýza je označení pro skupinu statistických metod, které modelují závislost nějaké veličiny na veličinách dalších. V této kapitole si představíme lineární regresní model a poté diskutujeme teoretické vlastnosti metody nejmenších čtverců. Těchto poznatků využijeme v dalších kapitolách pro představení robustnějších odhadů regresních koeficientů. Následně se budeme zabývat kvantilovou regresí, kde namísto podmíněné střední hodnoty modelujeme podmíněné kvantily.

Náš výklad lineární regrese založíme na knize Yan a Su (2009) a skriptech Komárek (2021), Kulich (2021), kvantilové regrese na knize Koenker (2005) a skriptech Omelka (2021).

2.1 Lineární regrese

V lineární regresi se snažíme modelovat podmíněnou střední hodnotu odezvy Y na základě předpokládaného lineárního vztahu s několika regresory $\boldsymbol{X}$. Jako v předchozí kapitole předpokládáme, že pozorujeme n nezávislých stejně rozdělených náhodných vektorů $\mathcal{D} = \{[Y_1, \boldsymbol{X}_1^\top]^\top, \dots, [Y_n, \boldsymbol{X}_n^\top]^\top\}$ s distribuční funkcí $F(y, \boldsymbol{x}) = F(y|\boldsymbol{x})F(\boldsymbol{x})$. V regresní analýze nás zajímá především podmíněné rozdělení Y při daném $\boldsymbol{X}$, zatímco marginální rozdělení $\boldsymbol{X}$ je blíže nespecifikováno. Budeme předpokládat model měření s aditivním šumem.

Definice 4. *Řekneme, že data $\mathcal{D}$ splňují* **lineární regresní model**, *pokud*

$$Y_i = \boldsymbol{X}_i^\top \boldsymbol{\beta}_0 + \varepsilon_i,$$

kde $\boldsymbol{\beta}_0$ je **vektor regresních koeficientů** *a $\varepsilon_1, \dots, \varepsilon_n$ jsou nezávislé* **chybové členy**[1], *nezávislé na regresorech $\boldsymbol{X}_i$*[2], *rozdělené jako obecná náhodná veličina ε splňující $\mathsf{E}\,\varepsilon = 0$ a $\mathsf{var}\,\varepsilon = \sigma^2$. Vektor chybových členů budeme značit $\boldsymbol{\varepsilon}$ a jejich distribuční funkci F_ε. Parametr $0 < \sigma^2 < \infty$ nazýváme* **reziduální rozptyl**.

Terminologie. Často předpokládáme $X_1 = 1$ skoro jistě. Parametr β_1 potom nazýváme **absolutním členem** a mluvíme o **lineárním modelu s absolutním členem**.

Definice 5. Hodností modelu *rozumíme číslo $r \leq k$ takové, že $\mathsf{rank}\,\mathbb{X} = r$ skoro jistě. Pokud $r = k$, mluvíme o* **modelu plné hodnosti** *(sloupce matice $\mathbb{X}$ jsou skoro jistě lineárně nezávislé v $\mathbb{R}^n$).*

Metoda nejmenších čtverců

Skutečná regresní funkce $\mathsf{E}[Y|\boldsymbol{X}]$ minimalizuje očekávané riziko s ℓ_2 ztrátou. Regresní koeficienty $\boldsymbol{\beta}_0$ můžeme tedy odhadnout minimalizací empirického rizika s touto ztrátovou funkcí.

[1] Náhodná veličina ε (které se také někdy říká šum) zachycuje všechny ostatní faktory, které ovlivňují odezvu, kromě regresorů $\boldsymbol{X}$.

[2] Obecně nemusíme uvažovat nezávislost chybových členů a regresorů, pro jednoduchost však uvažujeme tento předpoklad napříč celou prací.

Definice 6. *Odhad regresních koeficientů* **metodou nejmenších čtverců** (LS) *definujeme jako*

$$\widehat{\boldsymbol{\beta}}_n^{\mathsf{LS}} = \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^k} \mathsf{R}_{\mathcal{D}}(\boldsymbol{\beta}) = \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^k} \frac{1}{n}\sum_{i=1}^{n}(Y_i - \boldsymbol{X}_i^\top\boldsymbol{\beta})^2.$$

Pokud uvažujeme model o plné hodnosti, tak je matice $\mathbb{X}^\top\mathbb{X}$ regulární a existuje jednoznačně určené řešení $\widehat{\boldsymbol{\beta}}_n^{\mathsf{LS}} = (\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\boldsymbol{Y}$. Odhad metodou nejmenších čtverců je nestranným odhadem $\boldsymbol{\beta}_0$ a platí $\mathsf{var}\,\widehat{\boldsymbol{\beta}}_n^{\mathsf{LS}} = \sigma^2(\mathbb{X}^\top\mathbb{X})^{-1}$.

Definice 7. Vektorem vyrovnaných hodnot *(odhadnutých hodnot) rozumíme*

$$\hat{\boldsymbol{Y}} = \mathbb{X}\widehat{\boldsymbol{\beta}}_n^{\mathsf{LS}} = \mathbb{X}(\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\boldsymbol{Y}.$$

Odhad chybových členů $\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}_n^{\mathsf{LS}}$ *nazýváme* **vektorem reziduí**. *Někdy budeme uvažovat rezidua jako funkci regresních koeficientů, neboli*

$$\boldsymbol{e}(\boldsymbol{\beta}) = \boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}.$$

Inference a diagnostika lineárního regresního modelu závisí také na odhadu parametru σ^2. Nestranným odhadem reziduálního rozptylu je

$$\widehat{\sigma}_n^2 = \frac{\widehat{\boldsymbol{\varepsilon}}^\top\widehat{\boldsymbol{\varepsilon}}}{n-k} = \frac{1}{n-k}\sum_{i=1}^{n}(Y_i - \boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}_n^{\mathsf{LS}})^2.$$

Definice 8. *Řekneme, že data* $\mathcal{D}$ *splňují* **normální lineární model**, *pokud pro ně platí lineární model a chybové členy mají normální rozdělení,* $\varepsilon \sim \mathsf{N}(0, \sigma^2)$.

V normálním lineárním modelu má odhad metodou nejmenších čtverců přesné normální rozdělení,

$$\widehat{\boldsymbol{\beta}}_n^{\mathsf{LS}} \sim \mathsf{N}_k(\boldsymbol{\beta}_0,\ \sigma^2(\mathbb{X}^\top\mathbb{X})^{-1}),$$

na základě kterého lze zkonstruovat přesné testy a konfidenční intervaly. Za normality je navíc LS odhad $\widehat{\boldsymbol{\beta}}_n^{\mathsf{LS}}$ také maximálně věrohodným odhadem[3].

Odhad metodou nejmenších čtverců je konzistentní a asymptoticky normální. Statistická inference je tedy asymptoticky platná i bez splnění předpokladu normality.

Tvrzení 1. *Pokud platí lineární model a matice* $\mathbb{V} = \mathsf{E}\,\boldsymbol{X}\boldsymbol{X}^\top$ *je konečná a regulární, tak*

- $\widehat{\boldsymbol{\beta}}_n^{\mathsf{LS}} \xrightarrow{\mathsf{P}} \boldsymbol{\beta}_0$ *pro* $n \to \infty$,
- $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n^{\mathsf{LS}} - \boldsymbol{\beta}_0) \xrightarrow{D} \mathsf{N}_k(\boldsymbol{0}_k,\ \sigma^2\mathbb{V}^{-1})$ *pro* $n \to \infty$.

Poznámka. Matici $\mathbb{V}$ můžeme konzistentně odhadnout pomocí $\widehat{\mathbb{V}}_n = \frac{1}{n}\mathbb{X}^\top\mathbb{X}$.

Poznámka. Předchozí věta platí i za porušení předpokladu shody rozptylů, pokud předpokládáme, že rozptyl je funkcí regresorů, $\mathsf{var}[Y|\boldsymbol{X}] = \sigma^2(\boldsymbol{X})$.

[3] Maximálně věrohodným odhadem σ^2 je za normality $\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}_n)^2$. MLE reziduálního rozptylu je tedy vychýlený, ale asymptoticky nestranný.

2.2 Kvantilová regrese

Lineární kvantilová regrese souvisí s lineární regresí v tom smyslu, že studuje lineární vztah mezi odezvou a vysvětlujícími proměnnými. Nicméně, na rozdíl od lineární regrese, která modeluje podmíněnou střední hodnotu odezvy, kvantilová regrese modeluje podmíněný kvantil odezvy.

Nejprve připomeneme pojem kvantilu a problém hledání výběrového kvantilu formulujeme jako optimalizační úlohu. Těchto poznatků následně využijeme při definici regresních kvantilů.

Kvantily a optimalizace

Definice 9. *Pro reálnou náhodnou veličinu Z s distribuční funkcí $F_Z(z)$, $z \in \mathbb{R}$ definujeme* **kvantilovou funkci** *jako*

$$F_Z^{-1}(\tau) = \inf\{z \in \mathbb{R} : F_Z(z) \geq \tau\}, \quad \tau \in (0,1).$$

Pro pevné $\tau \in (0,1)$ nazýváme $F_Z^{-1}(\tau)$ $\boldsymbol{\tau}$**-kvantilem** *rozdělení F_Z. Speciálně, $F_Z^{-1}(\frac{1}{2})$ nazýváme* **mediánem** *rozdělení F_Z.*

Definice 10. Empirický $\boldsymbol{\tau}$-kvantil *reálné náhodné veličiny Z definujeme jako*

$$\widehat{F}_n^{-1}(\tau) = \inf\{z \in \mathbb{R} : \widehat{F}_n(z) \geq \tau\}, \quad \tau \in (0,1),$$

kde $\widehat{F}_n(z) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{Z_i \leq z\}$, $z \in \mathbb{R}$ je empirická distribuční funkce náhodného výběru $Z_1, \ldots, Z_n \sim F_Z$.

Konkrétní kvantily můžeme hledat minimalizací očekávaného rizika s vhodnou ztrátovou funkcí.

Definice 11. *Pro $\tau \in (0,1)$ definujme* **kvantilovou ztrátovou funkci** *ve tvaru*

$$\ell_\tau(u) = \tau u \mathbb{1}\{u \geq 0\} + (1-\tau)(-u)\mathbb{1}\{u < 0\}.$$

Potom je skórová funkce tvaru

$$\ell'_\tau(u) = \tau \mathbb{1}\{u > 0\} - (1-\tau)\mathbb{1}\{u < 0\}$$

pro u nenulové a dodefinujeme $\ell'_\tau(0) = 0$.

Lemma 2. *Buď Z reálná náhodná veličina s distribuční funkcí F_Z. Potom*

$$F_Z^{-1}(\tau) = \operatorname*{argmin}_{u \in \mathbb{R}} \mathsf{E}\, \ell_\tau(Z-u).$$

Důkaz. Podobně jako v Koenker (2005) využijeme derivaci integrálu podle horní meze. Z definice kvantilové ztrátové funkce

$$\begin{aligned} F_Z^{-1}(\tau) &= \operatorname*{argmin}_{u \in \mathbb{R}} \mathsf{E}\, \ell_\tau(Z-u) \\ &= \operatorname*{argmin}_{u \in \mathbb{R}} \Big\{ \mathsf{E}\left[\tau(Z-u)\mathbb{1}\{Z \geq u\} + (\tau-1)(Z-u)\mathbb{1}\{Z < u\}\right] \Big\} \\ &= \operatorname*{argmin}_{u \in \mathbb{R}} \left\{ \tau \int_u^\infty (z-u)\,\mathrm{d}F_Z(z) + (\tau-1)\int_{-\infty}^u (z-u)\,\mathrm{d}F_Z(z) \right\}. \end{aligned}$$

Derivací integrálu podle horní meze dostáváme

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}u}\int_u^\infty (z-u)\,\mathrm{d}F_Z(z) &= -\frac{\mathrm{d}}{\mathrm{d}u}\int_\infty^u (z-u)\,\mathrm{d}F_Z(z)\\ &= -\int_\infty^u \tfrac{\partial}{\partial u}(z-u)\,\mathrm{d}F_Z(z)\\ &= \int_\infty^u \mathrm{d}F_Z(z) = F_Z(u) - 1.\end{aligned}$$

Podobně

$$\frac{\mathrm{d}}{\mathrm{d}u}\int_{-\infty}^u (z-u)\,\mathrm{d}F_Z(z) = \int_{-\infty}^u \tfrac{\partial}{\partial u}(z-u)\,\mathrm{d}F_Z(z) = -F_Z(u),$$

a tedy dostáváme

$$0 = \tau F_Z(u) - \tau - \tau F_Z(u) + F_Z(u) = F_Z(u) - \tau. \tag{2.1}$$

Protože je distribuční funkce F_Z monotónní, libovolný prvek $\{z : F_Z(z) = \tau\}$ minimalizuje očekávané riziko. Pokud je řešení (2.1) jednoznačné, tak $u = F_Z^{-1}(\tau)$. V opačném případě dostáváme interval τ-kvantilů, z nichž volíme ten nejmenší. □

Nahradíme-li v předchozím lemmatu distribuční funkci F_Z pomocí empirické distribuční funkce $\widehat{F}_n$, dostaneme, že výběrový τ-kvantil můžeme hledat minimalizací empirického rizika s kvantilovou ztrátovou funkcí,

$$\widehat{F}_n^{-1}(\tau) = \operatorname*{argmin}_{u\in\mathbb{R}} \left\{ \int \ell_\tau(z-u)\,\mathrm{d}\widehat{F}_n(z) \right\} = \operatorname*{argmin}_{u\in\mathbb{R}} \left\{ \frac{1}{n}\sum_{i=1}^n \ell_\tau(Z_i - u) \right\}.$$

Pokud $\tau n \in \mathbb{N}$, dostáváme interval řešení $\{z : \widehat{F}_n(z) = \tau\}$, z nichž opět volíme to nejmenší.

V této sekci jsme převedli problém hledání výběrového τ-kvantilu na optimalizační problém, což motivuje definici regresních kvantilů.

Regresní kvantily

V lineární kvantilové regresi modelujeme podmíněný τ-kvantil odezvy na základě předpokládaného lineárního vztahu s regresory. Opět předpokládáme nezávislá, stejně rozdělená trénovací data $\mathcal{D}$ s distribuční funkcí $F_{Y,\boldsymbol{X}}$.

Definice 12. *Trénovací data splňují model* **lineární kvantilové regrese**, *pokud*

$$Y_i = \boldsymbol{X}_i^\top \boldsymbol{\beta}_0 + \varepsilon_i, \tag{2.2}$$

kde $\boldsymbol{\beta}_0 \in \mathbb{R}^k$ je vektor regresních koeficientů a $\varepsilon_1, \ldots, \varepsilon_n$ jsou nezávislé, stejně rozdělené náhodné veličiny s kvantilovou funkcí splňující $F_\varepsilon^{-1}(\tau) = 0$, $\tau \in (0,1)$. Navíc předpokládáme, že ε_i jsou nezávislé na $\boldsymbol{X}_i$.

Po vzoru předchozí sekce můžeme regresní koeficienty $\boldsymbol{\beta}_0$ odhadnout minimalizací empirického rizika s kvantilovou ztrátovou funkcí.

Definice 13. Regresní τ-kvantil *definujeme jako*

$$\widehat{\boldsymbol{\beta}}_n(\tau) = \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^k} \frac{1}{n}\sum_{i=1}^{n} \ell_\tau(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta}). \tag{2.3}$$

Poznámka. Regresní kvantil identifikuje parametr

$$\boldsymbol{\beta}(\tau) = \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^k} \mathsf{E}\, \ell_\tau(Y - \boldsymbol{X}^\top \boldsymbol{\beta}).$$

Díky lemmatu 2 platí

$$\begin{aligned}\mathsf{E}\, \ell_\tau(Y - \boldsymbol{X}^\top \boldsymbol{\beta}) &= \mathsf{E}\left\{ \mathsf{E}\left[\ell_\tau(Y - \boldsymbol{X}^\top \boldsymbol{\beta})|\boldsymbol{X}\right]\right\} \\ &\geq \mathsf{E}\left\{ \mathsf{E}\left[\ell_\tau(Y - F^{-1}_{Y|\boldsymbol{X}}(\tau))|\boldsymbol{X}\right]\right\} = \mathsf{E}\, \ell_\tau(Y - F^{-1}_{Y|\boldsymbol{X}}(\tau)).\end{aligned}$$

Pokud tedy platí model (2.2), neboli $F^{-1}_{Y|\boldsymbol{X}}(\tau) = \boldsymbol{X}^\top \boldsymbol{\beta}_0$, pak regresní τ-kvantil identifikuje správný parametr $\boldsymbol{\beta}(\tau) = \boldsymbol{\beta}_0$.

Poznámka. V uvažovaném modelu (2.2) se dva různé regresní kvantily liší pouze v absolutních členech. Tedy efekty prediktorů jsou stejné pro všechny kvantily odezvy. Regresní kvantily jsou však motivovány zejména v obecnějších situacích, kdy efekt prediktoru může být jiný pro různé kvantily odezvy.

Poznámka. Pro regresní kvantily nemáme analytické vyjádření jako v případě metody nejmenších čtverců. Minimalizační úloha (2.3) je úlohou lineárního programování a tedy je možné použít řešení založená na simplexové metodě.

Asymptotické vlastnosti regresních kvantilů shrnuje následující věta.

Věta 3. *Nechť platí následující předpoklady.*

(i) *Distribuční funkce F_ε je absolutně spojitá se spojitou hustotou f_ε splňující* $0 < f_\varepsilon(0) < \infty$.

(ii) *Matice $\mathbb{V} = \mathsf{E}\, \boldsymbol{X}\boldsymbol{X}^\top$ je pozitivně definitní.*

(iii) $\max_{i\leq n} \|\boldsymbol{X}_i\|_\infty = o_{\mathsf{P}}(n^{1/2})$, $n \to \infty$.

Potom $\widehat{\boldsymbol{\beta}}_n(\tau)$ je slabě konzistentním odhadem $\boldsymbol{\beta}(\tau)$ a platí

$$\sqrt{n}\big(\widehat{\boldsymbol{\beta}}_n(\tau) - \boldsymbol{\beta}(\tau)\big) \xrightarrow{D} \mathsf{N}_k\left(\mathbf{0}_k,\ \frac{\tau(1-\tau)}{f_\varepsilon^2(0)}\mathbb{V}^{-1}\right),\ n \to \infty.$$

Důkaz. V obecnější podobě je věta dokázána v Koenker (2005, věta 4.1). Pollard (1991) dokázal tvrzení pro náhodné regresory a $\tau = \frac{1}{2}$.

□

Poznámka. Slabá konzistence platí i za slabších předpokladů, viz například Koenker (2005), kapitola 4.1.2.

Poznámka. Matici $\mathbb{V}$ můžeme konzistentně odhadnout pomocí $\widehat{\mathbb{V}}_n = \frac{1}{n}\mathbb{X}^\top\mathbb{X}$. Pro odhad f_ε lze použít například jádrový odhad hustoty. Nabízí se také použití neparametrického bootstrapu.

3. Robustní regrese

Statistickou proceduru často nazýváme robustní, pokud je platná i při nesplněných předpokladech, za kterých byla odvozena. Naproti tomu, v této práci budeme v kontextu regrese nazývat **robustní procedurou** metodu, která není příliš ovlivněna **odlehlými pozorováními**, případně **vzdálenými pozorováními** (leverage points), tedy pozorováními s neobvyklou hodnotou odezvy, resp. neobvyklými hodnotami regresorů. Taková pozorování mohou být chybnými měřeními, mohou být naměřeny za výjimečných okolností nebo patřit do jiné populace. Pozorování tohoto druhu mohou zásadním způsobem ovlivnit výsledný model a je tedy důležité odlehlá pozorování rozpoznat.

V praxi se obvykle snažíme odlehlá pozorování detekovat na základě **diagnostických nástrojů** založených na klasických metodách, jako je metoda nejmenších čtverců. Nicméně, odlehlá pozorování mohou klasické metody ovlivnit natolik, že výsledný model neumožňuje odlehlá pozorování dobře rozpoznat (masking effect). Navíc, některá správná pozorování mohou být klasifikována jako odlehlá (swamping). **Robustní regrese** si naopak klade za cíl navrhnout regresní odhady, které nejsou tak silně ovlivněny odlehlými hodnotami. Ty následně můžeme identifikovat jako pozorování, která neodpovídají nalezenému robustnímu fitu. Protože odezva může být měřena v libovolných jednotkách, i-té pozorování většinou považujeme za odlehlé, pokud je $|\widehat{\varepsilon}_i/\widehat{\sigma}_n|$ velké, kde $\widehat{\sigma}_n$ je robustní odhad reziduální směrodatné odchylky.

Všimněme si, že diagnostika i robustní regrese mají stejné cíle, jen v opačném pořadí. Regresní diagnostika se snaží vypořádat s outliery a následně použít metodu nejmenších čtverců. Robustní regrese se nejprve snaží proložit majoritu dat a poté identifikovat odlehlá pozorování, následuje jejich studium.

V této kapitole zavedeme bod selhání a představíme několik robustních odhadů v kontextu lineární regrese. Motivace vychází z knihy Rousseeuw a Leroy (1987) a článků Hubert a kol. (2008), Rousseeuw a Hubert (2011). Výklad M-odhadů vychází ze skript Omelka (2021) a knihy Maronna a kol. (2006). Nejmenším ořezaným čtvercům se věnuje kniha Rousseeuw a Leroy (1987) a články Rousseeuw a Driessen (2006), Víšek (2006a), Víšek (2006b), Kalina (2015). U metody nejmenších vážených čtverců vycházíme z Čížek (2007), Kalina (2012), Kalina a Tichavský (2020), Víšek (2011) a Víšek (2002). Teorii k nejmenším adaptivně váženým čtvercům čerpáme především z článků Čížek (2011) a Čížek (2007).

3.1 Bod selhání

Užitečnou mírou robustnosti statistické metody je bod selhání, který zavedeme v kontextu lineární regrese. Předpokládejme tedy, že máme k dispozici nezávislá, stejně rozdělená trénovací data $\mathcal{D}$ o konečném rozsahu výběru n. Označme jako $\mathcal{D}_m$ kontaminovaná trénovací data, která získáme nahrazením m pozorování v $\mathcal{D}$ pomocí libovolných hodnot.

Definice 14. Bod selhání *odhadu regresních koeficientů* $\widehat{\boldsymbol{\beta}}_n\{\mathcal{D}\}$ *v lineární regresi s konečným rozsahem výběru* n *definujeme jako*

$$\varepsilon_n^*(\widehat{\boldsymbol{\beta}}_n) = \min_{m \geq 1} \left\{ \frac{m}{n} : \sup_{\mathcal{D}_m} \left\| \widehat{\boldsymbol{\beta}}_n\{\mathcal{D}_m\} - \widehat{\boldsymbol{\beta}}_n\{\mathcal{D}\} \right\| = \infty \right\}.$$

Intuitivně se jedná o nejmenší podíl $\frac{m}{n}$ kontaminace, který způsobí, že se odhad stane úplně nespolehlivým a neinformativním. Nejlepší bod selhání, kterého můžeme dosáhnout je $\frac{1}{2}$, neboť při vyšší kontaminaci již nejsme schopni rozlišit, která část pozorování je ta správná (viz větu 4).

Definice 15. Asymptotický bod selhání *odhadu* $\widehat{\boldsymbol{\beta}}_n\{\mathcal{D}\}$ *potom definujeme jako*

$$\varepsilon^*(\widehat{\boldsymbol{\beta}}_n) = \lim_{n \to \infty} \varepsilon_n^*(\widehat{\boldsymbol{\beta}}_n), \text{ pokud limita existuje.}$$

Poznámka. Podobně můžeme definovat i (asymptotický) bod selhání odhadu reziduální směrodatné odchylky $\widehat{\sigma}_n\{\mathcal{D}\}$, jen navíc požadujeme, aby byl odhad za kontaminace zdola omezený nulou.

Příklad. Jediné odlehlé pozorování může libovolně vychýlit odhad metodou nejmenších čtverců, tedy $\varepsilon_n^* = \frac{1}{n}$ a asymptotický bod selhání LS odhadu je 0 %.

Věta 4. *Pro libovolný odhad regresních koeficientů* $\widehat{\boldsymbol{\beta}}_n$*, který je ekvivariantní vzhledem k regresi*[1]*, platí*

$$\varepsilon_n^*(\widehat{\boldsymbol{\beta}}_n) \leq \big(\lfloor (n-k)/2 \rfloor + 1\big)/n.$$

Důkaz. Rousseeuw a Leroy (1987), kapitola 3, věta 4. □

Definice 16. *Řekneme, že pozorování jsou* **skoro jistě v obecné poloze**, *pokud libovolných* $k+1$ *bodů neleží na nadrovině skoro jistě. Tato podmínka automaticky platí, pokud data pochází ze spojitého rozdělení.*

3.2 Metoda nejmenších absolutních odchylek

Metoda nejmenších absolutních odchylek je robustnější alternativou k metodě nejmenších čtverců, která modeluje podmíněný medián odezvy. Předpokládáme, že pro trénovací data $\mathcal{D}$ platí model (2.2) s chybovými členy splňujícími $\mathsf{med}\,\varepsilon = 0$. Skutečná regresní funkce $\mathsf{med}[Y|\boldsymbol{X}]$ minimalizuje očekávané riziko s ℓ_1 ztrátou[2], což motivuje následující definici.

Definice 17. *Odhad regresních koeficientů* **metodou nejmenších absolutních odchylek** (LAD) *definujeme jako*

$$\widehat{\boldsymbol{\beta}}_n^{\mathsf{LAD}} = \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^k} \frac{1}{n} \sum_{i=1}^{n} |Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta}|.$$

[1] Viz definice 27 v dodatku.
[2] Proto mluvíme také o $\boldsymbol{L_1}$**-odhadu**.

Terminologie. Metoda nejmenších absolutních odchylek je speciálním případem kvantilové regrese pro $\tau = \frac{1}{2}$. Někdy se bavíme o **mediánové regresi** a odhad nazýváme **regresním mediánem**.

Poznámka. Regresní medián identifikuje parametr

$$\boldsymbol{\beta}^{\mathsf{LAD}} = \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^k} \mathsf{E}\,|Y - \boldsymbol{X}^\top\boldsymbol{\beta}|.$$

Pokud tedy platí uvažovaný model, neboli $\mathsf{med}[Y|\boldsymbol{X}] = \boldsymbol{X}^\top\boldsymbol{\beta}_0$, pak metoda identifikuje skutečný parametr $\boldsymbol{\beta}^{\mathsf{LAD}} = \boldsymbol{\beta}_0$.

Pozorování 5. *L_1-odhad je ekvivariantní vzhledem k regresi, měřítku i afinní transformaci*[3]*.*

Důkaz. Nahlédne se podobně, jako později v pozorování 13 pro LWS odhad. □

Tvrzení 6. *Uvažujme lineární regresní model a navíc předpokládejme, že chybové členy mají Laplaceovo rozdělení. Pak* LAD *odhad regresních koeficientů je současně maximálně věrohodným odhadem.*

Důkaz. Tvrzení dokážeme. Protože $Y|\boldsymbol{X} \sim \mathsf{Laplace}(\boldsymbol{X}^\top\boldsymbol{\beta}, \sigma^2)$, je podmíněná hustota tvaru

$$f_{Y|\boldsymbol{X}}(y|\boldsymbol{x}) = \frac{1}{2\sigma^2}\exp\Big\{-\frac{|y-\boldsymbol{x}^\top\boldsymbol{\beta}|}{\sigma^2}\Big\}, \quad y\in\mathbb{R}, \boldsymbol{x}\in\mathcal{X}.$$

Fixujme σ^2 a uvažujme věrohodnost pro parametr $\boldsymbol{\beta}$

$$L(\boldsymbol{\beta}) = (2\sigma^2)^{-n}\exp\Big\{-\frac{1}{\sigma^2}\sum_{i=1}^n |Y_i - \boldsymbol{X}_i^\top\boldsymbol{\beta}|\Big\}\prod_{i=1}^n f_{\boldsymbol{X}}(\boldsymbol{X}_i),$$

tedy logaritmická věrohodnost je tvaru

$$\log L(\boldsymbol{\beta}) = -n\log(2\sigma^2) - \frac{1}{\sigma^2}\sum_{i=1}^n |Y_i - \boldsymbol{X}_i^\top\boldsymbol{\beta}| + \sum_{i=1}^n \log f_{\boldsymbol{X}}(\boldsymbol{X}_i).$$

Maximalizace log-věrohodnosti tedy odpovídá optimalizačnímu problému

$$\widehat{\boldsymbol{\beta}}_n^{\mathsf{ML}} = \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^k}\sum_{i=1}^n |Y_i - \boldsymbol{X}_i^\top\boldsymbol{\beta}|,$$

neboť ostatní členy $\log L(\boldsymbol{\beta})$ nezávisí na parametru $\boldsymbol{\beta}$. □

Věta 7. *Nechť platí následující předpoklady.*

(i) Distribuční funkce F_ε je absolutně spojitá se spojitou hustotou f_ε splňující $0 < f_\varepsilon(0) < \infty$.

[3] Viz příslušné definice v dodatku A.1.

(ii) *Matice* $\mathbb{V} = \mathsf{E}\,\boldsymbol{X}\boldsymbol{X}^\top$ *je pozitivně definitní.*

(iii) $\max_{i\le n} \|\boldsymbol{X}_i\|_\infty = o_{\mathsf{P}}(n^{1/2}),\ n \to \infty.$

Potom $\widehat{\boldsymbol{\beta}}_n^{\mathsf{LAD}}$ *je slabě konzistentním odhadem* $\boldsymbol{\beta}^{\mathsf{LAD}}$ *a platí*

$$\sqrt{n}\big(\widehat{\boldsymbol{\beta}}_n^{\mathsf{LAD}} - \boldsymbol{\beta}^{\mathsf{LAD}}\big) \xrightarrow{D} \mathsf{N}_k\left(\mathbf{0}_k,\ \frac{1}{4f_\varepsilon^2(0)}\mathbb{V}^{-1}\right),\ n \to \infty.$$

Důkaz. Jedná se o důsledek věty 3. V této podobě větu dokázal Pollard (1991). □

Pozorování 8. *Metoda nejmenších absolutních odchylek je robustní vzhledem k odlehlým pozorováním, avšak je náchylná na vzdálená pozorování. Její asymptotický bod selhání je tedy* 0 %.

3.3 M-odhady, Huberova regrese

Jedním z prvních kroků směrem k robustnějším odhadům byly M-odhady. Spočívají v minimalizaci empirického rizika pro nějakou pomaleji rostoucí ztrátovou funkci, než je ℓ_2 ztráta. Vhodná ztrátová funkce by měla splňovat $\ell(e) \ge 0$, $\ell(0) = 0$, $\ell(e) = \ell(-e)$, $\ell(e_1) \ge \ell(e_2)$ pro $|e_1| \ge |e_2|$ a být diferencovatelná.

Definice 18. M-odhad *regresních koeficientů se ztrátovou funkcí* $\ell : \mathbb{R} \to \mathbb{R}_{\ge 0}$ *definujeme jako*

$$\widehat{\boldsymbol{\beta}}_n^{\mathsf{M}} = \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^k} \frac{1}{n}\sum_{i=1}^n \ell(Y_i - \boldsymbol{X}_i^\top\boldsymbol{\beta}). \tag{3.1}$$

Je-li ztrátová funkce diferencovatelná s derivací ψ*, odhad regresních koeficientů můžeme hledat jako řešení následující soustavy* k *rovnic*

$$\frac{1}{n}\sum_{i=1}^n \psi(Y_i - \boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}_n^{\mathsf{M}})\boldsymbol{X}_i = \mathbf{0}_k. \tag{3.2}$$

Poznámka. Pokud je ztrátová funkce konvexní, jsou řešení (3.1) a (3.2) ekvivalentní. V opačném případě může být volba optimálního řešení komplikovaná.

Příklad (Huberova regrese)**. Huberův odhad** regresních koeficientů je kompromisem mezi metodou nejmenších čtverců a metodou nejmenších absolutních odchylek. **Huberova ztrátová funkce** je tvaru

$$\ell_{\mathsf{H}}(e) = \begin{cases} \frac{e^2}{2}, & \text{pokud } |e| \le \delta, \\ \delta(|e| - \frac{\delta}{2}), & \text{pokud } |e| > \delta, \end{cases}$$

kde $\delta \in \mathbb{R}$ je zvolená konstanta, často volíme například $\delta = 1.345$. Tedy skórová funkce má tvar

$$\psi_{\mathsf{H}}(e) = \begin{cases} e, & \text{pokud } |e| \le \delta, \\ \delta\,\mathrm{sgn}(e), & \text{pokud } |e| > \delta. \end{cases}$$

Grafy těchto funkcí jsou k dispozici na obrázku 3.1. Huberův odhad regresních koeficientů potom definujeme jako M-odhad s Huberovou ztrátovou funkcí. Obecně je těžké říci, co je modelováno, jedná se o něco mezi $\mathsf{E}[Y|\boldsymbol{X}]$ a $\mathsf{med}[Y|\boldsymbol{X}]$.

Výpočet

Uvažujme M-odhad se skórovou funkcí ψ diferencovatelnou v 0 a definujme váhy $\widehat{w}_i = \psi(\widehat{\varepsilon}_i)/\widehat{\varepsilon}_i$, kde $\widehat{\varepsilon}_i = Y_i - \boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}_n^{\mathsf{M}}$, resp. $\widehat{w}_i = \psi'(0)$ pokud $\widehat{\varepsilon}_i$ je rovno nule. Položme $\hat{\mathbb{W}} = \mathsf{diag}(\widehat{w}_1, \ldots, \widehat{w}_n)$, potom můžeme pro odhadovací rovnice psát

$$\sum_{i=1}^{n} \psi(Y_i - \boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}_n^{\mathsf{M}})\boldsymbol{X}_i = \sum_{i=1}^{n} \widehat{w}_i(Y_i - \boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}_n^{\mathsf{M}})\boldsymbol{X}_i = \mathbf{0}_k.$$

V maticovém zápisu $\mathbb{X}^\top \hat{\mathbb{W}} \mathbb{X} \widehat{\boldsymbol{\beta}}_n^{\mathsf{M}} = \mathbb{X}^\top \hat{\mathbb{W}} \boldsymbol{Y}$ a tedy $\widehat{\boldsymbol{\beta}}_n^{\mathsf{M}} = (\mathbb{X}^\top \hat{\mathbb{W}} \mathbb{X})^{-1} \mathbb{X}^\top \hat{\mathbb{W}} \boldsymbol{Y}$.

Protože váhy závisí na reziduích, která závisí na odhadnutých koeficientech, které zase závisí na váhách, nemůžeme odhad spočítat přímo, ale potřebujeme iterativní řešení – **iterativně vážené nejmenší čtverce** (IWLS), které můžeme popsat následujícím pseudokódem.

Algoritmus 3: IWLS pro výpočet M-odhadů.

Inicializace: Zvol počáteční odhad $\widehat{\boldsymbol{\beta}}_n^{(1)}$, například LAD.
for *k = 1, 2, ...* **do**
  Spočítej rezidua $\widehat{\varepsilon}_i^{(k)} \leftarrow Y_i - \boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}_n^{(k)}$ a váhy $\widehat{w}_i^{(k)}$ jako výše.
  Polož $\hat{\mathbb{W}}^{(k)} \leftarrow \mathsf{diag}(\widehat{w}_1^{(k)}, \ldots, \widehat{w}_n^{(k)})$.
  Spočítej $\widehat{\boldsymbol{\beta}}_n^{(k+1)} \leftarrow (\mathbb{X}^\top \hat{\mathbb{W}}^{(k)} \mathbb{X})^{-1} \mathbb{X}^\top \hat{\mathbb{W}}^{(k)} \boldsymbol{Y}$.
end

Iterujeme do konvergence odhadu, tedy například dokud $\|\widehat{\boldsymbol{\beta}}_n^{(k+1)} - \widehat{\boldsymbol{\beta}}_n^{(k)}\| < \kappa$, kde κ je předem specifikovaná tolerance.

Příklad (Huberova regrese, pokr.)**.** V případě Huberovy regrese dostáváme váhy

$$\widehat{w}_i = \frac{\psi_{\mathsf{H}}(\widehat{\varepsilon}_i)}{\widehat{\varepsilon}_i} = \begin{cases} 1, & \text{pokud } |\widehat{\varepsilon}_i| \le \delta, \\ \delta \operatorname{sgn}(\widehat{\varepsilon}_i)/\widehat{\varepsilon}_i, & \text{pokud } |\widehat{\varepsilon}_i| > \delta. \end{cases}$$

Studentizace

Řešení (3.2) nemusí být ekvivariantní vzhledem k měřítku[4], tedy musíme standardizovat rezidua vhodným odhadem σ a řešit

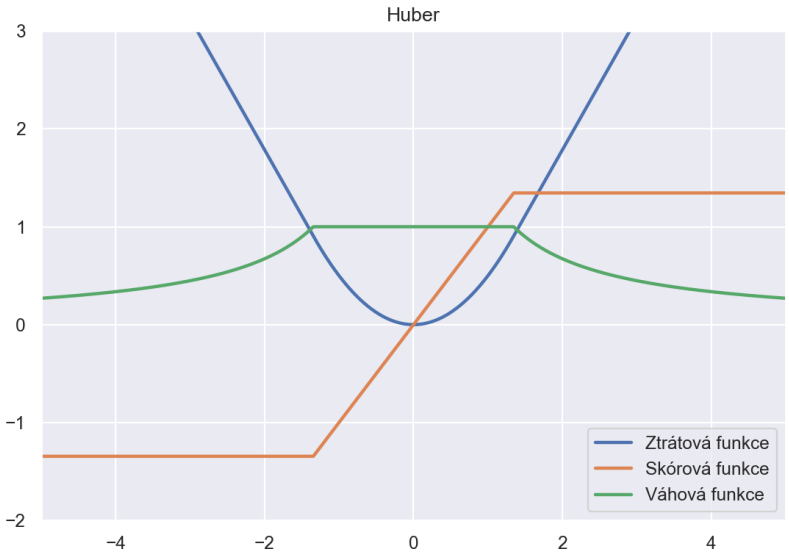
$$\frac{1}{n}\sum_{i=1}^{n} \psi\left(\frac{Y_i - \boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}_n^{\mathsf{M}}}{\widehat{\sigma}_n}\right)\boldsymbol{X}_i = \mathbf{0}_k. \tag{3.3}$$

Parametr měřítka σ můžeme robustně odhadnout například pomocí[5]

$$\mathsf{MAD} = c \cdot \mathsf{med}_{1\le i\le n} |e_i - \mathsf{med}_{1\le i\le n} e_i|,$$

kde e_i jsou rezidua získaná L_1-regresí.

[4] Viz definice 28 v dodatku.
[5] Konstanta c závisí na rozdělení chybových členů. Pro normální rozdělení $c = 1.483$.



Obrázek 3.1. Grafy ztrátové, skórové a váhové funkce pro Huberovu regresi s volbou $\delta = 1.345$. Na horizontální ose jsou hodnoty x a na vertikální ose hodnoty příslušných funkcí $f(x)$.

Vlastnosti

Za určitých předpokladů regularity (Huber a Ronchetti (2009), sekce 6.3) se dá ukázat, že M-odhad regresních koeficientů $\widehat{\boldsymbol{\beta}}_n^{\mathsf{M}}$ (3.3) v lineárním modelu (definice 4) je slabě konzistentním odhadem $\boldsymbol{\beta}_0$ a platí

$$\sqrt{n}\big(\widehat{\boldsymbol{\beta}}_n^{\mathsf{M}} - \boldsymbol{\beta}_0\big) \xrightarrow{\mathsf{D}} \mathsf{N}_k\left(\mathbf{0}_k,\ \sigma^2 \frac{\mathsf{E}\,\psi(\varepsilon/\sigma)^2}{[\mathsf{E}\,\psi'(\varepsilon/\sigma)]^2}[\mathsf{E}\,\boldsymbol{X}\boldsymbol{X}^\top]^{-1}\right),\ n \to \infty,$$

viz například kapitolu 5.4.3 v knize Maronna a kol. (2006).

M-odhady jsou robustní vůči odlehlým pozorováním, nicméně díky jejich náchylnosti na vzdálená pozorování je jejich bod selhání roven $\frac{1}{n}$, tedy asymptotický bod selhání je stále 0 %. Jako možné řešení byly představeny **zobecněné M-odhady** (GM-odhady), které se snaží omezit vliv vzdáleného pozorování pomocí váhové funkce w a hledají odhad jako řešení

$$\sum_{i=1}^{n} w(\boldsymbol{X}_i)\psi\left(\frac{Y_i - \boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}_n^{\mathsf{GM}}}{w(\boldsymbol{X}_i)\widehat{\sigma}_n}\right)\boldsymbol{X}_i = \mathbf{0}_k.$$

3.4 Nejmenší ořezané čtverce

Nadále budeme uvažovat lineární model, ale tentokrát v obecnější podobě.

Definice 19. *Řekneme, že obecná data* $(\boldsymbol{Y}, \mathbb{X})$ *splňují* **lineární model**, *pokud*

$$\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon},$$

kde $\boldsymbol{\beta}_0$ *je vektor skutečných regresních koeficientů a* $\boldsymbol{\varepsilon}$ *je vektor chybových členů splňujících* $\mathsf{E}\,\boldsymbol{\varepsilon} = \mathbf{0}_n$ *a* $\mathsf{var}\,\boldsymbol{\varepsilon} = \sigma^2\mathbb{I}_n$.

Uvažujme nyní rezidua jako funkci regresních koeficientů, $\boldsymbol{e}(\boldsymbol{\beta}) = \boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}$. Jako $e^2_{(i)}(\boldsymbol{\beta})$ budeme značit i-tou nejmenších hodnotu mezi čtverci reziduí, neboli

$$0 \leq e^2_{(1)}(\boldsymbol{\beta}) \leq \cdots \leq e^2_{(n)}(\boldsymbol{\beta}).$$

Definice 20. *Odhad regresních koeficientů* **metodou nejmenších ořezaných čtverců** (LTS) *definujeme jako*

$$\widehat{\boldsymbol{\beta}}^{\mathsf{LTS}}_n = \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^k} \sum_{i=1}^{\delta} e^2_{(i)}(\boldsymbol{\beta}) = \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^k} \sum_{i=1}^{n} e^2_i(\boldsymbol{\beta}) \mathbb{1}\{e^2_i(\boldsymbol{\beta}) \leq e^2_{(\delta)}(\boldsymbol{\beta})\},$$

kde $\frac{n}{2} < \delta \leq n$ *kontroluje robustnost (bod selhání), neboť z definice vyplývá, že* $n - \delta$ *pozorování s největšími čtverci reziduí nemá vliv na* LTS *odhad.*

Poznámka. Problém z definice je ekvivalentní nalezení δ-podmnožiny trénovacích dat s nejmenším součtem čtverců. LTS odhad je potom odhadem metodou nejmenších čtverců na základě těchto δ pozorování. Vidíme tedy, že pro $\delta = n$ dostáváme metodu nejmenších čtverců. Naopak nejmenší ořezané čtverce jsou speciálním případem metody nejmenších vážených čtverců, které se budeme věnovat později.

Poznámka. Volba δ přirozeně závisí na rozsahu výběru n. Proto když zkoumáme asymptotické vlastnosti LTS odhadu, fixujeme $\frac{1}{2} \leq \alpha \leq 1$ a pro dané n položíme $\delta_n = \lfloor \alpha n \rfloor$.

Poznámka. Při praktické aplikaci LTS odhadu se nabízí otázka, jak volit hodnotu parametru δ. Můžeme využít dostupnou apriorní informaci, nebo například křížovou validaci, viz sekci 1.3. Subjektivní způsob volby δ diskutuje také Kalina (2015). Alternativou může být volba δ na základě nějakého robustního počátečního odhadu, jak uvidíme později v kapitole 3.6.

Poznámka. Rousseeuw a Driessen (2006) navrhli rychlý aproximativní algoritmus pro výpočet LTS odhadu, vhodný pro data o velkém rozsahu. Pro malé výběry typicky nachází přesné řešení. Tento algoritmus je implementován například ve funkci `ltsReg` ve výpočetním prostředí `R`.

Pozorování 9. *Odhad metodou nejmenších ořezaných čtverců je ekvivariantní vzhledem k regresi, měřítku i afinní transformaci.*

Věta 10. *Uvažujme nezávislá, stejně rozdělená trénovací data* $\mathcal{D}$*, která jsou skoro jistě v obecné poloze a* $\delta = \lfloor n/2 \rfloor + \lfloor (k+1)/2 \rfloor$*. Potom*

$$\varepsilon^*_n\left(\widehat{\boldsymbol{\beta}}^{\mathsf{LTS}}_n\right) = \frac{\lfloor (n-k)/2 \rfloor + 1}{n} \to \frac{1}{2}, \; n \to \infty.$$

Důkaz. Rousseeuw a Leroy (1987), věta 6.

□

Poznámka. Když používáme LTS regresi, můžeme reziduální rozptyl odhadnout pomocí

$$\widehat{\sigma}^2_n = \frac{\gamma}{\delta} \sum_{i=1}^{\delta} (Y_i - \boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}^{\mathsf{LTS}}_n)^2_{(i)},$$

kde γ je konstanta, díky které je odhad konzistentní za normálně rozdělených náhodných chyb.

Asymptotické vlastnosti

Nejprve zjemníme předpoklady nutné pro odvození asymptotických výsledků, kdy budeme předpokládat jistou formu závislosti v datech. Následně formulujeme konzistenci a asymptotickou normalitu pro odhad regresních koeficientů LTS metodou.

Značení. Distribuční funkce, resp. hustoty (pokud existují) chybových členů ε_i a ε_i^2 budeme značit F_ε a G_ε, resp. f_ε a g_ε. Příslušné kvantilové funkce budeme značit F_ε^{-1} a G_ε^{-1}.

Pozorování 11. *Jelikož G_ε je distribuční funkce ε_i^2, dostáváme*

$$G_\varepsilon(e^2) = \{F_\varepsilon(e) - F_\varepsilon(-e)\}\mathbb{1}\{e^2 > 0\}.$$

Tudíž, je-li F_ε absolutně spojitá, je také G_ε absolutně spojitá s hustotou

$$g_\varepsilon(e^2) = \frac{f_\varepsilon(e) + f_\varepsilon(-e)}{2e}\mathbb{1}\{e^2 > 0\}.$$

Předpoklady, za kterých formulujeme asymptotické výsledky se opírají o pojem absolutně regulární posloupnosti, viz dodatek A.3, především definice 33.

Předpoklad 1.

(i) Náhodné vektory $\{[\boldsymbol{X}_i^\top, \varepsilon_i]^\top : i \in \mathbb{N}\}$ tvoří slabě stacionární absolutně regulární posloupnost s koeficienty $\{\beta_m\}$ splňujícími

$$m^{r/(r-2)}(\log m)^{2(r-1)/(r-2)}\beta_m \to 0,\ m \to \infty$$

pro nějaké $r > 2$. Dále nechť mají tyto náhodné vektory konečné r-té momenty. Navíc, matice $\mathbb{D} = \mathsf{E}\,\boldsymbol{X}_i\boldsymbol{X}_i^\top$ je regulární a

$$n^{-1/4}\max_{1\le i\le n}\|\boldsymbol{X}_i\|_\infty = \mathcal{O}_\mathsf{P}(1).$$

(ii) Nechť $\{\varepsilon_i : i \in \mathbb{N}\}$ je posloupnost symetricky a stejně rozdělených náhodných veličin s konečnými druhými momenty, $\mathsf{E}\,\varepsilon_i = 0$ a $\mathsf{var}\,\varepsilon_i = \sigma^2$. Dále buď ε_i a $\boldsymbol{X}_i$ nezávislé. Distribuční funkce F_ε náhodných veličin ε_i je absolutně spojitá s kladnou, omezenou a spojitě diferencovatelnou hustotou na jejím nosiči.

Poznámka. První část jsou standardní předpoklady (stejnoměrné) centrální limitní věty. Pro nezávislé a stejně rozdělené $[\boldsymbol{X}_i^\top, \varepsilon_i]$ je postačující existence druhých momentů. Druhá část jsou klasické předpoklady na chybové členy ε_i a jejich rozdělení. Veličiny ε_i a $\boldsymbol{X}_i$ nemusí být nutně nezávislé, ale ε_i podmíněně na $\boldsymbol{X}_i$ musí být symetricky rozdělené.

Věta 12. *Nechť platí obecný lineární model a předpoklad 1. Uvažujme* LTS *odhad regresních koeficientů takový, že $\frac{\delta_n}{n} \to \alpha$ pro $n \to \infty$, $\frac{1}{2} \le \alpha \le 1$.*

Potom $\widehat{\boldsymbol{\beta}}_n^{\mathsf{LTS}}$ je slabě konzistentním odhadem $\boldsymbol{\beta}_0$ a platí

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n^{\mathsf{LTS}} - \boldsymbol{\beta}_0\right) \xrightarrow{D} \mathsf{N}_k\left(\mathbf{0}_k,\ \frac{\mathbb{D}^{-1}\int_{-u_\alpha}^{u_\alpha}\varepsilon^2\,\mathrm{d}F_\varepsilon(\varepsilon)}{[2u_\alpha f_\varepsilon(u_\alpha) - \alpha]^2}\right)$$

pro $n \to \infty$, kde $u_\alpha^2 = G_\varepsilon^{-1}(\alpha)$, pokud je jmenovatel kladný.

Důkaz. Větu ukážeme s využitím věty 14, proto čtenáři doporučujeme se k důkazu vrátit až po přečtení následující podkapitoly.

Uvažujme váhovou funkci $\psi(t) = \mathbb{1}\{t \leq \alpha\},\ t \in [0,1]$, která zřejmě splňuje předpoklad 2. Potom podle věty 14 tvrzení platí a asymptotická varianční matice je tvaru

$$\mathbb{V} = \frac{\mathbb{D}^{-1} \operatorname{var}\left[\boldsymbol{X}_1 \varepsilon_1 \psi\left\{G_\varepsilon(\varepsilon_1^2)\right\}\right] \mathbb{D}^{-1}}{\left[\int \varepsilon \psi\left\{G_\varepsilon(\varepsilon^2)\right\} f'_\varepsilon(\varepsilon)\,\mathrm{d}\varepsilon\right]^2},$$

pokud je jmenovatel kladný. Nejprve si všimněme, že

$$\mathbb{1}\{G_\varepsilon(\varepsilon_1^2) \leq \alpha\} = \mathbb{1}\{-u_\alpha \leq \varepsilon_1 \leq u_\alpha\} =: \mathbb{1}_\alpha.$$

Tedy můžeme psát

$$\operatorname{var}\left[\boldsymbol{X}_1 \varepsilon_1 \mathbb{1}_\alpha\right] = \mathsf{E}\left[\boldsymbol{X}_1 \varepsilon_1 \mathbb{1}_\alpha\right]^{\otimes 2} - \left[\mathsf{E}\,\boldsymbol{X}_1 \varepsilon_1 \mathbb{1}_\alpha\right]^{\otimes 2}.$$

Nyní využijeme nezávislost $\boldsymbol{X}_1$ a ε_1. Pro první člen dostáváme

$$\mathsf{E}\left[\boldsymbol{X}_1 \varepsilon_1 \mathbb{1}_\alpha\right]^{\otimes 2} = \mathsf{E}\left[\boldsymbol{X}_1 \boldsymbol{X}_1^\top\right] \mathsf{E}\left[\varepsilon_1^2 \mathbb{1}_\alpha\right] = \mathbb{D} \int_{-u_\alpha}^{u_\alpha} \varepsilon^2\,\mathrm{d}F_\varepsilon(\varepsilon).$$

Podobně pro druhý člen platí

$$\left[\mathsf{E}\,\boldsymbol{X}_1 \varepsilon_1 \mathbb{1}_\alpha\right]^{\otimes 2} = \left[\mathsf{E}\,\boldsymbol{X}_1\right]^{\otimes 2} \left[\mathsf{E}\,\varepsilon_1 \mathbb{1}\{-u_\alpha \leq \varepsilon_1 \leq u_\alpha\}\right]^2 = \mathbf{0},$$

neboť podle předpokladu 1 mají $\boldsymbol{X}_i$ konečné druhé momenty a ε_i jsou symetricky rozdělené. Zbývá spočítat integrál ve jmenovateli. Integrací per partes dostáváme

$$\begin{aligned}
\int_{-u_\alpha}^{u_\alpha} \varepsilon f'_\varepsilon(\varepsilon)\,\mathrm{d}\varepsilon &= \left[\varepsilon f_\varepsilon(\varepsilon)\right]_{-u_\alpha}^{u_\alpha} - \int_{-u_\alpha}^{u_\alpha} f_\varepsilon(\varepsilon)\,\mathrm{d}\varepsilon \\
&= \left[u_\alpha f_\varepsilon(u_\alpha) + u_\alpha f_\varepsilon(-u_\alpha)\right] - \left[F_\varepsilon(u_\alpha) - F_\varepsilon(-u_\alpha)\right] \\
&= 2u_\alpha f_\varepsilon(u_\alpha) - \alpha,
\end{aligned}$$

kde jsme využili pozorování 11 a že hustota f_ε je podle předpokladu 1 symetrická. □

Odhad asymptotické varianční matice najdeme později pro obecnou váhovou funkci, viz větu 15.

Poznámka. Jako speciální případ dostáváme zobecnění tvrzení 1 pro metodu nejmenších čtverců. V tomto případě je jmenovatel asymptotické varianční matice roven jedné.

3.5 Nejmenší vážené čtverce

Nyní představíme zobecnění nejmenších ořezaných čtverců pomocí implicitního vážení, kdy potenciálním odlehlým pozorováním přiřadíme menší váhy.

Definice 21. *Nerostoucí deterministickou funkci* $\psi : [0,1] \to \mathbb{R}_{\geq 0}$ *budeme nazývat* **váhovou funkcí**.

Definice 22. *Odhad regresních koeficientů* **metodou nejmenších vážených čtverců** (LWS) *s váhovou funkcí ψ definujeme jako*

$$\widehat{\boldsymbol{\beta}}_n^{\mathsf{LWS}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^k}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \psi\Big(\tfrac{2i-1}{2n}\Big)\Big(Y_i - \boldsymbol{X}_i^\top\boldsymbol{\beta}\Big)^2_{(i)}.$$

Na hodnoty $w_i = \psi\Big(\frac{2i-1}{2n}\Big)$, $i \in \{1,\dots,n\}$ se můžeme dívat jako na váhy. Dostáváme alternativní definici LWS odhadu.

Definice 23. *Uvažujme nerostoucí posloupnost nezáporných vah $w_1,\dots,w_n$. Potom můžeme definovat* LWS *odhad jako*

$$\widehat{\boldsymbol{\beta}}_n^{\mathsf{LWS}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^k}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} w_i \left(Y_i - \boldsymbol{X}_i^\top\boldsymbol{\beta}\right)^2_{(i)}.$$

Poznámka. Metoda nejmenších čtverců a metoda nejmenších ořezaných čtverců jsou speciálními případy LWS pro volby váhových funkcí $\psi(t) = 1$, respektive $\psi(t) = \mathbb{1}\{t \le \delta/n\}$, $t \in [0,1]$. Pro dosažení asymptotického bodu selhání $\frac{1}{2}$ však musíme zanedbat stejný počet pozorování jako v případě LTS[6] a položit $\psi(t) = 0$ pro $t > \frac{1}{2}$.

Poznámka. Zásadním rozdílem mezi LWS a WLS (vážené nejmenší čtverce) je, že váhy jsou přiřazovány pořádkovým statistikám čtverců reziduí, namísto jednotlivým čtvercům reziduí.

Příklad. Nyní představíme několik možných váhových funkcí, jejichž grafy jsou k dispozici na obrázku 3.2.

(i) **Lineární váhy:** $\psi(t) = 1 - t$, $t \in [0,1]$.

(ii) **Logistické váhy:** $\psi(t) = (1+\exp\{-s/2\})/(1+\exp\{s(t-1/2)\})$. V této práci budeme uvažovat hodnotu parametru $s = 10$, která je kompromisem mezi tím, že všechny pozorování dostanou velké váhy (malé s) a tím, že jsou přiřazeny velmi malé váhy velkým reziduím (velké s).

(iii) **Ořezané lineární váhy:** $\psi(t) = (1 - \frac{t}{\alpha})\mathbb{1}\{t \le \alpha\}$, $t \in [0,1]$ pro pevné $\frac{1}{2} \le \alpha < 1$. Význam parametru α je takový, že $\lfloor \alpha n \rfloor$ pozorování je zachováno a ostatní jsou ignorovány.

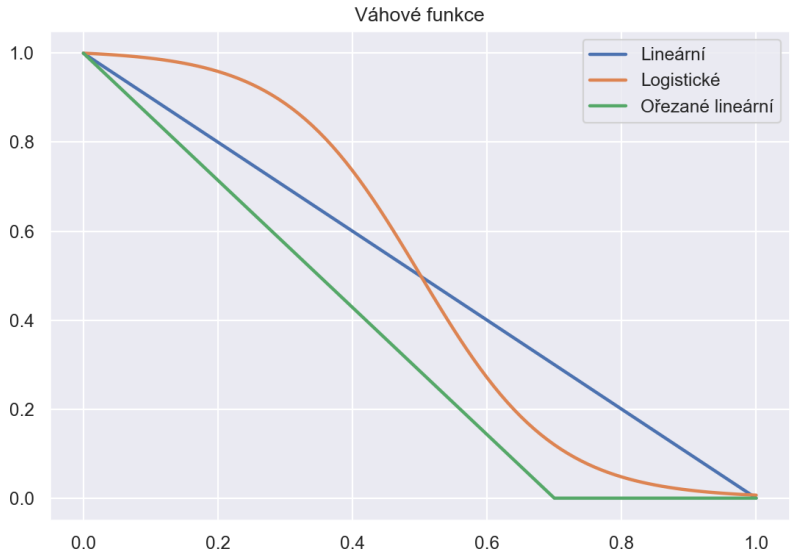
Poznámka. Rychlý algoritmus pro výpočet LWS odhadu získáme váženou analogií FAST-LTS algoritmu (Rousseeuw a Driessen, 2006).

Pozorování 13. *Odhad metodou nejmenších vážených čtverců je ekvivariantní vzhledem k regresi, měřítku i afinní transformaci.*

Důkaz. Pozorování ukážeme, pro další odhady by se postupovalo analogicky. Uvažujme LWS odhad jako v definici 23, potom

$$\sum_{i=1}^{n} w_i\Big((Y_i + \boldsymbol{X}_i^\top\boldsymbol{a}) - (\boldsymbol{X}_i^\top(\boldsymbol{\beta}+\boldsymbol{a}))\Big)^2_{(i)} = \sum_{i=1}^{n} w_i\Big(Y_i - \boldsymbol{X}_i^\top\boldsymbol{\beta}\Big)^2_{(i)},$$

[6]Tento fakt spolu s faktem, že LTS ani LWS neumí kombinovat vysoký bod selhání a asymptotickou eficienci, motivuje použití adaptivních vah, které představíme v další sekci.



Obrázek 3.2. Ilustrace představených váhových funkcí. Na horizontální ose jsou hodnoty $t \in [0, 1]$ a na vertikální ose hodnoty příslušných váhových funkcí $\psi(t)$.

odkud plyne ekvivariance vzhledem k regresi a podobně vzhledem k měřítku,

$$\sum_{i=1}^{n} w_i \Big(aY_i - a\boldsymbol{X}_i^\top \boldsymbol{\beta}\Big)^2_{(i)} = a^2 \sum_{i=1}^{n} w_i \Big(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta}\Big)^2_{(i)}.$$

Pro ekvivarianci vzhledem k afinní transformaci stačí nahlédnout, že

$$\sum_{i=1}^{n} w_i \Big(Y_i - (\boldsymbol{X}_i^\top \mathbb{A})(\mathbb{A}^{-1}\boldsymbol{\beta})\Big)^2_{(i)} = \sum_{i=1}^{n} w_i \Big(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta}\Big)^2_{(i)}.$$

□

Poznámka. Podobně jako u LTS můžeme reziduální rozptyl odhadnout pomocí

$$\widehat{\sigma}_n^2 = \frac{1}{n\gamma} \sum_{i=1}^{n} \psi\Big(\tfrac{2i-1}{2n}\Big) \Big(Y_i - \boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}_n^{\mathsf{LWS}}\Big)^2_{(i)},$$

kde γ je konstanta, díky které je odhad konzistentní za daného rozdělení[7].

Asymptotické vlastnosti

Asymptotické vlastnosti LWS odhadu formulujeme za stejných předpokladů jako pro LTS odhad, budeme se tedy držet značení zavedeného v předchozí sekci. Nejprve však musíme formulovat předpoklady kladené na váhovou funkci.

Předpoklad 2. *Váhová funkce $\psi : [0, 1] \to \mathbb{R}_{\geq 0}$ je nezáporná, omezená, zleva spojitá funkce s omezenou derivací skoro všude, tedy až na konečnou množinu $D = \{d_1, \ldots, d_J\}$ bodů nespojitosti. Tedy existuje rozklad $\psi = \psi_s + \psi_c$, kde ψ_s je schodovitá funkce (konečná lineární kombinace indikátorových funkcí intervalů) a ψ_c je spojitá, diferencovatelná funkce.*

[7] Viz například Víšek (2010).

Věta 14. *Nechť platí obecný lineární model, předpoklad 1 a předpoklad 2 s váhovou funkcí ψ. Potom $\widehat{\boldsymbol{\beta}}_n^{\mathsf{LWS}}$ je slabě konzistentním odhadem $\boldsymbol{\beta}_0$ a platí*

$$\sqrt{n}\big(\widehat{\boldsymbol{\beta}}_n^{\mathsf{LWS}} - \boldsymbol{\beta}_0\big) \xrightarrow{D} \mathsf{N}_k\left(\mathbf{0}_k,\ \frac{\mathbb{D}^{-1}\,\mathsf{var}\left[\boldsymbol{X}_1\varepsilon_1\psi\Big\{G_\varepsilon(\varepsilon_1^2)\Big\}\right]\mathbb{D}^{-1}}{\Big[\int \varepsilon\psi\Big\{G_\varepsilon(\varepsilon^2)\Big\}f_\varepsilon'(\varepsilon)\,\mathrm{d}\varepsilon\Big]^2}\right)$$

pro $n \to \infty$, pokud je jmenovatel kladný.

Důkaz. Čížek (2007), věta 5.1.

□

Nyní představíme konzistentní odhad asymptotické varianční matice z předchozí věty.

Značení. Připomeňme naše značení $\widehat{\varepsilon}_i = e_i(\widehat{\boldsymbol{\beta}}_n^{\mathsf{LWS}})$. V kontextu následující věty budeme raději využívat značení $\widehat{\varepsilon}_{n,i}$, aby bylo zřejmé, že $\{\widehat{\varepsilon}_{n,i}\}$ je trojúhelníkové schéma náhodných veličin.

Dále nechť $\hat{q}_{n,j}^2 = \hat{\varepsilon}_{(d_j n)}^2$ značí d_j-tý empirický kvantil čtverců reziduí $\{\hat{\varepsilon}_{n,i}^2\}_{i=1}^n$ pro $j \in \{1, \ldots, J\}$ a položme $d_{J+1} = 1$.

Věta 15. *Nechť platí předpoklad 1 a předpoklad 2 s rozkladem $\psi = \psi_s + \psi_c$ takovým, že schodovitá funkce splňuje $\psi_s(1) = 0$. Potom*

$$\widehat{\mathbb{D}}_n^{-1}\widehat{\mathbb{V}}_n\widehat{\mathbb{D}}_n^{-1}/\widehat{\gamma}_n^2$$

je slabě konzistentní odhad asymptotické varianční matice z věty 14, kde

- $\widehat{\mathbb{D}}_n = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^\top$,
- $\widehat{\mathbb{V}}_n = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^\top\hat{\varepsilon}_{n,i}^2\psi^2\Big\{\widehat{G}_n(\hat{\varepsilon}_{n,i}^2) - \frac{1}{2n}\Big\}$,
- $\widehat{\gamma}_n = \widehat{\gamma}_n^s + \widehat{\gamma}_n^c$, *kde*

$$\widehat{\gamma}_n^s = \sum_{j=1}^J \Big\{\psi_s(d_j) - \psi_s(d_{j+1})\Big\}\Big\{d_j - 2\hat{q}_{n,j}^2\widehat{g}_n(\hat{q}_{n,j}^2)\Big\},$$
$$\widehat{\gamma}_n^c = \frac{1}{n}\sum_{i=1}^n \psi_c\Big\{\widehat{G}_n(\hat{\varepsilon}_{n,i}^2) - \frac{1}{2n}\Big\} + \frac{2}{n}\sum_{i=1}^n \hat{\varepsilon}_{n,i}^2\psi_c'\Big\{\widehat{G}_n(\hat{\varepsilon}_{n,i}^2) - \frac{1}{2n}\Big\}\widehat{g}_n(\hat{\varepsilon}_{n,i}^2),$$

- *$\widehat{G}_n$ značí stejnoměrně konzistentní odhad distribuční funkce G_ε,*
- *$\widehat{g}_n$ je stejnoměrně konzistentní odhad hustoty g_ε.*

Důkaz. Podrobněji rozepíšeme některé kroky důkazu z článku Čížek (2011). Základním nástrojem, který budeme využívat je zákon velkých čísel pro $\mathcal{L}^1$-mixingaly (věta 26, viz dodatek A.4) a jeho varianta pro trojúhelníkové schéma náhodných veličin.

(i) Nejprve ukážeme, že

$$\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^\top \xrightarrow{\mathsf{P}} \mathsf{E}\,\boldsymbol{X}_1\boldsymbol{X}_1^\top \text{ pro } n \to \infty.$$

Podívejme se na prvky (j,l) matic výše

$$\frac{1}{n}\sum_{i=1}^{n} X_{ij}X_{il} \xrightarrow{\mathsf{P}} \mathsf{E}\, X_{1j}X_{1l} \text{ pro } n\to\infty.$$

Označme $Z_{ijl} = X_{ij}X_{il}$, $i\in\mathbb{N}$, pak $\{Z_{ijl} - \mathsf{E}\, Z_{ijl}\}$ je podle předpokladu 1 α-mixing posloupnost s konečnými r-tými momenty, $r>2$, neboť se jedná o měřitelnou funkci α-mixing posloupnosti (viz tvrzení 25). Podle příkladu z dodatku A.4 se jedná o stejnoměrně integrovatelný $\mathcal{L}^1$-mixingal. Aplikací zákona velkých čísel pro $\mathcal{L}^1$-mixingaly dostáváme

$$\frac{1}{n}\sum_{i=1}^{n} Z_{ijl} \xrightarrow{\mathsf{P}} 0 \;\text{ a tedy }\; \frac{1}{n}\sum_{i=1}^{n} X_{ij}X_{il} \xrightarrow{\mathsf{P}} \mathsf{E}\, X_{1j}X_{1l},\; n\to\infty.$$

(ii) Nyní potřebujeme ukázat

$$\widehat{\mathbb{V}}_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i\boldsymbol{X}_i^\top \hat{\varepsilon}_{n,i}^2 \psi^2\Big\{\widehat{G}_n(\hat{\varepsilon}_{n,i}^2) - \tfrac{1}{2n}\Big\} \xrightarrow{\mathsf{P}} \mathsf{var}\Big[\boldsymbol{X}_1\varepsilon_1\psi\Big\{G_\varepsilon(\varepsilon_1^2)\Big\}\Big].$$

Základní myšlenkou je, že $\Big\{\boldsymbol{X}_i\hat{\varepsilon}_{n,i}\psi\{\widehat{G}_n(\hat{\varepsilon}_{n,i}^2)\}\Big\}_{i=1}^n$ formuje posloupnost martingalových diferencí s konečnými r-tými momenty, $r>2$. Tedy podle příkladu z dodatku A.4 se jedná o stejnoměrně integrovatelný $\mathcal{L}^1$-mixingal. Aplikací zákona velkých čísel pro trojúhelníkové schéma náhodných veličin a lemmatu A.1 (Čížek, 2004) dostaneme

$$\frac{1}{n}\sum_{i=1}^{n} X_{ij}X_{il}\hat{\varepsilon}_{n,i}^2\psi^2\Big\{\widehat{G}_n(\hat{\varepsilon}_{n,i}^2) - \tfrac{1}{2n}\Big\} \xrightarrow{\mathsf{P}} \mathsf{E}\Big[X_{1j}X_{1l}\varepsilon_1^2\psi^2\Big\{G_\varepsilon(\varepsilon_1^2)\Big\}\Big] \text{ pro } n\to\infty.$$

Pro podrobnosti viz důkaz věty 5.4 a věty 5.1, speciálně rozklad (23) – (24), v článku Čížek (2007).

(iii) Zbývá odhadnout jmenovatel variační matice,

$$-\int \varepsilon\psi\Big\{G_\varepsilon(\varepsilon^2)\Big\}f_\varepsilon'(\varepsilon)\,\mathrm{d}\varepsilon = -\int \varepsilon\Big[\psi_s\Big\{G_\varepsilon(\varepsilon^2)\Big\} + \psi_c\Big\{G_\varepsilon(\varepsilon^2)\Big\}\Big]f_\varepsilon'(\varepsilon)\,\mathrm{d}\varepsilon = \gamma^s + \gamma^c.$$

Nejprve se zaměříme na odhad členu γ^c. Integrací per partes dostáváme

$$\gamma^c = \int \psi_c\Big\{G_\varepsilon(\varepsilon^2)\Big\}f_\varepsilon(\varepsilon)\,\mathrm{d}\varepsilon + \int 2\varepsilon^2\psi_c'\Big\{G_\varepsilon(\varepsilon^2)\Big\}g_\varepsilon(\varepsilon^2)f_\varepsilon(\varepsilon)\,\mathrm{d}\varepsilon \tag{3.4}$$

$$= \mathsf{E}\Big[\psi_c\Big\{G_\varepsilon(\varepsilon_1^2)\Big\}\Big] + 2\,\mathsf{E}\Big[\varepsilon_1^2\psi_c'\Big\{G_\varepsilon(\varepsilon_1^2)\Big\}g_\varepsilon(\varepsilon_1^2)\Big], \tag{3.5}$$

neboť ε_i mají konečné druhé momenty (předpoklad 1) a funkce ψ_c je omezená, spojitě diferencovatelná (předpoklad 2). Při integraci per partes (3.4) dostaneme navíc následující člen, který ukážeme, že je roven nule,

$$-\Big[\varepsilon\psi_c\Big\{G_\varepsilon(\varepsilon^2)\Big\}f_\varepsilon(\varepsilon)\Big]_{-a}^{a} = -2a\psi_c\Big\{G_\varepsilon(a^2)\Big\}f_\varepsilon(a).$$

Budeme uvažovat rozdělení s nosičem $\mathbb{R}$. Podle předpokladu 1 je hustota f_ε spojitá, omezená a kladná na $\mathbb{R}$, tedy $f_\varepsilon(\varepsilon)\to 0$, $\varepsilon\to\infty$. Pro dostatečně velká ε je $f_\varepsilon(\varepsilon)$ nerostoucí, odkud

$$\int_{\varepsilon/2}^{\infty} f_\varepsilon(u)\,\mathrm{d}u \geq \int_{\varepsilon/2}^{\varepsilon} f_\varepsilon(u)\,\mathrm{d}u \geq \int_{\varepsilon/2}^{\varepsilon} f_\varepsilon(\varepsilon)\,\mathrm{d}u = \tfrac{1}{2}\varepsilon f_\varepsilon(\varepsilon) \geq 0,$$

kde integrál na levé straně konverguje k nule pro $\varepsilon \to \infty$. Proto i $\varepsilon f_\varepsilon(\varepsilon)$ pro $\varepsilon \to \infty$. Podle předpokladu 2 je váhová funkce ψ_c omezená, odkud dostáváme

$$-2a\psi_c\Big\{G_\varepsilon(a^2)\Big\}f_\varepsilon(a) \to 0, \quad a \to \infty.$$

(iv) Nyní konzistentně odhadneme první střední hodnotu v (3.5). Máme k dispozici $\widehat{G}_n$ a $\widehat{g}_n$, stejnoměrně konzistentní odhady G_ε, resp g_ε. Z definice

$$\forall \delta > 0\ \forall \varepsilon > 0\ \exists n_0 \in \mathbb{N}\ \forall n \geq n_0 : \quad \mathsf{P}\Big[\sup_z |\widehat{g}_n(z) - g_\varepsilon(z)| < \delta\Big] > 1 - \varepsilon$$

a podobně

$$\forall \delta > 0\ \forall \varepsilon > 0\ \exists n_0 \in \mathbb{N}\ \forall n \geq n_0 : \quad \mathsf{P}\Big[\sup_z |\widehat{G}_n(z) - G_\varepsilon(z)| + \tfrac{1}{2n} < \delta\Big] > 1 - \varepsilon.$$

Funkce ψ_c je spojitá na uzavřeném intervalu $[0,1]$, odkud plyne její stejnoměrná spojitost, neboli

$$\forall \varepsilon > 0\ \exists \delta > 0\ \forall t, t' \in [0,1] : \quad \sup_{|t-t'|<\delta} |\psi_c(t) - \psi_c(t')| < \varepsilon.$$

Proto můžeme pro $n \to \infty$ psát

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \psi_c\Big\{\widehat{G}_n(\hat{\varepsilon}_{n,i}^2) - \tfrac{1}{2n}\Big\} &= \frac{1}{n}\sum_{i=1}^n \psi_c\Big\{G_\varepsilon(\hat{\varepsilon}_{n,i}^2)\Big\} \\
&\quad + \frac{1}{n}\sum_{i=1}^n \Big[\psi_c\Big\{\widehat{G}_n(\hat{\varepsilon}_{n,i}^2) - \tfrac{1}{2n}\Big\} - \psi_c\Big\{G_\varepsilon(\hat{\varepsilon}_{n,i}^2)\Big\}\Big] \\
&= \frac{1}{n}\sum_{i=1}^n \psi_c\Big\{G_\varepsilon(\hat{\varepsilon}_{n,i}^2)\Big\} + o_{\mathsf{P}}(1).
\end{aligned}$$

Protože $\hat{\varepsilon}_{n,i} \xrightarrow{\mathsf{P}} \varepsilon_i$ pro $n \to \infty$ z věty 14 a funkce ψ_c a G_ε jsou spojité, aplikací zákona velkých čísel pro trojúhelníkové schéma náhodných veličin dostáváme

$$\frac{1}{n}\sum_{i=1}^n \psi_c\Big\{G_\varepsilon(\hat{\varepsilon}_{n,i}^2)\Big\} \xrightarrow{\mathsf{P}} \mathsf{E}\Big[\psi_c\Big\{G_\varepsilon(\varepsilon_1^2)\Big\}\Big],\ n \to \infty.$$

Celkem

$$\frac{1}{n}\sum_{i=1}^n \psi_c\Big\{\widehat{G}_n(\hat{\varepsilon}_{n,i}^2) - \tfrac{1}{2n}\Big\} \xrightarrow{\mathsf{P}} \mathsf{E}\Big[\psi_c\Big\{G_\varepsilon(\varepsilon_1^2)\Big\}\Big],\ n \to \infty,$$

což jsme chtěli ukázat. Druhá střední hodnota v (3.5) by se ukázala obdobně.

(v) Nakonec najdeme konzistentní odhad γ^s. Pro $j = 1, \ldots, J$ budeme značit $q_j^2 = G_\varepsilon^{-1}(d_j)$. Podle předpokladu 2 je ψ_s schodovitá, zleva spojitá funkce se skoky v bodech $\{d_1, \ldots, d_J\}$ a případně v $d_{J+1} = 1$, která navíc splňuje $\psi_s(1) = 0$. Tedy můžeme psát

$$\psi_s(t) = \sum_{j=1}^J \Big[\psi_s(d_j) - \psi_s(d_{j+1})\Big]\mathbb{1}\{t \leq d_j\}. \tag{3.6}$$

Počítejme

$$\begin{aligned}
\int \varepsilon\psi_s\Big\{G_\varepsilon(\varepsilon^2)\Big\}f_\varepsilon'(\varepsilon)\,\mathrm{d}\varepsilon &= \sum_{i=1}^J \Big[\psi_s(d_j) - \psi_s(d_{j+1})\Big]\int_{-q_j}^{q_j} \varepsilon f_\varepsilon'(\varepsilon)\,\mathrm{d}\varepsilon \\
&= \sum_{j=1}^J \Big[\psi_s(d_j) - \psi_s(d_{j+1})\Big]\Big\{\Big[\varepsilon f_\varepsilon(\varepsilon)\Big]_{-q_j}^{q_j} - \int_{-q_j}^{q_j} f_\varepsilon(\varepsilon)\,\mathrm{d}\varepsilon\Big\},
\end{aligned}$$

kde jsme postupně využili (3.6) a integraci per partes. Nyní

$$\begin{aligned}\Big[\varepsilon f_\varepsilon(\varepsilon)\Big]_{-q_j}^{q_j} - \int_{-q_j}^{q_j} f_\varepsilon(\varepsilon)\,\mathrm{d}\varepsilon &= q_j\Big\{f_\varepsilon(q_j) + f_\varepsilon(-q_j)\Big\} - \Big\{F_\varepsilon(q_j) - F_\varepsilon(-q_j)\Big\} \\ &= 2q_j^2 g_\varepsilon(q_j^2) - d_j,\end{aligned}$$

neboť podle pozorování 11 platí $F_\varepsilon(q_j) - F_\varepsilon(-q_j) = G_\varepsilon(q_j^2) = G_\varepsilon(G_\varepsilon^{-1}(d_j)) = d_j$ a $f_\varepsilon(q_j) + f_\varepsilon(-q_j) = 2q_j g_\varepsilon(q_j^2)$. Neznámé kvantity jsou pouze hustota g_ε a kvantily q_j^2, které odhadneme pomocí $\widehat{g}_n$, resp. pomocí $\hat{q}_{n,j}^2 = e_{(d_j n)}^2(\widehat{\boldsymbol{\beta}}_n)$. Nyní $\hat{q}_{n,j}^2 \xrightarrow{\mathsf{P}} q_j^2$ pro $n \to \infty$ (Čížek, 2004, lemma A.2) a $\widehat{g}_n$ je stejnoměrně konzistentní a omezená, odkud plyne

$$\hat{q}_{n,j}^2 \widehat{g}_n(\hat{q}_{n,j}^2) - q_j^2 g_\varepsilon(q_j^2) = \Big(\hat{q}_{n,j}^2 - q_j^2\Big)\widehat{g}_n(\hat{q}_{n,j}^2) + q_j^2\Big\{\widehat{g}_n(\hat{q}_{n,j}^2) - g_\varepsilon(q_j^2)\Big\} \xrightarrow{\mathsf{P}} 0,$$

pro $n \to \infty$. Celkem tedy dostáváme, že

$$\widehat{\gamma}_n^s = \sum_{j=1}^{J}\Big[\psi_s(d_j) - \psi_s(d_{j+1})\Big]\Big[d_j - 2\hat{q}_{n,j}^2\widehat{g}_n(\hat{q}_{n,j}^2)\Big] \xrightarrow{\mathsf{P}} \gamma^s, \quad n \to \infty.$$

□

Poznámka. Předchozí věta nespecifikuje konkrétní odhady $\widehat{G}_n$ a $\widehat{g}_n$. Distribuční funkci můžeme odhadnout například (vyhlazenou) empirickou distribuční funkcí. Podobně, hustotu g můžeme odhadnout například jádrovým odhadem hustoty.

Také poznamenejme, že navržený odhad varianční matice nemusí být přesný pro malé rozsahy výběru, neboť se jedná o asymptotickou aproximaci a navíc zahrnuje neparametrický odhad hustoty.

Asymptotická inference

Asymptotická normalita (věta 14) spolu s konzistentním odhadem asymptotické varianční matice (věta 15) umožňuje přirozeně zkonstruovat testy pro regresní koeficienty založené na LWS metodě.

Mějme $\boldsymbol{c} \in \mathbb{R}^k$ nenulový vektor a označme $\lambda = \boldsymbol{c}^\top\boldsymbol{\beta}_0$ a $\widehat{\lambda}_n = \boldsymbol{c}^\top\widehat{\boldsymbol{\beta}}_n^{\mathsf{LWS}}$. Podobně uvažujme matici $\mathbb{L} \in \mathbb{R}^{m\times k}$ s $m \leq k$ lineárně nezávislými, nenulovými řádky a označme $\boldsymbol{\theta} = \mathbb{L}\boldsymbol{\beta}_0$, resp. $\widehat{\boldsymbol{\theta}}_n = \mathbb{L}\widehat{\boldsymbol{\beta}}_n^{\mathsf{LWS}}$.

Tvrzení 16. *Nechť platí předpoklady věty 15 a označme asymptotickou varianční matici z věty 14 jako $\mathbb{V}$, její konzistentní odhad jako $\widehat{\mathbb{V}}_n$. Potom*

- $T_n = \frac{\sqrt{n}(\widehat{\lambda}_n - \lambda)}{\sqrt{\boldsymbol{c}^\top\widehat{\mathbb{V}}_n\boldsymbol{c}}} \xrightarrow{D} \mathsf{N}(0,1)$ *pro* $n \to \infty$,
- $Q_n = n(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})^\top(\mathbb{L}\widehat{\mathbb{V}}_n\mathbb{L}^\top)^{-1}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{D} \chi_m^2$ *pro* $n \to \infty$.

Důkaz. Tvrzení dokážeme. Pro první část tvrzení máme

$$T_n = \frac{\sqrt{n}(\widehat{\lambda}_n - \lambda)}{\sqrt{\boldsymbol{c}^\top\mathbb{V}\boldsymbol{c}}}\sqrt{\frac{\boldsymbol{c}^\top\mathbb{V}\boldsymbol{c}}{\boldsymbol{c}^\top\widehat{\mathbb{V}}_n\boldsymbol{c}}},$$

kde první člen konverguje v distribuci ke standardnímu normálnímu rozdělení pro $n \to \infty$, neboť $\widehat{\boldsymbol{\beta}}_n^{\mathsf{LWS}}$ je asymptoticky normální (věta 14) a aplikací Cramérovy-Woldovy věty dostáváme

$$\sqrt{n}(\widehat{\lambda}_n - \lambda) \xrightarrow{\mathsf{D}} \mathsf{N}(0, \boldsymbol{c}^\top \mathbb{V} \boldsymbol{c}), \ n \to \infty.$$

Druhý člen konverguje v pravděpodobnosti k 1 pro $n \to \infty$ z konzistence $\widehat{\mathbb{V}}_n$. Aplikací Cramérovy-Slutského věty dostáváme první část tvrzení.

Pro druhý bod máme

$$Q_n = \sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})^\top (\mathbb{L}\widehat{\mathbb{V}}_n \mathbb{L}^\top)^{-1} \sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}),$$

kde první a třetí člen konvergují v distribuci k $\mathsf{N}_m(\mathbf{0}_m, \mathbb{L}\mathbb{V}\mathbb{L}^\top)$ pro $n \to \infty$ opět z Cramérovy-Woldovy věty. Prostřední člen konverguje v pravděpodobnosti k $(\mathbb{L}\mathbb{V}\mathbb{L}^\top)^{-1}$ z konzistence $\widehat{\mathbb{V}}_n$. Tedy Q_n jakožto kvadratická forma konverguje v distribuci k χ^2_m pro $n \to \infty$.

□

Inference pro regresní koeficienty

Předpokládejme, že chceme testovat nulovou hypotézu $\mathcal{H}_0 : \beta_{0,j} = b$ proti alternativě $\mathcal{H}_1 : \beta_{0,j} \neq b$. Pak můžeme položit $\boldsymbol{c} = \boldsymbol{e}_j$ a uvažovat testovou statistiku

$$T_n = \frac{\sqrt{n}(\widehat{\beta}_{n,j}^{\mathsf{LWS}} - b)}{\sqrt{v_{jj}}},$$

kde v_{jj} je j-tý diagonální prvek matice $\widehat{\mathbb{V}}_n$. Podle tvrzení 16 zamítáme nulovou hypotézu, pokud $|T_n| \geq u_{1-\frac{\alpha}{2}}$, kde u_α je α-kvantil $\mathsf{N}(0,1)$ rozdělení. Dostáváme konfidenční interval pro $\beta_{0,j}$ s asymptotickým pokrytím $1-\alpha$ tvaru

$$\left(\widehat{\beta}_{n,j}^{\mathsf{LWS}} - u_{1-\frac{\alpha}{2}} \sqrt{\frac{v_{jj}}{n}}, \ \widehat{\beta}_{n,j}^{\mathsf{LWS}} + u_{1-\frac{\alpha}{2}} \sqrt{\frac{v_{jj}}{n}} \right).$$

Simultánní inference pro vektor regresních koeficientů

Uvažujme testovou statistiku

$$Q_n = n(\widehat{\boldsymbol{\beta}}_n^{\mathsf{LWS}} - \boldsymbol{b})^\top \widehat{\mathbb{V}}_n^{-1} (\widehat{\boldsymbol{\beta}}_n^{\mathsf{LWS}} - \boldsymbol{b})$$

za účelem testování nulové hypotézy $\mathcal{H}_0 : \boldsymbol{\beta}_0 = \boldsymbol{b}$ proti alternativě $\mathcal{H}_1 : \boldsymbol{\beta}_0 \neq \boldsymbol{b}$. Pak podle tvrzení 16 zamítáme $\mathcal{H}_0$ pokud $Q_n \geq \chi^2_k(1-\alpha)$, kde $\chi^2_k(\alpha)$ je α-kvantil χ^2 rozdělení s k stupni volnosti. Konfidenční množina s asymptotickým pokrytím $1-\alpha$ pro vektor $\boldsymbol{\beta}_0$ je potom tvaru

$$\left\{ \boldsymbol{\beta} \in \mathbb{R}^k : n(\widehat{\boldsymbol{\beta}}_n^{\mathsf{LWS}} - \boldsymbol{\beta})^\top \widehat{\mathbb{V}}_n^{-1} (\widehat{\boldsymbol{\beta}}_n^{\mathsf{LWS}} - \boldsymbol{\beta}) < \chi^2_k(1-\alpha) \right\}.$$

Test podmodelu

Uvažujme model $\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon} = \mathbb{X}_A\boldsymbol{\beta}_A + \mathbb{X}_B\boldsymbol{\beta}_B + \boldsymbol{\varepsilon}$, kde $\mathbb{X}_A \in \mathbb{R}^{n\times(k-m)}$, $\mathbb{X}_B \in \mathbb{R}^{n\times m}$, $\boldsymbol{\beta}_A \in \mathbb{R}^{k-m}$ a $\boldsymbol{\beta}_B \in \mathbb{R}^m$. Předpokládejme, že chceme testovat platnost podmodelu $\boldsymbol{Y} = \mathbb{X}_A\boldsymbol{\beta}_A + \boldsymbol{\varepsilon}$, neboli nulovou hypotézu $\mathcal{H}_0 : \boldsymbol{\beta}_B = \mathbf{0}_m$ proti alternativě $\mathcal{H}_1 : \boldsymbol{\beta}_B \neq \mathbf{0}_m$. Potom můžeme položit $\mathbb{L} = (\mathbb{O}_{m\times(k-m)}, \mathbb{I}_m)$ a označme $\mathbb{L}\widehat{\mathbb{V}}_n\mathbb{L}^\top = \widehat{\mathbb{V}}_B$. Testová statistika z tvrzení 16 má tvar

$$Q_n = n\widehat{\boldsymbol{\beta}}_B^\top \widehat{\mathbb{V}}_B^{-1}\widehat{\boldsymbol{\beta}}_B,$$

kde $\widehat{\boldsymbol{\beta}}_n^{\mathsf{LWS}} = (\widehat{\boldsymbol{\beta}}_A^\top, \widehat{\boldsymbol{\beta}}_B^\top)^\top$ je odhad metodou nejmenších vážených čtverců. Nulovou hypotézu zamítáme, pokud $Q_n \geq \chi^2_m(1-\alpha)$.

3.6 Adaptivní váhy

Nyní představíme na datech závislé adaptivní váhy pro odhad metodou nejmenších vážených čtverců. Výhodou tohoto přístupu je, že váhy mohou záviset na neparametrických odhadech neznámých funkcí, jako je distribuční nebo kvantilová funkce chybových členů ε_i.

Uvažujme opět obecný lineární regresní model a počáteční robustní odhady regresních koeficientů $\widehat{\boldsymbol{\beta}}_n^0$ a reziduální směrodatné odchylky $\widehat{\sigma}_n^0$. Vektor počátečních reziduí je tedy $\boldsymbol{e}^0 = \boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}_n^0$.

Definice 24. *Definujme váhy* $w_i = \widehat{\psi}_n\big\{(2i-1)/(2n)\big\}$, *kde* $\widehat{\psi}_n$ *je váhová funkce, která může záviset na* $\widehat{\boldsymbol{\beta}}_n^0$, $\widehat{\sigma}_n^0$ *nebo* $\boldsymbol{e}^0$, *ale předpokládáme, že konverguje k po částech spojité funkci* $\psi : [0,1] \to \mathbb{R}_{\geq 0}$, $\widehat{\psi}_n(t) \to \psi(t)$ *pro* $n \to \infty$ *a všechna* $t \in [0,1]$. *Potom definujeme odhad* **metodou nejmenších adaptivně vážených čtverců** (AW) *jako*

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_n^{\mathsf{AW}} &= \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^k} \frac{1}{n}\sum_{i=1}^n \widehat{\psi}_n\Big(\tfrac{2i-1}{2n}\Big)\Big(Y_i - \boldsymbol{X}_i^\top\boldsymbol{\beta}\Big)^2_{(i)} \\ &= \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^k} \frac{1}{n}\sum_{i=1}^n \widehat{\psi}_n\Big(\widehat{G}_n\big\{e_i^2(\boldsymbol{\beta})\big\} - \tfrac{1}{2n}\Big)e_i^2(\boldsymbol{\beta}).\end{aligned}$$

Poznámka. Na rozdíl od LWS, váhová funkce $\widehat{\psi}_n$ závisí na datech a může být odhadem neznámé funkce. Protože AW odpovídá LWS pro váhovou funkci $\widehat{\psi}_n \equiv \psi$ nezávislou na datech, může se zdát, že není asymptoticky žádný rozdíl mezi AW s váhovou funkcí $\widehat{\psi}_n \to \psi$ a LWS s váhovou funkcí ψ. Zásadním rozdílem je, že váhy $\widehat{\psi}_n$ mohou konvergovat k neznámé funkci ψ, která může záviset například na distribuční funkci ε_i. Naproti tomu, LWS lze použít pouze v případě, kdy je váhová funkce ψ známá.

Váhové funkce

Nyní představíme konkrétní váhové funkce pro AW odhad.

Binární váhová funkce

Předpokládejme, že $\varepsilon_i \sim \mathsf{N}(0, \sigma^2)$. Potom $|\varepsilon_i/\sigma|$ mají složené normální rozdělení $|\mathsf{N}(0,1)|$, označme jeho distribuční funkci jako F_0^+. Skutečnou distribuční funkci chybových členů $|\varepsilon_i/\sigma|$ budeme značit F_ε^+. Definujme

$$d_0 = \sup_{t \geq c} \left\{ \max \left\{ 0, F_0^+(t) - F_\varepsilon^+(t) \right\} \right\},$$

kde c je velký kvantil F_0^+. Pro naši volbu $|\mathsf{N}(0,1)|$ můžeme uvažovat např. $c = 2.5$. Potom d_0 měří největší rozdíl mezi F_0^+ a F_ε^+ na chvostu rozdělení a můžeme definovat váhy

$$\psi^{\mathsf{R}}(t) = \mathbb{1}\{t \leq 1 - d_0\}, \quad t \in [0,1]. \tag{3.7}$$

V praxi však distribuční funkci F_ε^+ neznáme a musíme ji nahradit empirickou distribuční funkcí $\widehat{F}_n^+$ absolutních standardizovaných reziduí $|e_i(\widehat{\boldsymbol{\beta}}_n^0)/\widehat{\sigma}_n^0|$, kde podobně jako u M-odhadů můžeme volit například

$$\widehat{\sigma}_n^0 = 1.483 \cdot \mathsf{med}_{1 \leq i \leq n} \, |e_i(\widehat{\boldsymbol{\beta}}_n^0) - \mathsf{med}_{1 \leq i \leq n} \, e_i(\widehat{\boldsymbol{\beta}}_n^0)|. \tag{3.8}$$

Dostáváme analogii váhové funkce (3.7).

Definice 25. *Definujme váhovou funkci*

$$\widehat{\psi}_n^{\mathsf{R}}(t) = \mathbb{1}\{t \leq 1 - d_n\}, \quad t \in [0,1],$$

kde

$$d_n = \sup_{t \geq c} \left\{ \max \left\{ 0, F_0^+(t) - \widehat{F}_n^+(t) \right\} \right\}$$

a c reprezentuje velký kvantil $|\mathsf{N}(0,1)|$ rozdělení, volíme například $c = 2.5$. Příslušný odhad s touto váhovou funkcí budeme značit AWR.

Poznámka. Všimněme si, že AWR odhad je vlastně LTS odhadem s adaptivně zvoleným hyperparametrem

$$\delta_n = \lfloor (1 - d_n)n \rfloor = \sum_{i=1}^{n} \widehat{\psi}_n^{\mathsf{R}}\left\{ \tfrac{2i-1}{2n} \right\}.$$

Tedy zanedbá pouze adaptivně zvolený počet pozorování.

Poznámka. Jak uvidíme v tvrzení 20, za určitých předpokladů by měla váhová funkce $\widehat{\psi}_n^{\mathsf{R}}$ konvergovat k $\psi^{\mathsf{R}}(t) = \mathbb{1}\{t \leq 1 - d_0\}$.

Kvantilová váhová funkce

Nyní představíme striktně kladné váhy, které garantují vysoký bod selhání a současně vysokou relativní eficienci. Účelová funkce je založena na metodě nejmenších čtverců, která je eficientní za normálně rozdělených chybových členů ε_i. Tohoto faktu nyní využijeme k představení vah takových, že vážená rezidua budou normálně rozdělená (tzn. optimálně pro LS) v počátečním odhadu regresních koeficientů $\widehat{\boldsymbol{\beta}}_n^0$.

Značení. Označme F_χ distribuční funkci χ^2 rozdělení s jedním stupněm volnosti. Připomeňme, že $\widehat{G}_n$ značí stejnoměrně konzistentní odhad distribuční funkce G_ε. Podobně, $\widehat{G}_n^0$ bude značit empirickou distribuční funkci $e_i^2(\widehat{\boldsymbol{\beta}}_n^0)$. Nakonec, $G_{\boldsymbol{\beta}}$ bude značit distribuční funkci čtverců reziduí $e_i^2(\boldsymbol{\beta})$. Příslušné kvantily budeme značit $(\widehat{G}_n^0)^{-1}$, resp. $G_{\boldsymbol{\beta}}^{-1}$.

Předpokládejme nejprve, že rezidua $e_i(\boldsymbol{\beta})$ mají standardní normální rozdělení pro nějaké $\boldsymbol{\beta} \in \mathbb{R}^k$. Potom čtverce reziduí $e_i^2(\boldsymbol{\beta})$ mají χ^2 rozdělení s 1 stupněm volnosti. Pro obecně rozdělená rezidua toho docílíme transformací $F_\chi^{-1}(G_{\boldsymbol{\beta}}(e_i^2(\boldsymbol{\beta})))$. Pro AW odhad tedy definujeme následující váhy, viz definici 24.

Definice 26. *Definujme váhovou funkci tvaru*

$$\widehat{\psi}_n^{\mathsf{Q}}(t) = (\widehat{\sigma}_n^0)^2 \frac{F_\chi^{-1}(\max\{t, c_n\})}{(\widehat{G}_n^0)^{-1}(\max\{t, c_n\})}, \quad t \in [0,1], \tag{3.9}$$

kde uvažujeme $c_n = \min\{\frac{m}{n} : e_{(m)}^2(\widehat{\boldsymbol{\beta}}_n^0) > 0\}$*, abychom se vyhnuli dělení nulou a* $\widehat{\sigma}_n^0$ *je počáteční odhad reziduální směrodatné odchylky, viz (3.8). Příslušný odhad s touto váhovou funkcí budeme značit* AWQ*.*

Výhodou AWQ je, že váhová funkce je všude kladná a tedy nezanedbává žádná pozorování. Váhová funkce $\widehat{\psi}_n^{\mathsf{Q}}$ navíc nemusí záviset na odhadu rozptylu, neboť v (3.9) nemusíme nutně uvažovat člen $(\widehat{\sigma}_n^0)^2$.

Poznámka. Za určitých předpokladů by měla váhová funkce $\widehat{\psi}_n^{\mathsf{Q}}$ konvergovat k $\psi^{\mathsf{Q}}(t) = \sigma^2 F_\chi^{-1}(t)/G_{\beta_0}^{-1}(t)$, jak uvidíme v tvrzení 20.

Poznámka. Čížek (2007) diskutuje i alternativní váhové funkce. Všechna tvrzení, která postupně formulujeme, potom platí i pro tyto váhové funkce.

Poznámka. Představené váhové funkce můžeme dále kombinovat. Například můžeme uvažovat součinové váhy $\widehat{\psi}_n^{\mathsf{Q}}(t) \cdot \widehat{\psi}_n^{\mathsf{R}}(t),\ t \in (0,1)$.

Základní vlastnosti

Následující tvrzení ukazuje, že za normálně rozdělených chybových členů konvergují představené váhové funkce bodově ke konstantní funkci. Pro normálně rozdělená data je tedy účelová funkce asymptoticky ekvivalentní nejmenším čtvercům, z čehož později vyplyne asymptotická eficience AW odhadu za normality.

Tvrzení 17. *Nechť* $\{[\boldsymbol{X}_1^\top, \varepsilon_1]^\top, \ldots, [\boldsymbol{X}_n^\top, \varepsilon_n]^\top\}$ *je posloupnost nezávislých stejně rozdělených náhodných vektorů a* $\varepsilon_i \sim \mathsf{N}(0, \sigma^2)$*. Dále uvažujme slabě konzistentní počáteční odhady* $\widehat{\boldsymbol{\beta}}_n^0$ *a* $\widehat{\sigma}_n^0$*. Potom pro všechna* $t \in (0,1)$ *platí*

$$\widehat{\psi}_n^{\mathsf{R}}(t) \xrightarrow{\mathsf{P}} 1 \quad a \quad \widehat{\psi}_n^{\mathsf{Q}}(t) \xrightarrow{\mathsf{P}} 1 \quad \text{pro } n \to \infty.$$

Důkaz. Uvažujme libovolné $t \in (0,1)$. Podle lemmatu 4.1 v Gervini a Yohai (2002) platí $\widehat{\psi}_n^{\mathsf{R}}(t) \xrightarrow{\mathsf{P}} \psi^{\mathsf{R}}(t)$. Pro normálně rozdělené chybové členy jsou však distribuční funkce F_0^+ a F_ε^+ totožné a tedy $\widehat{\psi}_n^{\mathsf{R}}(t) \xrightarrow{\mathsf{P}} 1$ pro $n \to \infty$.

Důkaz pro kvantilovou váhovou funkci přebíráme z článku Čížek (2007). Z předpokladů $\widehat{\boldsymbol{\beta}}_n^0 \xrightarrow{\mathsf{P}} \boldsymbol{\beta}_0$ a $(\widehat{\sigma}_n^0)^2 \xrightarrow{\mathsf{P}} \sigma^2$, tedy i $e_i(\widehat{\boldsymbol{\beta}}_n^0) \xrightarrow{\mathsf{P}} e_i(\boldsymbol{\beta}_0) \equiv \varepsilon_i$. Navíc

$$\sup_{\beta \in U(\boldsymbol{\beta}_0, \delta)} |\widehat{G}_n^{-1}(t) - G_\beta^{-1}(t)| \xrightarrow{\mathsf{P}} 0$$

na nějakém okolí $U(\boldsymbol{\beta}_0, \delta)$, $\delta > 0$ (Čížek, 2004, lemma A.2). Předpoklad normality nyní implikuje, že

$$\frac{(\widehat{G}_n^0)^{-1}(t)}{(\widehat{\sigma}_n^0)^2} \xrightarrow{\mathsf{P}} \frac{G_{\boldsymbol{\beta}_0}^{-1}(t)}{\sigma^2} = F_\chi^{-1}(t).$$

Odtud plyne $\widehat{\psi}_n^{\mathsf{Q}}(t) \xrightarrow{\mathsf{P}} 1$ pro $n \to \infty$.

□

Bod selhání představených metod se rovná minimu bodů selhání počátečních odhadů regresních koeficientů a reziduální směrodatné odchylky.

Tvrzení 18. *Uvažujme nezávislá, stejně rozdělená*[8] *trénovací data $\mathcal{D}$, která jsou skoro jistě v obecné poloze pro $n > k$. Označme ε_n^{0*} bod selhání pro počáteční odhad $(\widehat{\boldsymbol{\beta}}_n^0, \widehat{\sigma}_n^0)$ s limitou $\varepsilon_n^{0*} \to \varepsilon^{0*}$ pro $n \to \infty$. Potom body selhání pro* AWR *a* AWQ *odhady jsou větší nebo rovno*

$$\min\left\{\varepsilon_n^{0*}, \frac{\lfloor (n+1)/2 \rfloor - (k+1)}{n}\right\},$$

a asymptoticky konvergují k ε^{0*}.

Důkaz. Čížek (2011), tvrzení 2.

□

Asymptotické vlastnosti

Představené váhové funkce závisí na odhadech distribuční, resp. kvantilové funkce regresních reziduí, a tedy konvergují ke konkrétní deterministické funkci, $\widehat{\psi}_n \to \psi$ pro $n \to \infty$. Za chvíli uvidíme, že asymptotické výsledky pro AW odhad s náhodnou váhovou funkcí $\widehat{\psi}_n$ jsou ekvivalentní asymptotickým výsledkům pro LWS odhad s váhovou funkcí ψ, které jsme formulovali ve větě 14.

Předpoklady uvažované v předchozí sekci rozšíříme o několik požadavků na náhodnou váhovou funkci $\widehat{\psi}_n$.

Předpoklad 3. *Nechť platí předpoklad 2 s váhovou funkcí ψ a uvažujme* AW *odhad $\widehat{\boldsymbol{\beta}}_n^{\mathsf{AW}}$ s omezenou váhovou funkcí $\widehat{\psi}_n$ založenou na počátečních odhadech $\widehat{\boldsymbol{\beta}}_n^0$ a $\widehat{\sigma}_n^0$. Navíc předpokládejme, že $\widehat{\psi}_n(t) \xrightarrow{\mathsf{P}} \psi(t)$ na $[0,1]$ a $n^{-\alpha}|\widehat{\psi}_n(t) - \psi(t)| \xrightarrow{\mathsf{P}} 0$ stejnoměrně na libovolné kompaktní podmnožině $(0,1), \alpha > 0$ pro $n \to \infty$.*

[8] Bod selhání pro závislá pozorování obecně závisí na volbě modelu.

Věta 19. *Nechť platí obecný lineární model, předpoklad 1 a předpoklad 3.*

Potom $\widehat{\boldsymbol{\beta}}_n^{\mathsf{AW}}$ *je slabě konzistentním odhadem* $\boldsymbol{\beta}_0$ *a platí*

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n^{\mathsf{AW}} - \boldsymbol{\beta}_0\right) \xrightarrow{D} \mathsf{N}_k\left(\mathbf{0}_k, \mathbb{V}\right)$$

pro $n \to \infty$*, kde* $\mathbb{V}$ *je asymptotická varianční matice z věty 14.*

Důkaz. Čížek (2007), důsledek 5.2.

□

Poznámka. Odhad regresních koeficientů AW metodou má tedy asymptoticky normální rozdělení nezávislé na počátečních odhadech. Asymptotická varianční matice se shoduje s LWS odhadem a tedy ji můžeme konzistentně odhadnout jako ve větě 15. Testy hypotéz, které jsme zkonstruovali v kapitole 3.5 pro LWS, jsou tedy platné i pro AW odhad.

Nyní uvedeme příklad jednoduchých podmínek regularity, za kterých splňují představené váhové funkce předpoklady věty 19.

Tvrzení 20. *Nechť* $\{[\boldsymbol{X}_1^\top, \varepsilon_1]^\top, \ldots, [\boldsymbol{X}_n^\top, \varepsilon_n]^\top\}$ *je posloupnost nezávislých stejně rozdělených náhodných vektorů, distribuční funkce* F_ε *veličin* ε_i *splňuje druhou podmínku z předpokladu 1 a* $z^2 f_\varepsilon'(z)$ *je omezená. Pokud je počáteční odhad* $(\widehat{\boldsymbol{\beta}}_n^0, \widehat{\sigma}_n^0)$ n^α*-konzistentní,* $\alpha \geq 1/4$*, potom*

$$\sup_{t\in[a,b]} |\widehat{\psi}_n^{\mathsf{R}}(t) - \psi^{\mathsf{R}}(t)| = \mathcal{O}_{\mathsf{P}}\left(n^{-\frac{1}{2}}\right) \quad a \quad \sup_{t\in[a,b]} |\widehat{\psi}_n^{\mathsf{Q}}(t) - \psi^{\mathsf{Q}}(t)| = \mathcal{O}_{\mathsf{P}}\left(n^{-\frac{1}{2}}\right)$$

pro libovolný interval $[a, b] \subseteq (0, 1)$ *a* $n \to \infty$*.*

Důkaz. Čížek (2011), tvrzení 4.

□

Nyní jsme připraveni ukázat, že za normality mají odhady regresních koeficientů AW metodou a metodou nejmenších čtverců stejná asymptotická rozdělení.

Důsledek 21. *Buď* $\{[\boldsymbol{X}_1^\top, \varepsilon_1]^\top, \ldots, [\boldsymbol{X}_n^\top, \varepsilon_n]^\top\}$ *posloupnost nezávislých stejně rozdělených náhodných vektorů s konečnými druhými momenty a* $\varepsilon_i \sim \mathsf{N}(0, \sigma^2)$*. Dále nechť jsou* $\boldsymbol{X}_i$ *a* ε_i *nezávislé, matice* $\mathbb{D} = \mathsf{E}\,\boldsymbol{X}_i\boldsymbol{X}_i^\top$ *je regulární a platí* $n^{-1/4}\max_{1\leq i\leq n}\|\boldsymbol{X}_i\|_\infty = \mathcal{O}_{\mathsf{P}}(1)$*. Uvažujme* n^α*-konzistentní počáteční odhady* $\widehat{\boldsymbol{\beta}}_n^0$ *a* $\widehat{\sigma}_n^0$ *pro nějaké* $\alpha \geq \frac{1}{4}$*. Potom* LS*,* AWR *a* AWQ *mají asymptoticky stejná rozdělení.*

Důkaz. Důsledek dokážeme. Označme funkci[9]

$$h(z) = z^2 f_\varepsilon'(z) = -\frac{1}{\sqrt{2\pi}}\left(\frac{z}{\sigma}\right)^3 \exp\left\{-\frac{1}{2}\left(\frac{z}{\sigma}\right)^2\right\},\ \sigma \in (0, \infty)$$

a uvažujme její reparametrizaci $h(x) = -\frac{1}{\sqrt{2\pi}}x^3\exp\{-\frac{1}{2}x^2\}$.
Funkce $x^3\exp\{-x^2/2\}$ je spojitá a platí

$$\lim_{x\to\pm\infty}\frac{x^3}{\exp\{x^2/2\}} = \lim_{x\to\pm\infty}\frac{3x}{\exp\{x^2/2\}} = \lim_{x\to\pm\infty}\frac{3}{x\exp\{x^2/2\}} = 0,$$

[9]Využili jsme faktu, že pro hustotu $\mathsf{N}(\mu, \sigma^2)$ rozdělení platí $f'(x) = -f(x)(\frac{x-\mu}{\sigma^2})$.

kde jsme opakovaně použili L'Hospitalovo pravidlo. Tedy existuje $K \in \mathbb{R}$ takové, že $|x^3 \exp\{-\frac{1}{2}x^2\}| \leq K$ pro všechna $x \in \mathbb{R}$. Dostáváme, že funkce h je omezená, $|h(x)| \leq (2\pi)^{-1/2}K$, pro všechna $x \in \mathbb{R}$.

Tedy jsou splněny předpoklady tvrzení 20, díky kterému platí věta 19 a tedy asymptotická varianční matice je tvaru

$$\mathbb{V} = \frac{\mathbb{D}^{-1}\,\mathsf{var}[\boldsymbol{X}_1\varepsilon_1\psi\{G_\varepsilon(\varepsilon_1^2)\}]\mathbb{D}^{-1}}{[\int \varepsilon\psi\{G_\varepsilon(\varepsilon^2)\}f'_\varepsilon(\varepsilon)\,\mathrm{d}\varepsilon]^2} = \frac{\mathbb{D}^{-1}\,\mathsf{var}[\boldsymbol{X}_1\varepsilon_1]\mathbb{D}^{-1}}{[\int \varepsilon f'_\varepsilon(\varepsilon)\,\mathrm{d}\varepsilon]^2},$$

neboť podle tvrzení 17 je limitní váhová funkce $\psi(t)$ na intervalu $(0,1)$ konstantní. Nyní využijeme nezávislost $\boldsymbol{X}_1$ a ε_1, konečnost momentů a předpoklad $\varepsilon_i \sim \mathsf{N}(0,\sigma^2)$,

$$\mathsf{var}[\boldsymbol{X}_1\varepsilon_1] = \mathsf{E}[\boldsymbol{X}_1\varepsilon_1]^{\otimes 2} - [\mathsf{E}\,\boldsymbol{X}_1\varepsilon_1]^{\otimes 2} = \mathsf{E}\,\boldsymbol{X}_1^{\otimes 2}\,\mathsf{E}\,\varepsilon_1^2 - [\mathsf{E}\,\boldsymbol{X}_1]^{\otimes 2}[\mathsf{E}\,\varepsilon_1]^2 = \sigma^2\mathbb{D}.$$

Dále

$$-\int \varepsilon f'_\varepsilon(\varepsilon)\,\mathrm{d}\varepsilon = \sigma^{-2}\int \varepsilon^2 f_\varepsilon(\varepsilon)\,\mathrm{d}\varepsilon = \sigma^{-2}\,\mathsf{var}\,\varepsilon_1 = 1,$$

kde f_ε je hustota $\mathsf{N}(0,\sigma^2)$ rozdělení.

Tedy dostáváme, že $\mathbb{V} = \sigma^2\mathbb{D}^{-1}$, což je asymptotická varianční matice LS odhadu, viz tvrzení 1.

□

4. Neuronové sítě

Umělá neuronová síť je výpočetní model inspirovaný strukturou neuronových sítí v mozku. Skládá se z neuronů, které jsou vzájemně propojeny synaptickými vazbami, navzájem si předávají signály a transformují je pomocí aktivačních funkcí. Neuronové sítě lze popsat jako orientované grafy, jejichž vrcholy odpovídají neuronům a hrany vazbám mezi nimi. V této práci se budeme zabývat pouze dopřednými neuronovými sítěmi, jejichž grafy neobsahují cykly.

Předpokládejme, že pozorujeme spojitou odezvu $Y_i \in \mathcal{Y} \subseteq \mathbb{R}$ a vektor regresorů $\boldsymbol{X}_i \in \mathcal{X} \subseteq \mathbb{R}^k$ pro celkem n pozorování. V této kapitole se budeme zabývat úlohou **nelineární regrese**, kdy předpokládáme

$$Y_i = \varphi(\boldsymbol{X}_i; \boldsymbol{w}) + \varepsilon_i, \quad i = 1, \dots, n,$$

kde φ je neznámá nelineární funkce a $\varepsilon_1, \dots, \varepsilon_n$ jsou chybové členy. Na rozdíl od lineární regrese budeme parametry modelu nazývat **váhami** a značit $\boldsymbol{w}$. Nejprve zadefinujeme dopřednou neuronovou síť a představíme běžně používaný heuristický přístup trénování neuronových sítí. Následně diskutujeme, jak se pomocí zpětné propagace počítá gradient ztrátové funkce. Nakonec představíme několik robustních neuronových sítí v kontextu nelineární regrese, které jsou motivovány robustními odhady diskutovanými v předchozí kapitole.

V této kapitole vycházíme především z knihy Shalev-Shwartz a Ben-David (2014), kterou doplníme knihou Goodfellow a kol. (2016).

4.1 Definice neuronové sítě

Dopřednou neuronovou síť můžeme popsat jako orientovaný acyklický graf $\mathcal{G} = (V, E)$ spolu s váhovou funkcí hran $w : E \to \mathbb{R}$, kde V je neprázdná množina vrcholů a E je množina hran. Vrcholy grafu potom nazýváme **neurony** a modelujeme jednoduchou nelineární funkcí $\sigma : \mathbb{R} \to \mathbb{R}$, které říkáme **aktivační funkce**. Každá hrana takového grafu spojuje výstup nějakého neuronu se vstupem jiného neuronu. Vstup neuronu získáme jako vážený součet výstupů všech neuronů, které jsou s ním spojeny hranou, na základě váhové funkce w.

Předpokládáme, že síť je uspořádána do vrstev, tedy že množina vrcholů je sjednocením disjunktních podmnožin $V = \bigcup_{t=0}^{T} V_t$ tak, že každá hrana z E spojuje pro nějaké $t \in \{0, \dots, T\}$ vrchol z V_{t-1} s vrcholem z V_t.

Terminologie. Spodní vrstvu V_0 nazýváme **vstupní vrstva**. Pod **skrytými vrstvami** rozumíme vrstvy $V_1, \dots, V_{T-1}$. Někdy se jim říká také vnitřní vrstvy. O vrchní vrstvě V_T mluvíme jako o **výstupní vrstvě**.

Poznámka. V regresních problémech má výstupní vrstva jediný neuron, typicky s identitou jako aktivační funkcí. Při obecném popisu neuronových sítí však budeme uvažovat, že výstup může být vektorem. Tedy pokrýváme i případ, kdy chceme modelovat více odezev současně.

Smyslem aktivační funkce je vnést do modelu nelinearitu a případně normalizovat procházející data. Následující příklady shrnují běžně používané aktivační funkce.

Příklad. Nejprve uvedeme aktivační funkce používané ve výstupní vrstvě.

(i) **Identita:** Používá se v regresních úlohách. Bez skrytých vrstev odpovídá lineární regresi.[1]

(ii) **Exponenciála:** $\sigma(x) = e^x$, pokud neuvažujeme skryté vrstvy, tak odpovídá Poissonově regresi.[2]

(iii) **Logistická funkce:** $\sigma(x) = 1/(1 + e^{-x})$, bez skrytých vrstev odpovídá logistické regresi.[3] Uvádíme pro úplnost, ačkoliv se jedná o klasifikaci.

Příklad. Nyní shrneme aktivační funkce používané ve skrytých vrstvách včetně derivací a oborů hodnot.

(i) **Logistická funkce:** Je nesymetrická, moc se nepoužívá.

$$\sigma(x) = 1/(1 + e^{-x}), \quad \sigma'(x) = \sigma(x)(1 - \sigma(x)), \quad H(\sigma) = [0, 1].$$

(ii) **Hyperbolický tangens:**

$$\sigma(x) = \tanh(x), \quad \sigma'(x) = 1 - \tanh^2(x), \quad H(\sigma) = (-1, 1).$$

(iii) **ReLU:** Nejpoužívanější nelineární aktivační funkce. Není diferencovatelná v nule, ale má subdiferenciál $\partial\sigma(0) = [0, 1]$, viz diskuze v kapitole 1.4.

$$\sigma(x) = \max\{0, x\}, \quad \sigma'(x) = \mathbb{1}\{x > 0\} \text{ pro } x \neq 0, \quad H(\sigma) = [0, \infty).$$

(iv) **Varianty ReLU:** Leaky ReLU, Softplus, ELU, GELU a další.

Poznámka. Pro jednoduchost uvažujeme, že každý neuron modelujeme stejnou aktivační funkcí σ. Obecně bychom však měli uvažovat pro každou vrstvu neuronové sítě její aktivační funkci σ_t. Ve vstupní vrstvě se používá jako aktivační funkce identita. Volba aktivační funkce ve výstupní vrstvě zase závisí na povaze úlohy.

Značení. Označme i-tý neuron v t-té vrstvě jako $v_{t,i}$. Dále označme výstup neuronu $v_{t,i}$ pro vstupní vektor $\boldsymbol{x}$ jako $o_{t,i}(\boldsymbol{x})$. V poslední řadě označme jako $a_{t,i}(\boldsymbol{x})$ vstup i-tého neuronu v t-té vrstvě, pokud je vstupní vektor naší sítě $\boldsymbol{x}$.

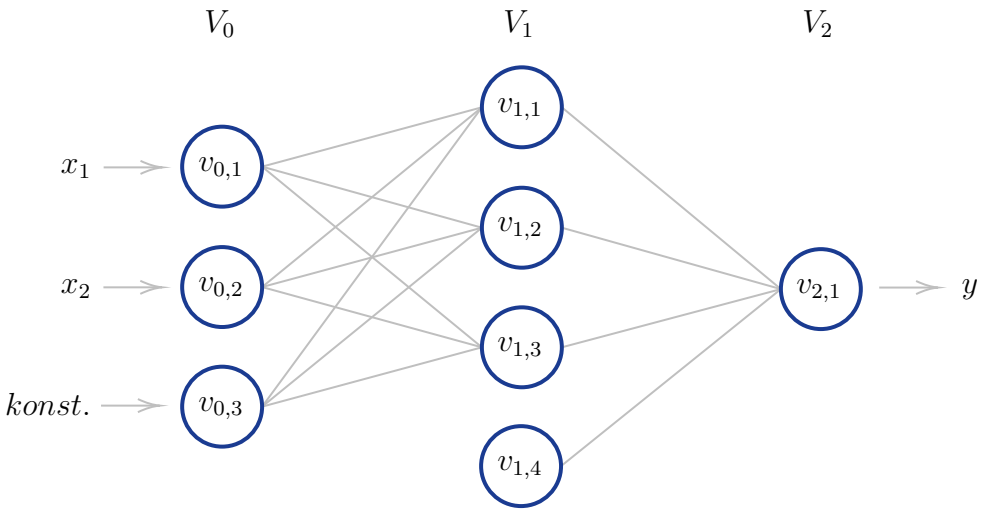
Vstupní vrstva V_0 obsahuje $k + 1$ neuronů, kde k je dimenzionalita vstupního prostoru $\mathcal{X}$. Máme $o_{0,i}(\boldsymbol{x}) = x_i$ pro $i \in \{1, \ldots, k\}$ a $o_{0,k+1} = 1$. Dále postupujeme ve výpočtu vrstvu po vrstvě. Předpokládejme, že máme k dispozici výstupy neuronů z t-té vrstvy. Pak spočítáme výstupy neuronů ve vrstvě $t + 1$ následovně. Vezměme nějaké $v_{t+1,j} \in V_{t+1}$, potom

$$a_{t+1,j}(\boldsymbol{x}) = \sum_{r:(v_{t,r}, v_{t+1,j}) \in E} w\{(v_{t,r}, v_{t+1,j})\} o_{t,r}(\boldsymbol{x}),$$

a tedy

$$o_{t+1,j}(\boldsymbol{x}) = \sigma\{a_{t+1,j}(\boldsymbol{x})\}.$$

[1] Pokud uvažujeme ℓ_2 ztrátovou funkci.
[2] Pro Poissonovo rozdělení a negativní logaritmickou věrohodnost (NLL) jako ztrátu.
[3] Pokud uvažujeme alternativní rozdělení a NLL jako ztrátovou funkci.



Obrázek 4.1. Ilustrace dopředné neuronové sítě s hloubkou 2, velikostí 8 a šířkou 4. Všimněme si osamoceného neuronu $v_{1,4}$ ve skryté vrstvě, do kterého nevstupují žádné hrany. Tento neuron má výstup $o_{1,4}(\boldsymbol{x}) = \sigma(0)$ a hraje roli absolutního členu.

Terminologie. Počet vrstev neuronové sítě bez vstupní vrstvy, tedy číslo T, nazýváme **hloubka sítě**. **Velikostí sítě** rozumíme počet neuronů $|V|$. **Šířka sítě** se potom definuje jako $\max_t |V_t|$. Graf jednoduché neuronové sítě ilustruje obrázek 4.1.

Neuronová síť (V, E, σ, w) tedy definuje zobrazení $\boldsymbol{f}_{V,E,\sigma,w} : \mathbb{R}^{|V_0|-1} \to \mathbb{R}^{|V_T|}$. Trojici (V, E, σ) říkáme **architektura** sítě. Zpravidla volíme architekturu sítě pevnou a uvažujeme model tvaru

$$\mathcal{F}_{V,E,\sigma} = \{\boldsymbol{f}_{V,E,\sigma,w} : w \text{ je zobrazení z } E \text{ do } \mathbb{R}\}.$$

Neboli predikční funkce z našeho modelu jsou parametrizovány váhami přes hrany neuronové sítě.

Věta o univerzální aproximaci říká, že neuronové sítě jsou schopny aproximovat libovolnou spojitou funkci. Nyní formulujeme její duální verzi pro **hluboké neuronové sítě**, tedy sítě s omezenou šířkou a libovolnou hloubkou. Její výhodou je, že platí v podstatě pro libovolné aktivační funkce.

Věta 22. *Uvažujme spojitou aktivační funkci $\sigma : \mathbb{R} \to \mathbb{R}$, která není afinní a je spojitě diferencovatelná alespoň v jednom bodě, ve kterém má navíc nenulovou derivaci. Pro $k, m \in \mathbb{N}$ označme $\mathcal{F}^{\sigma}_{k,m,k+m+2}$ prostor funkcí z $\mathbb{R}^k$ do $\mathbb{R}^m$ definovaných dopřednou neuronovou sítí s k neurony ve vstupní vrstvě, m neurony s identitou jako aktivační funkcí ve výstupní vrstvě a libovolným počtem skrytých vrstev, každou s $k + m + 2$ neurony a aktivační funkcí σ.*

Buď $\mathcal{X}$ kompaktní podprostor $\mathbb{R}^k$. Potom pro libovolné $\varepsilon > 0$ a každou funkci $\boldsymbol{f} \in \mathcal{C}(\mathcal{X}, \mathbb{R}^m)$ existuje $\widehat{\boldsymbol{f}} \in \mathcal{F}^{\sigma}_{k,m,k+m+2}$ taková, že

$$\sup_{\boldsymbol{x} \in \mathcal{X}} \|\widehat{\boldsymbol{f}}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x})\| < \varepsilon.$$

Důkaz. Kidger a Lyons (2020), věta 3.2. □

4.2 Trénování neuronových sítí

Problém nalezení predikční funkce z rodiny $\mathcal{F}_{V,E,\sigma}$ s malým očekávaným rizikem odpovídá nalezení optimálních vah w. V této sekci představíme běžně používaný heuristický přístup trénování neuronových sítí, založený na stochastickém gradientu. Úskalím trénování neuronových sítí je, že účelová funkce je silně nekonvexní. I přes to však můžeme použít SGD algoritmus a doufat, že nalezené řešení bude rozumné, jako je tomu v mnoha praktických aplikacích.

Jelikož množina hran E je konečná, můžeme se na váhovou funkci dívat jako na vektor $\boldsymbol{w} \in \mathbb{R}^{|E|}$. Předpokládejme, že síť má k vstupních neuronů a m výstupních neuronů a označme jako $\boldsymbol{f}(\,\cdot\,;\boldsymbol{w}) : \mathbb{R}^k \to \mathbb{R}^m$ predikční funkci spočítanou sítí, jsou-li váhy definované vektorem $\boldsymbol{w}$. Dále označme ztrátu predikce $\boldsymbol{f}(\boldsymbol{x};\boldsymbol{w})$, je-li skutečný výstup $\boldsymbol{y} \in \mathcal{Y}$, jako $\ell(\boldsymbol{y}, \boldsymbol{f}(\boldsymbol{x};\boldsymbol{w}))$. Konkrétní volby ztrátových funkcí budeme diskutovat později. Cílem trénování je nalézt váhy $\hat{\boldsymbol{w}}$ minimalizující očekávané riziko $R(\boldsymbol{w}) = \mathsf{E}[\ell(\boldsymbol{Y}, \boldsymbol{f}(\boldsymbol{X};\boldsymbol{w}))]$, respektive empirické riziko

$$R_{\mathcal{S}}(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{Y}_i, \boldsymbol{f}(\boldsymbol{X}_i;\boldsymbol{w})).$$

Tyto účelové funkce můžeme minimalizovat pomocí stochastického gradientu, jak jsme diskutovali v sekci 1.4. Počáteční volba parametrů má velký vliv na trénování neuronových sítí. Počáteční váhy mohou ovlivnit konvergenci algoritmu a její rychlost nebo zda dokonverguje k bodu s malou či velkou ztrátou. Navíc, body s podobnou hodnotou ztráty mohou mít odlišnou generalizační chybu, tedy počáteční váhy mohou mít vliv i na schopnost zobecňovat.

Počáteční váhy $\hat{\boldsymbol{w}}_1 \in \mathbb{R}^{|E|}$ volíme náhodně, z rozdělení takového, že $\hat{\boldsymbol{w}}_1$ je blízko $\mathbf{0}$. Zvolení stejných vah u skrytých neuronů spojených se stejnými vstupy a se stejnou aktivační funkcí by vedlo k aktualizaci obou těchto neuronů stejným způsobem. Navíc doufáme, že pokud SGD proceduru několikrát zopakujeme, jedna z iterací povede k dobrému lokálnímu minimu. Například u plně propojené vrstvy s m vstupy můžeme heuristicky volit každou počáteční váhu z rovnoměrného rozdělení na intervalu $\left(\frac{-1}{\sqrt{m}}, \frac{1}{\sqrt{m}}\right)$. Stochastický gradient pro trénování neuronové sítě shrnuje algoritmus 4.

Algoritmus 4: Stochastický gradient.

Vstup: Neuronová síť s predikční funkcí $\boldsymbol{f}(\,\cdot\,;\boldsymbol{w})$.
Vstup: Posloupnost kroků $\{\alpha_k : k \in \mathbb{N}\}$.
Vstup: Batch size m.
Náhodně zvol počáteční váhy $\hat{\boldsymbol{w}}_1$ tak, že $\hat{\boldsymbol{w}}_1$ je blízko $\mathbf{0}$.
for *k = 1, 2, ...* **do**
  Náhodně vyber m vzorků $[\boldsymbol{Y}_i, \boldsymbol{X}_i^\top]^\top$ z trénovacích dat $\mathcal{D}$.
  Spočítej odhad gradientu $\widehat{\boldsymbol{g}}_k \leftarrow \frac{1}{m}\sum_{i=1}^{m} \nabla\ell(\boldsymbol{Y}_i, \boldsymbol{f}(\boldsymbol{X}_i;\hat{\boldsymbol{w}}_k))$.
  Polož $\hat{\boldsymbol{w}}_{k+1} \leftarrow \hat{\boldsymbol{w}}_k - \alpha_k\widehat{\boldsymbol{g}}_k$.
end

4.3 Zpětná propagace

Nyní zbývá popsat, jak pomocí algoritmu **zpětné propagace** spočítáme gradient ztrátové funkce pro vzorek $(\boldsymbol{x}, \boldsymbol{y})$ vzhledem k vektoru vah $\boldsymbol{w}$. Řetízkové pravidlo (věta o derivaci složené funkce) z matematické analýzy je připomenuto v dodatku A.2 spolu s Jakobiho maticemi, které budeme používat.

Budeme opět předpokládat, že množinu vrcholů můžeme rozložit na jednotlivé vrstvy $V = \bigcup_{t=0}^{T} V_t$. Neurony každé vrstvy budeme značit $V_t = \{v_{t,1}, \ldots, v_{t,k_t}\}$, kde $k_t = |V_t|$ pro $t \in \{1, \ldots, t\}$. Jako $\mathbb{W}_t \in \mathbb{R}^{k_{t+1} \times k_t}$ označme matici vah všech potenciálních hran mezi V_t a V_{t+1}. Pokud $(v_{t,j}, v_{t+1,i}) \in E$, potom $W_{t,i,j}$ je váha této hrany daná vektorem $\boldsymbol{w}$. V opačném případě hranu přidáme a položíme její váhu rovnu 0. Bez újmy na obecnosti tedy můžeme předpokládat, že všechny hrany existují. Nyní potřebujeme spočítat parciální derivace vzhledem ke složkám matice $\mathbb{W}_{t-1}$. Protože fixujeme všechny ostatní váhy sítě, neurony ve vrstvě V_{t-1} nezávisí na vahách ve $\mathbb{W}_{t-1}$. Označme výstupy všech neuronů ve vrstvě V_{t-1} jako $\boldsymbol{o}_{t-1} \in \mathbb{R}^{k_{t-1}}$. Ztrátovou funkci podsítě definované vrstvami $V_t, \ldots, V_T$ označme jako funkci neuronů ve V_t, $\ell_t : \mathbb{R}^{k_t} \to \mathbb{R}$. Pro vstupy neuronů v t-té vrstvě můžeme psát $\boldsymbol{a}_t = \mathbb{W}_{t-1}\boldsymbol{o}_{t-1} \in \mathbb{R}^{k_t}$ a pro výstupy podobně $\boldsymbol{o}_t = \boldsymbol{\sigma}(\boldsymbol{a}_t)$, kde $\boldsymbol{\sigma}$ značí naši aktivační funkci aplikovanou po složkách, neboli $o_{t,j} = \sigma(a_{t,j})$ pro všechna j.

Tedy můžeme psát

$$g_t(\mathbb{W}_{t-1}) = \ell_t(\boldsymbol{o}_t) = \ell_t(\boldsymbol{\sigma}(\boldsymbol{a}_t)) = \ell_t(\boldsymbol{\sigma}(\mathbb{W}_{t-1}\boldsymbol{o}_{t-1})).$$

Dále bude užitečné přepsat g_t jako funkci vektoru $\boldsymbol{w}_{t-1} = \mathsf{vec}(\mathbb{W}_{t-1}^\top) \in \mathbb{R}^{k_{t-1}k_t}$, který získáme vektorizací matice $\mathbb{W}_{t-1}^\top$ po sloupcích. Definujme matici $\mathbb{O}_{t-1}$ s rozměry $k_t \times (k_{t-1}k_t)$ následujícím způsobem

$$\mathbb{O}_{t-1} = \begin{bmatrix} \boldsymbol{o}_{t-1}^\top & 0 & \ldots & 0 \\ 0 & \boldsymbol{o}_{t-1}^\top & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \boldsymbol{o}_{t-1}^\top \end{bmatrix}.$$

Potom platí $\mathbb{W}_{t-1}\boldsymbol{o}_{t-1} = \mathbb{O}_{t-1}\boldsymbol{w}_{t-1}$ a můžeme psát

$$g_t(\boldsymbol{w}_{t-1}) = \ell_t(\boldsymbol{\sigma}(\mathbb{O}_{t-1}\boldsymbol{w}_{t-1})).$$

Aplikací řetízkového pravidla (věta 23) dostáváme

$$\mathbb{J}_{g_t}(\boldsymbol{w}_{t-1}) = \mathbb{J}_{\ell_t}(\boldsymbol{\sigma}(\mathbb{O}_{t-1}\boldsymbol{w}_{t-1})) \cdot \mathbb{J}_{\boldsymbol{\sigma}}(\mathbb{O}_{t-1}\boldsymbol{w}_{t-1}) \cdot \mathbb{J}_{\mathbb{O}_{t-1}\boldsymbol{w}_{t-1}}(\boldsymbol{w}_{t-1}).$$

Nyní využijeme, že $\boldsymbol{o}_t = \boldsymbol{\sigma}(\boldsymbol{a}_t)$, $\boldsymbol{a}_t = \mathbb{O}_{t-1}\boldsymbol{w}_{t-1}$ a $\mathbb{J}_{\boldsymbol{\sigma}}(\boldsymbol{a}_t) = \mathsf{diag}(\boldsymbol{\sigma}'(\boldsymbol{a}_t))$. Potom

$$\mathbb{J}_{g_t}(\boldsymbol{w}_{t-1}) = \mathbb{J}_{\ell_t}(\boldsymbol{o}_t) \cdot \mathsf{diag}(\boldsymbol{\sigma}'(\boldsymbol{a}_t)) \cdot \mathbb{O}_{t-1}.$$

Pokud označíme $\boldsymbol{\delta}_t = \mathbb{J}_{\ell_t}(\boldsymbol{o}_t)$, můžeme výsledek psát ve tvaru

$$\mathbb{J}_{g_t}(\boldsymbol{w}_{t-1}) = \left(\delta_{t,1}\sigma'(a_{t,1})\boldsymbol{o}_{t-1}^\top, \ldots, \delta_{t,k_t}\sigma'(a_{t,k_t})\boldsymbol{o}_{t-1}^\top\right). \tag{4.1}$$

Zbývá už jen pro každé t spočítat vektor $\boldsymbol{\delta}_t$. Budeme postupovat rekurzivně. Protože pro výstupní vrstvu máme $\ell_T(\boldsymbol{u}) = \ell(\boldsymbol{u}, \boldsymbol{y})$, nejprve pro uvažovanou

ztrátovou funkci ℓ spočítáme $\boldsymbol{\delta}_T = \mathbb{J}_\ell(\boldsymbol{o}_T)$. Například pro ℓ_2 ztrátovou funkci $\ell(\boldsymbol{u}, \boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{u} - \boldsymbol{y}\|_2^2$ dostáváme $\boldsymbol{\delta}_T = \boldsymbol{o}_T - \boldsymbol{y}$. Dále platí

$$\ell_t(\boldsymbol{u}) = \ell_{t+1}(\boldsymbol{\sigma}(\mathbb{W}_t\boldsymbol{u})).$$

Opět aplikujeme řetízkové pravidlo,

$$\mathbb{J}_{\ell_t}(\boldsymbol{u}) = \mathbb{J}_{\ell_{t+1}}(\boldsymbol{\sigma}(\mathbb{W}_t\boldsymbol{u})) \cdot \mathsf{diag}(\boldsymbol{\sigma}'(\mathbb{W}_t\boldsymbol{u})) \cdot \mathbb{W}_t.$$

Speciálně dostáváme, že

$$\begin{aligned}\boldsymbol{\delta}_t = \mathbb{J}_{\ell_t}(\boldsymbol{o}_t) &= \mathbb{J}_{\ell_{t+1}}(\boldsymbol{\sigma}(\mathbb{W}_t\boldsymbol{o}_t)) \cdot \mathsf{diag}(\boldsymbol{\sigma}'(\mathbb{W}_t\boldsymbol{o}_t)) \cdot \mathbb{W}_t \\ &= \mathbb{J}_{\ell_{t+1}}(\boldsymbol{o}_{t+1}) \cdot \mathsf{diag}(\boldsymbol{\sigma}'(\boldsymbol{a}_{t+1})) \cdot \mathbb{W}_t = \boldsymbol{\delta}_{t+1} \cdot \mathsf{diag}(\boldsymbol{\sigma}'(\boldsymbol{a}_{t+1})) \cdot \mathbb{W}_t,\end{aligned}$$

kde jsme využili, že $\boldsymbol{\sigma}(\mathbb{W}_t\boldsymbol{o}_t) = \boldsymbol{\sigma}(\boldsymbol{a}_{t+1}) = \boldsymbol{o}_{t+1}$.

Shrnutí

Nejprve průchodem sítě zdola nahoru spočítáme vektory $\boldsymbol{a}_t$ a $\boldsymbol{o}_t$. Následně průchodem shora dolů spočítáme vektory $\boldsymbol{\delta}_t$. Potom už můžeme spočítat parciální derivace pomocí (4.1). Tudíž výpočet gradientu ztrátové funkce pomocí zpětné propagace můžeme popsat pomocí pseudokódu shrnutého v algoritmu 5.

Algoritmus 5: Zpětná propagace.

Vstup: Neuronová síť s architekturou (V, E, σ), vektor vah $\boldsymbol{w}$.
Vstup: Pozorování $(\boldsymbol{x}, \boldsymbol{y})$.
Inicializace: Pro každou vrstvu definuj matici vah $\mathbb{W}_t$ jako výše.
Dopředná propagace:
  Polož $\boldsymbol{o}_0 \leftarrow \boldsymbol{x}$.
  for *t = 1, ..., T* **do**
    Polož $\boldsymbol{a}_t \leftarrow \mathbb{W}_{t-1}\boldsymbol{o}_{t-1}$.
    Polož $\boldsymbol{o}_t \leftarrow \boldsymbol{\sigma}(\boldsymbol{a}_t)$.
  end
Zpětná propagace:
  Polož $\boldsymbol{\delta}_T \leftarrow \mathbb{J}_\ell(\boldsymbol{o}_T)$.
  for *t = T-1, ..., 1* **do**
    Polož $\boldsymbol{\delta}_t \leftarrow \boldsymbol{\delta}_{t+1}\,\mathsf{diag}(\boldsymbol{\sigma}'(\boldsymbol{a}_{t+1}))\mathbb{W}_t$.
  end
Výstup:
  foreach $(v_{t-1,j}, v_{t,i}) \in E$ **do**
    Polož parciální derivaci rovnu $\delta_{t,i}\sigma'(a_{t,i})o_{t-1,j}$.
  end

4.4 Metody

Nyní představíme neuronové sítě založené na robustních ztrátových funkcích inspirovaných předchozí kapitolou. Zatímco robustní neuronové sítě založené na LTS nebo LWS s jednoduchými volbami váhových funkcí již byly v literatuře studovány (Kalina a Vidnerová, 2020), zde nově představíme metody založené na LWS odhadu s adaptivními váhovými funkcemi (AW). Budeme uvažovat dopřednou neuronovou síť s pevnou architekturou, která definuje zobrazení

$$f(\,\cdot\,;\boldsymbol{w}) : \mathbb{R}^k \to \mathbb{R}.$$

Klasické neuronové sítě používané v regresních úlohách bývají založeny na metodě nejmenších čtverců, tedy

$$\hat{\boldsymbol{w}} = \operatorname*{argmin}_{\boldsymbol{w}} \frac{1}{n} \sum_{i=1}^{n} \Big(Y_i - f(\boldsymbol{X}_i;\boldsymbol{w})\Big)^2.$$

Jak bylo diskutováno v kontextu lineární regrese, LS odhad není robustní vůči odlehlým pozorováním a tedy je přirozené i v případě neuronových sítí nahradit ℓ_2 ztrátovou funkci nějakou robustnější variantou. Příklady robustních ztrátových funkcí implementovaných v knihovnách `TensorFlow` a `PyTorch` jsou ℓ_1 a Huberova ztrátová funkce.

Neuronové sítě založené na LTS odhadu hledají optimální váhy jako řešení optimalizačního problému

$$\hat{\boldsymbol{w}} = \operatorname*{argmin}_{\boldsymbol{w}} \frac{1}{\delta} \sum_{i=1}^{\delta} \Big(Y_i - f(\boldsymbol{X}_i;\boldsymbol{w})\Big)^2_{(i)},$$

kde $\delta = \lfloor \alpha n \rfloor$ pro nějaké $\frac{1}{2} \leq \alpha \leq 1$. Parametr α můžeme volit například pomocí křížové validace (sekce 1.3). Nakonec můžeme zobecnit LTS pomocí implicitního vážení podobně jako v případě lineární regrese, dostáváme

$$\hat{\boldsymbol{w}} = \operatorname*{argmin}_{\boldsymbol{w}} \frac{1}{n} \sum_{i=1}^{n} \psi\Big(\tfrac{2i-1}{2n}\Big) \Big(Y_i - f(\boldsymbol{X}_i;\boldsymbol{w})\Big)^2_{(i)}.$$

Poznámka. U diskutovaných metod můžeme také uvažovat nahrazení (pořádkových statistik) čtverců reziduí pomocí jejich absolutních hodnot. Například pro LTS dostaneme LTAD metodu, které se v kontextu neuronových sítí věnuje například Rusiecki (2013).

Také můžeme uvažovat dva parametry $0 < \tau_1 < \tau_2 < 1$ a pomocí neuronových sítí s kvantilovou ztrátovou funkcí spočítat odhady příslušných nelineárních regresních kvantilů. Následně můžeme natrénovat klasickou neuronovou síť s ℓ_2 ztrátovou funkcí na trénovacích datech, které leží nad (resp. pod) příslušným dolním (resp. horním) kvantilem. Nevýhodou je potom velká výpočetní náročnost (trénujeme celkem tři sítě). Podobně jako u LTS je potřeba vhodně zvolit parametry τ_1 a τ_2.

Poznámka. Při trénování neuronových sítí používáme zpětnou propagaci pro výpočet gradientu účelové funkce. Explicitně zde neuvádíme derivace uvažovaných ztrátových funkcí, neboť knihovny jako `TensorFlow` nebo `PyTorch` umožňují

automatic differentiation pomocí `tf.GradientTape`, resp. `torch.autograd` API. Derivace pro LTS a LWS odhady jsou k dispozici například v článku Kalina a Vidnerová (2020).

V této práci nás bude zajímat zejména metoda nejmenších adaptivně vážených čtverců,

$$\hat{\boldsymbol{w}} = \operatorname*{argmin}_{\boldsymbol{w}} \frac{1}{n} \sum_{i=1}^{n} \widehat{\psi}_n\Big(\tfrac{2i-1}{2n}\Big)\Big(Y_i - f(\boldsymbol{X}_i; \boldsymbol{w})\Big)^2_{(i)},$$

spolu s adaptivními váhovými funkcemi představenými v předchozí kapitole, tzn. binární váhovou funkcí (AWR, definice 25) a kvantilovou váhovou funkcí (AWQ, definice 26). Počáteční odhad obvykle volíme s vysokým bodem selhání, tedy například LTS s volbou $\delta = \lfloor \frac{n}{2} \rfloor + 1$. Jako počáteční odhad reziduální směrodatné odchylky se typicky volí

$$\widehat{\sigma}_n^0 = 1.483 \cdot \mathsf{med}_{1\le i\le n} |e_i^0 - \mathsf{med}_{1\le i\le n}\, e_i^0|,$$

kde e_i^0 jsou rezidua z počátečního odhadu.

Poznámka. V praxi se neuronové sítě často trénují na datech o velkých rozsazích výběru n a používá se polo-dávkový stochastický gradient, kde batch size $m \ll n$. S tím se pojí v kontextu robustních metod řada praktických problémů, se kterými se v lineární regresi nesetkáme a které nejsou často v literatuře diskutovány. V rámci stochastického gradientu se v každém kroku náhodně volí podmnožina dat, na základě které odhadujeme (nerobustně) gradient účelové funkce. V daném kroku však může být proporce odlehlých hodnot vyšší, než je v kompletních trénovacích datech, což může vychýlit odhad gradientu a model může ztratit robustní vlastnosti. Pokud tedy nemáme příliš mnoho pozorování, je vhodnější ve spojitosti s robustními ztrátovými funkcemi používat klasickou metodu největšího spádu. V případě LTS odhadu může pomoci nadhodnotit volbu parametru α, resp. δ. Větší problém nastává u adaptivních metod, které poměrně dobře odhadují skutečnou proporci odlehlých hodnot v trénovacích datech. Z tohoto důvodu v našich simulacích uvažujeme $m = n$, jelikož nám to naše rozsahy výběrů umožňují.

Další nevýhodou adaptivních metod je jejich závislost na počátečním odhadu. Je důležité, aby u počátečního odhadu nedošlo k přeučení, resp. nedoučení. V menší míře si však adaptivní metody dokáží poradit a podávají lepší výsledky než počáteční odhad, jak uvidíme v následujících simulacích. I přes to je však potřeba si v praxi dát pozor a u počátečního odhadu pečlivě volit příslušné hyperparametry.

5. Simulační studie

V této sekci porovnáme představené neuronové sítě pomocí Monte Carlo studie. Zaměříme se jak na predikční vlastnosti jednotlivých metod, tak na schopnosti detekovat odlehlá pozorování obsažená v trénovacích datech. Při výpočtech používáme programovací jazyk `Python` (verze 3.10.8) spolu se známými knihovnami `Numpy` (verze 1.23.5; Harris a kol. 2020) a `TensorFlow` (verze 2.10.0; Abadi a kol. 2015). Součástí práce je implementace všech představených (robustních) ztrátových funkcí a příslušných metrik. Ztrátové funkce dědí od třídy `tf.keras.losses.Loss`, metriky od `tf.keras.metrics.Mean` a tedy jsou připraveny pro trénování neuronových sítí s knihovnou `TensorFlow`. Implementovány jsou také všechny diskutované váhové funkce, jak deterministické ze sekce 3.5, tak adaptivní z podkapitoly 3.6.

5.1 Data

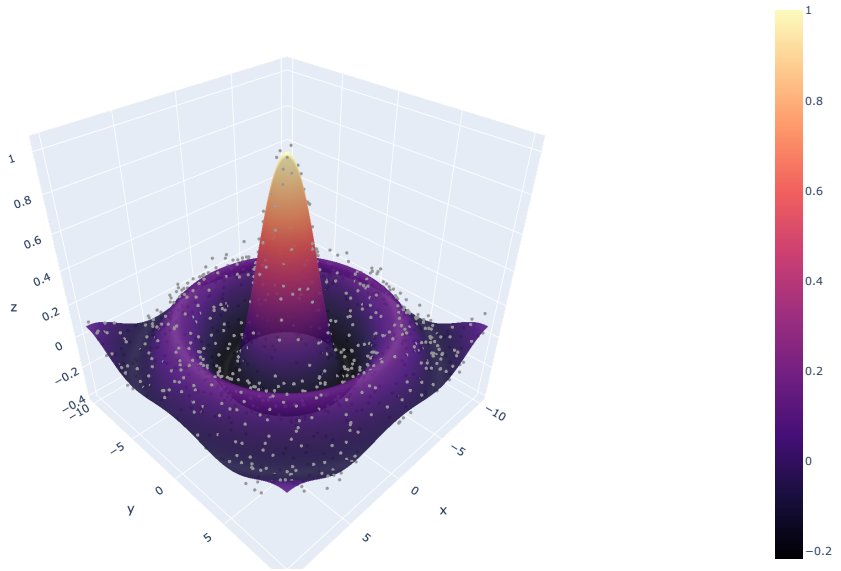
Předpokládejme, že skutečná regresní funkce je tvaru $\varphi(\boldsymbol{x}) = \frac{\sin(\|\boldsymbol{x}\|)}{\|\boldsymbol{x}\|}, \boldsymbol{x} \in \mathbb{R}^k$. Postupně budeme uvažovat $k \in \{1, 2, 10\}$ nezávisle proměnných, které budeme generovat z rovnoměrného rozdělení na intervalu $[-10, 10)$, tedy $X_{ij} \sim \mathsf{Unif}[-10, 10)$ pro $i = 1, \ldots, n$ a $j = 1, \ldots, k$. Rozsah výběru n uvažujeme v každé simulaci v závislosti na počtu regresorů k. Hodnotu odezvy určíme jako $Y_i = \varphi(\boldsymbol{X}_i) + \varepsilon_i$, kde $\varepsilon_i \sim \mathsf{N}(0, \sigma^2)$, $\sigma = 0.05$. Skutečnou regresní funkci a trénovací data pro $k = 2$ ilustruje obrázek 5.1. Následně kontaminujeme příslušný podíl $\delta \in \{0, 0.1\}$ pozorování následujícími způsoby.

Odlehlá pozorování

Pro $1-\delta$ pozorování uvažujeme chybové členy jako výše a zbylých δ pozorování kontaminujeme pomocí $\varepsilon_i \sim \mathsf{N}(5, 0.5^2)$.

Vzdálená pozorování

Každé z δ pozorování bude odlehlé alespoň v prvním regresoru, navíc zvolíme náhodně další regresory (každý s pravděpodobností 1/2), ve kterých bude pozorování také odlehlé. Příslušné hodnoty kontaminovaných regresorů volíme náhodně z rovnoměrného rozdělení na intervalu $[5, 20)$. Nakonec, vzdálená pozorování budou současně i odlehlými, tedy uvažujeme opět $Y_i = \varphi(\boldsymbol{X}_i) + \varepsilon_i$, kde $\varepsilon_i \sim \mathsf{N}(5, 0.5^2)$.



Obrázek 5.1. Skutečná regresní funkce $\varphi(\boldsymbol{x}) = \sin(\|\boldsymbol{x}\|)/\|\boldsymbol{x}\|$ pro případ $k = 2$ spolu s $n = 1200$ trénovacími daty (bez kontaminace). Osa x, resp. y odpovídá první, resp. druhé složce $\boldsymbol{x}$. Osa z potom odpovídá odezvě.

5.2 Metodologie

Budeme srovnávat predikční schopnosti, spolu se schopností správně identifikovat kontaminovaná pozorování v trénovacích datech. Zaměříme se na klasickou metodu nejmenších čtverců, Huberův odhad ($\delta = 1.345$), metodu nejmenších absolutních odchylek, LTS odhad (volbu $\alpha = 0.9$ budeme značit LTS1, volbu $\alpha = 0.8$ jako LTS2) a zejména na nejmenší adaptivně vážené čtverce, kde uvažujeme za počáteční odhad LTS (pro volby parametru α výše). V praxi je žádoucí volit počáteční odhad co nejvíce robustní. Nicméně, u neuronových sítí je důležité, aby u počátečního odhadu nedošlo k přeučení, případně nedoučení. Pro nižší hodnoty α měl LTS odhad tendence pomaleji konvergovat, proto jsme se uchýlili k tomuto kompromisu.

V každé simulaci uvažujeme pro všechny metody stejnou architekturu a pokud není uvedeno jinak, tak i stejný konstantní krok. Za optimalizační algoritmus byl zvolen Adam a batch size je vždy roven rozsahu trénovacích dat. Používáme early stopping, tedy každou metodu trénujeme do momentu, kdy po určitý počet epoch již nedochází ke zlepšení. Zejména nerobustní metody však měly v případě kontaminace tendenci se přeučovat, proto jsme v těchto případech ukončili trénování dříve.

V každé části simulační studie uvažujeme $m = 100$ Monte Carlo (MC) iterací, kdy vygenerujeme n trénovacích pozorování obsahujících šum, případně kontaminaci. Nezávisle na trénovacích datech vygenerujeme dalších n testovacích dat, tentokrát bez šumu a kontaminace. Tedy regresory generujeme z $\mathsf{Unif}[-10, 10)$ a následně pro ně dopočítáme skutečnou hodnotu odezvy φ. Pro každou MC ite-

raci trénujeme metody na trénovacích datech a následně na testovacích datech spočítáme RMSE,

$$\mathsf{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}.$$

Ve výsledcích potom uvádíme průměrnou RMSE spočítanou přes všechny MC iterace spolu s příslušným odhadem MC chyby.

Pomocí každé metody se navíc pokusíme detekovat kontaminovaná pozorování následovně. Nejprve spočítáme odhad reziduální směrodatné odchylky jako

$$\widehat{\sigma}_n = \sqrt{\frac{1}{n-k}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$

pro metodu nejmenších čtverců, resp. jako

$$\widehat{\sigma}_n = 1.4826 \cdot \mathsf{med}_{1\le i\le n}\,|\widehat{\varepsilon}_i - \mathsf{med}_{1\le i\le n}\,\widehat{\varepsilon}_i|$$

pro robustní metody s rezidui $\widehat{\varepsilon}_i$. Následně klasifikujeme i-té pozorování jako odlehlé, pokud $|\widehat{\varepsilon}_i/\widehat{\sigma}_n| > 2.5$ a pro každou metodu spočítáme přesnost a F_1-skóre (viz dodatek A.5). V našem případě, kdy jsou počty dobrých a odlehlých pozorování nevyvážené, je F_1-skóre vhodnější metrikou. Ve výsledcích opět uvádíme průměrné hodnoty těchto metrik spočítané přes všechny MC iterace.

Na závěr poznamenejme, že neuronové sítě jsou vysoce ovlivněny počáteční volbou vah. Z tohoto důvodu inicializujeme počáteční váhy v každé MC iteraci jiným způsobem (ale stejným pro všechny metody v rámci jedné iterace). Pro každou vrstvu neuronové sítě generujeme počáteční váhy z $\mathsf{N}(0, \sigma^2)$ rozdělení, kde $\sigma^2 = 2/(m_1 + m_2)$, m_1 je počet vstupů a m_2 počet výstupů dané vrstvy, viz Glorot a Bengio (2010).

5.3 Výsledky

Dvourozměrný případ

Výsledky simulace pro jediný regresor a 400 pozorování shrnuje tabulka 5.1. V tomto případě uvažujeme neuronové sítě s jedinou skrytou vrstvou s 20 neurony a hyperbolický tangens jako aktivační funkci. Jak bychom očekávali, v případě bez kontaminovaných dat se zdá být optimální metoda nejmenších čtverců. LAD metoda lehce zaostává, podobně LTS, zejména pokud zanedbáme hodně pozorování. Jak jsme diskutovali v kontextu lineární regrese, kvantilové adaptivní váhy jsou zkonstruovány tak, aby garantovaly vysokou relativní eficienci pro normálně rozdělené chyby. Podobný trend pozorujeme i v nelineární regresi, kdy se zdá být metoda AWQ srovnatelná s metodou nejmenších čtverců. Totéž platí pro metodu AWR, která zanedbá pouze adaptivně zvolený počet pozorování. Všimněme si, že ačkoli u metody LTS2 pozorujeme potíže s konvergencí, adaptivní metody s tímto počátečním odhadem jsou nadále srovnatelné s LS. Přesto je však důležité si dát u počátečních odhadů obzvlášť pozor při volbě hyperparametrů, neboť počáteční odhad může výrazně ovlivnit chování adaptivních metod. Predikční funkce pro jednu z prvních MC iterací jsou vizualizovány na obrázku 5.2.

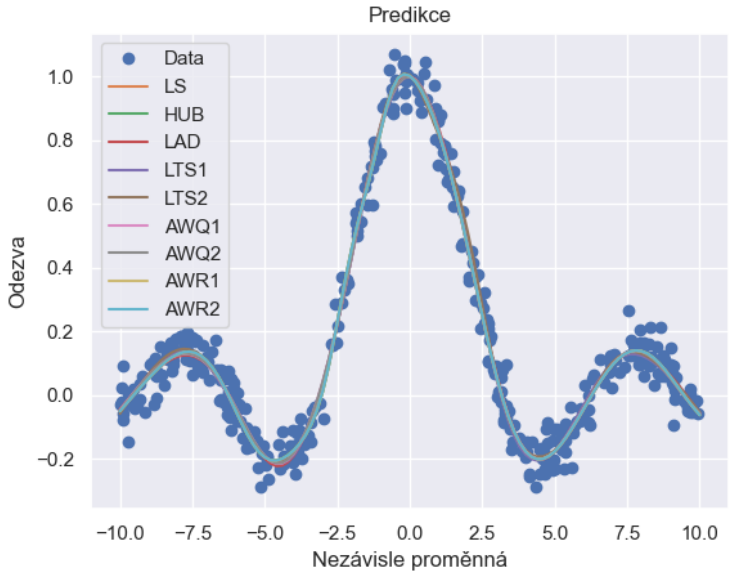
V případě odlehlých pozorování začínají mít LS a Huberův odhad značné problémy s predikcí. Odhad směrodatné odchylky založený na LS je sice vychýlený, ale v porovnání s Huberovým odhadem metoda nejmenších čtverců relativně zvládá detekovat kontaminovaná pozorování, ačkoliv ne tak dobře, jako robustní metody. Všimněme si, že přestože přesnost je u Huberovy regrese relativně vysoká, F_1-skóre je podstatně nižší v porovnání s ostatními robustními odhady. Metoda nejmenších absolutních odchylek se zdá být lehce horší než ostatní robustní přístupy, což může být způsobeno pomalejší konvergencí. Zajímavé je sledovat, že LTS si oproti nekontaminovanému případu polepšila. I zde se zdají být adaptivní metody nejvhodnější a z hlediska RMSE dávají výsledky srovnatelné s případem bez kontaminace, avšak za cenu vyšší výpočetní náročnosti, spojenou zejména s počátečním odhadem. Také si všimněme, že LTS1 metoda, která zanedbá správný počet odlehlých pozorování patří mezi nejlepší. Predikce pro tento případ ilustruje obrázek 5.3.

U vzdálených pozorování se daří nejmenším čtvercům z pohledu predikce podobně, ale začínají mít větší problémy s diagnostikou, zejména F_1-skóre je nejhorší ze všech prezentovaných metod. Zhoršení také pozorujeme u Huberova odhadu a LAD, což není překvapující, neboť se nejedná o odhady robustní vzhledem ke vzdáleným pozorováním. Metoda LAD začíná mít značené problémy také s identifikací kontaminovaných dat. Nejlepší se zdají být metody, které přiřazují kontaminovaným pozorováním nulové váhy, viz LTS1 a AWR. Naopak metoda AWQ, která nezanedbává žádná pozorování, má o něco větší RMSE. Tento trend pozorujeme také v lineární regresi, viz simulační studie v článku Čížek (2011). Jak se daří jednotlivým metodám aproximovat skutečnou regresní funkci je znázorněno na obrázku 5.4.

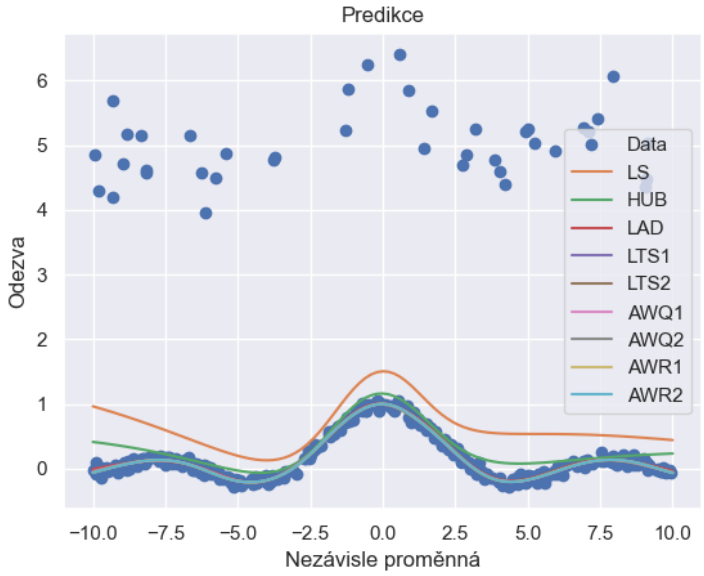
Jakým způsobem penalizují adaptivní metody kontaminovaná pozorování se snaží ilustrovat obrázek 5.5, jedná se o případ s počátečním odhadem LTS1. Pozorujeme, že v tomto konkrétním případě se oběma metodám podařilo spolehlivě identifikovat všechna odlehlá i vzdálená pozorování. Zároveň se nezdá, že by zbytečně penalizovaly běžná pozorování. Přesto, že v tabulce 5.1 pozorujeme zhoršení počátečního odhadu LTS2 oproti LTS1, obrázky pro tento počáteční odhad by vypadaly obdobně. To odpovídá i faktu, že ve zmíněné tabulce nepozorujeme výrazné rozdíly pro oba počáteční odhady.

Metrika	LS	Huber ($\delta = 1.345$)	LAD	LTS ($\alpha = 0.9$)	LTS ($\alpha = 0.8$)	AWQ ($\alpha = 0.9$)	AWQ ($\alpha = 0.8$)	AWR ($\alpha = 0.9$)	AWR ($\alpha = 0.8$)
400 pozorování, bez kontaminace:									
RMSE	0.011 (0.0002)	0.011 (0.0002)	0.016 (0.0005)	0.017 (0.0010)	0.030 (0.0016)	0.011 (0.0004)	0.011 (0.0002)	0.011 (0.0002)	0.011 (0.0002)
Sigma	0.049 (0.0001)	0.049 (0.0003)	0.049 (0.0003)	0.049 (0.0003)	0.049 (0.0003)	0.050 (0.0003)	0.049 (0.0003)	0.049 (0.0003)	0.049 (0.0003)
Přesnost	0.989 (0.0005)	0.988 (0.0008)	0.983 (0.0010)	0.980 (0.0012)	0.964 (0.0019)	0.987 (0.0009)	0.986 (0.0009)	0.988 (0.0008)	0.987 (0.0008)
F_1-skóre	—	—	—	—	—	—	—	—	—
400 pozorování, 10 % odlehlých:									
RMSE	0.546 (0.0032)	0.181 (0.0012)	0.020 (0.0005)	0.011 (0.0002)	0.020 (0.0012)	0.012 (0.0003)	0.012 (0.0002)	0.011 (0.0002)	0.012 (0.0002)
Sigma	1.498 (0.0027)	0.120 (0.0018)	0.057 (0.0003)	0.056 (0.0003)	0.056 (0.0003)	0.056 (0.0003)	0.055 (0.0003)	0.056 (0.0003)	0.056 (0.0003)
Přesnost	0.990 (0.0004)	0.922 (0.0045)	0.993 (0.0005)	0.996 (0.0004)	0.990 (0.0009)	0.995 (0.0005)	0.995 (0.0005)	0.996 (0.0003)	0.996 (0.0003)
F_1-skóre	0.948 (0.0024)	0.737 (0.0115)	0.968 (0.0024)	0.983 (0.0017)	0.953 (0.0040)	0.976 (0.0023)	0.974 (0.0023)	0.983 (0.0016)	0.983 (0.0016)
400 pozorování, 10 % vzdálených:									
RMSE	0.433 (0.0079)	0.257 (0.0061)	0.075 (0.0065)	0.011 (0.0002)	0.020 (0.0012)	0.014 (0.0004)	0.014 (0.0004)	0.012 (0.0002)	0.012 (0.0002)
Sigma	0.929 (0.0086)	0.115 (0.0016)	0.056 (0.0003)	0.056 (0.0003)	0.056 (0.0003)	0.056 (0.0003)	0.056 (0.0003)	0.056 (0.0003)	0.056 (0.0003)
Přesnost	0.937 (0.0008)	0.912 (0.0014)	0.959 (0.0009)	0.996 (0.0004)	0.990 (0.0009)	0.995 (0.0005)	0.994 (0.0006)	0.996 (0.0004)	0.996 (0.0003)
F_1-skóre	0.539 (0.0078)	0.647 (0.0038)	0.796 (0.0045)	0.983 (0.0017)	0.953 (0.0040)	0.976 (0.0022)	0.971 (0.0026)	0.983 (0.0017)	0.983 (0.0016)

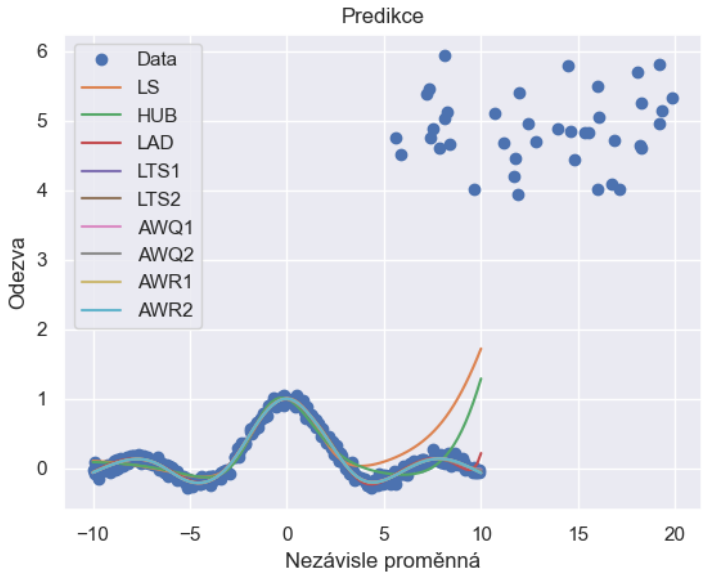
Tabulka 5.1. Výsledky simulace pro $k = 1$ a $n = 400$ pozorování. Uvedeny jsou průměrné hodnoty metrik pro $m = 100$ Monte Carlo iterací, včetně příslušných odhadů Monte Carlo chyb. Neuvedené F_1-skóre znamená, že není v daném případě definováno. U adaptivních metod se uvedená hodnota α vztahuje k počátečnímu LTS odhadu.



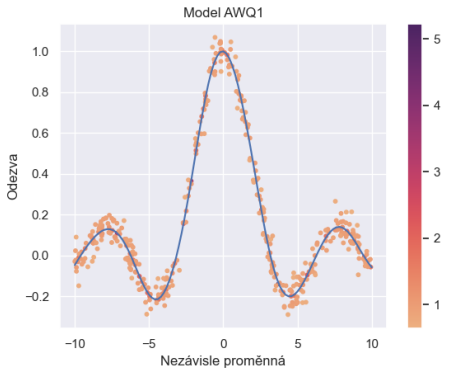
Obrázek 5.2. Predikce daných metod spolu s trénovacími daty, případ $k = 1$, $n = 400$ pozorování, bez kontaminace. Výsledky z jedné z prvních MC iterací. Pozorujeme, že v tomto případě všechny metody přesně aproximují skutečnou regresní funkci.



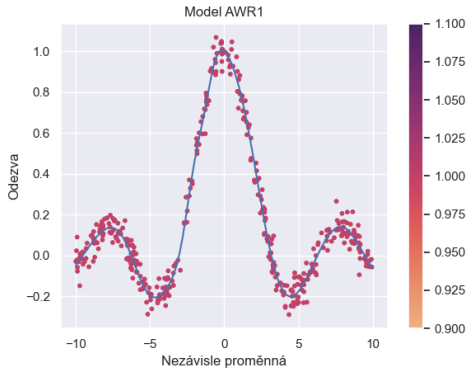
Obrázek 5.3. Trénovací data a predikce jednotlivých metod pro případ s jediným regresorem, 400 pozorováními a 10 % odlehlých hodnot. U metody nejmenších čtverců a Huberovy regrese pozorujeme vychýlení způsobené odlehlými pozorováními.



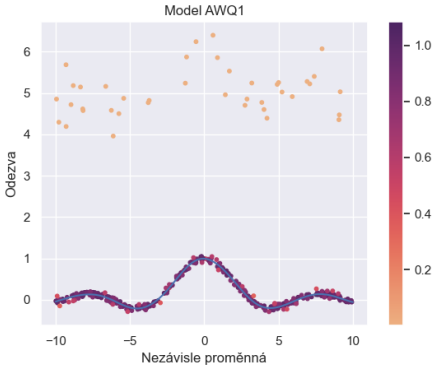
Obrázek 5.4. Predikce a trénovací data v případě kontaminace vzdálenými pozorováními. Predikční funkce uvažujeme pouze na $[-10, 10]$, neboť trénovací data neobsahují žádné běžné pozorování pro regresory s hodnotami v $[10, 20]$. Pozorujeme, že LAD není robustní vzhledem ke kontaminaci v regresorech.



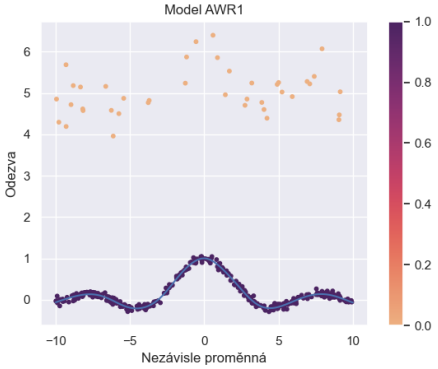
(a) AW-Q ($\alpha = 0.9$)



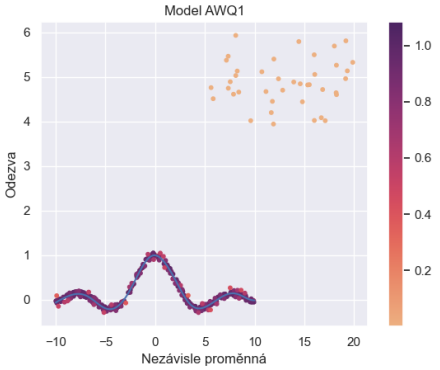
(b) AW-R ($\alpha = 0.9$)



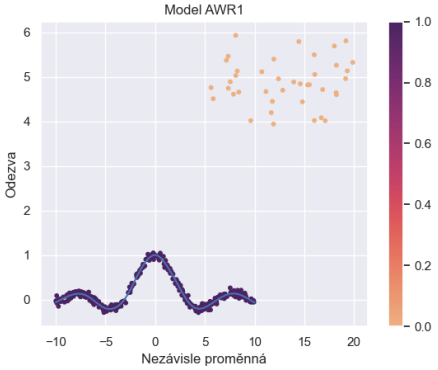
(c) AW-Q ($\alpha = 0.9$)



(d) AW-R ($\alpha = 0.9$)



(e) AW-Q ($\alpha = 0.9$)



(f) AW-R ($\alpha = 0.9$)

Obrázek 5.5. Vizualizace přiřazených vah adaptivními metodami jednotlivým trénovacím datům spolu s predikcemi pro případ $k = 1$ a $n = 400$. Jedná se o výsledky z některé z prvních MC iterací.

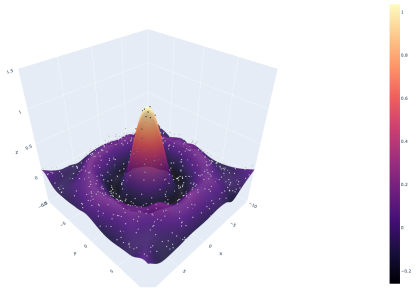
Třírozměrný případ

Druhá simulace se věnuje situaci se dvěma nezávisle proměnnými a 1200 trénovacími daty. Uvažujeme sítě se dvěma skrytými vrstvami, každou s 20 neurony a hyperbolický tangens jako aktivaci. Výsledky jsou k dispozici v tabulce 5.2, obecně můžeme říct, že odpovídají předchozí části pro $k = 1$. Bez kontaminace se zdá být nejvhodnější metoda nejmenších čtverců a Huberova regrese, ačkoli odhady reziduální směrodatné odchylky se zdají být o něco horší než pro ostatní metody. Podobně jako v předchozí simulaci, LTS2 odhad, který zanedbává zbytečně mnoho pozorování se jeví jako nejslabší. Větší MC chyba v porovnání s ostatními odhady napovídá, že má pro některé volby počátečních vah potíže konvergovat. Opět si všimněme, že adaptivní metody si dokáží oproti počátečnímu odhadu LTS2 poměrně polepšit.

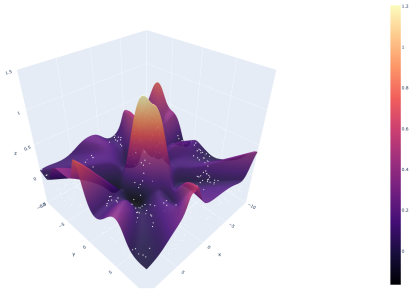
Při kontaminaci odlehlými pozorováními se všechny robustní metody chovají rozumně, až na Huberův odhad, který modeluje něco mezi LS a LAD. Pro oba druhy kontaminace se chovají nejhůře LS a Huber, pozorujeme vysokou RMSE, odhady σ jsou vychýlené a potíže mají také s diagnostikou. Zejména adaptivní metody a LTS1 však dávají výsledky srovnatelné s daty bez kontaminace, podobně pro vzdálená pozorování. Opět pozorujeme nerobustnost LAD vzhledem ke vzdáleným pozorováním. Predikce vybraných metod pro různé druhy kontaminace ilustruje obrázek 5.6.

Metrika	LS	Huber ($\delta = 1.345$)	LAD	LTS ($\alpha = 0.9$)	LTS ($\alpha = 0.8$)	AWQ ($\alpha = 0.9$)	AWQ ($\alpha = 0.8$)	AWR ($\alpha = 0.9$)	AWR ($\alpha = 0.8$)
1200 pozorování, bez kontaminace:									
RMSE	0.027 (0.0006)	0.027 (0.0005)	0.033 (0.0004)	0.033 (0.0005)	0.054 (0.0035)	0.027 (0.0003)	0.029 (0.0004)	0.026 (0.0003)	0.028 (0.0004)
Sigma	0.046 (0.0003)	0.046 (0.0004)	0.050 (0.0002)	0.050 (0.0002)	0.049 (0.0005)	0.051 (0.0002)	0.050 (0.0002)	0.050 (0.0003)	0.051 (0.0002)
Přesnost	0.987 (0.0003)	0.986 (0.0004)	0.970 (0.0007)	0.962 (0.0008)	0.935 (0.0015)	0.982 (0.0005)	0.978 (0.0006)	0.986 (0.0004)	0.982 (0.0006)
F_1-skóre	—	—	—	—	—	—	—	—	—
1200 pozorování, 10 % odlehlých:									
RMSE	0.664 (0.0046)	0.225 (0.0029)	0.038 (0.0004)	0.030 (0.0006)	0.035 (0.0005)	0.028 (0.0004)	0.030 (0.0004)	0.030 (0.0007)	0.028 (0.0005)
Sigma	1.441 (0.0027)	0.158 (0.0012)	0.058 (0.0003)	0.052 (0.0005)	0.056 (0.0003)	0.057 (0.0003)	0.057 (0.0003)	0.052 (0.0005)	0.056 (0.0005)
Přesnost	0.980 (0.0004)	0.951 (0.0013)	0.985 (0.0005)	0.994 (0.0003)	0.979 (0.0006)	0.991 (0.0003)	0.989 (0.0003)	0.995 (0.0003)	0.995 (0.0002)
F_1-skóre	0.886 (0.0025)	0.805 (0.0043)	0.932 (0.0021)	0.973 (0.0012)	0.905 (0.0024)	0.959 (0.0013)	0.950 (0.0015)	0.974 (0.0012)	0.974 (0.0011)
1200 pozorování, 10 % vzdálených:									
RMSE	0.352 (0.0057)	0.202 (0.0052)	0.134 (0.0094)	0.030 (0.0007)	0.036 (0.0005)	0.028 (0.0004)	0.030 (0.0004)	0.030 (0.0007)	0.028 (0.0005)
Sigma	0.725 (0.0061)	0.129 (0.0009)	0.054 (0.0003)	0.052 (0.0005)	0.056 (0.0003)	0.056 (0.0003)	0.056 (0.0003)	0.052 (0.0005)	0.056 (0.0004)
Přesnost	0.992 (0.0004)	0.934 (0.0009)	0.948 (0.0007)	0.994 (0.0003)	0.979 (0.0006)	0.992 (0.0003)	0.989 (0.0003)	0.994 (0.0003)	0.995 (0.0002)
F_1-skóre	0.383 (0.0048)	0.658 (0.0040)	0.732 (0.0035)	0.973 (0.0012)	0.905 (0.0024)	0.960 (0.0014)	0.948 (0.0015)	0.973 (0.0012)	0.974 (0.0011)

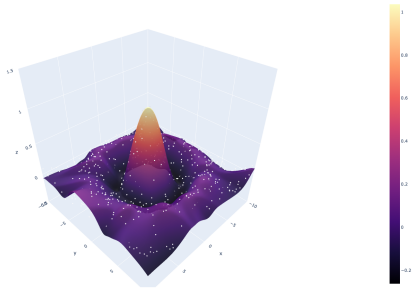
Tabulka 5.2. Průměrné hodnoty metrik pro případ $k = 2$, $n = 1200$ pozorování a $m = 100$ Monte Carlo iterací, včetně příslušných odhadů Monte Carlo chyb. F_1-skóre pro data bez kontaminace nejsou definována. Hodnota α u adaptivních metod značí počáteční LTS odhad.



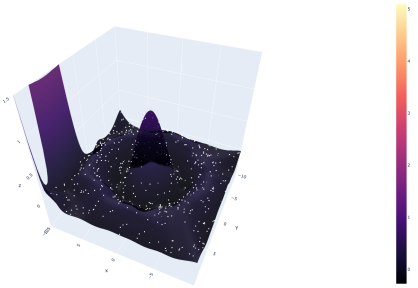
(a) LS, data bez kontaminace



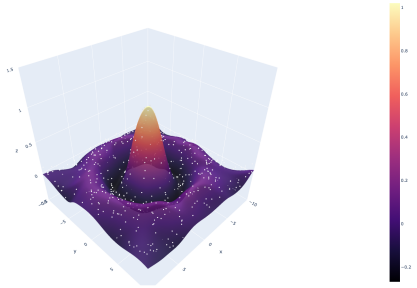
(b) Huber, odlehlá pozorování



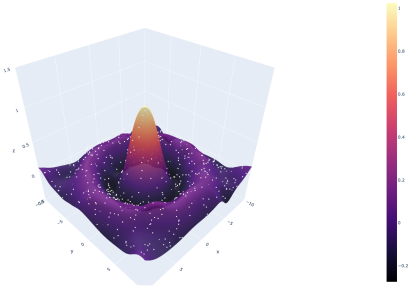
(c) LAD, odlehlá pozorování



(d) LAD, vzdálená pozorování



(e) AWQ1, odlehlá pozorování



(f) AWR1, vzdálená pozorování

Obrázek 5.6. Predikce vybraných metod spolu s testovacími daty pro různé druhy kontaminace, případ $k = 2$ a $n = 1200$. Výsledky z některé z prvních MC iterací. Osy x, resp. y odpovídají první, resp. druhé složce $\boldsymbol{x}$. Osy z potom odpovídají odezvě.

Vícerozměrný případ

Nakonec uvažujeme případ s 10 regresory a $n = 2000$ pozorování. Architektura sítí je volena stejná jako v předchozím případě pro $k = 2$. Výsledky shrnuje tabulka 5.3. Bez kontaminace se opět zdá LS spolu s Huberovým odhadem jako nejoptimálnější, jak z hlediska predikce tak diagnostiky, přičemž adaptivní metody jsou stále velmi blízko. Zejména u LTS2 pozorujeme ve vyšších dimenzích horší přesnost. V případě kontaminace jsou metoda nejmenších čtverců i Huberova regrese velmi nespolehlivé. Všimněme si, že zatímco v nižších dimenzích jsme u LS pozorovali relativně vysoké F_1-skóre v případě odlehlých pozorování, pro vícerozměrný případ tomu tak není. Také pozorujeme výraznější zhoršení F_1-skóre u LAD, LTS2 a AWQ2. Nejvhodnější se potom zdá být použití adaptivních metod a LTS s vhodnou volbou parametru α, přičemž praktická nevýhoda LTS spočívá právě v nutnosti zvolit tento hyperparametr. Pokud je α příliš malé, ztrácíme robustnost a pokud naopak zanedbáme příliš mnoho pozorování, pozorujeme horší konvergenci, jako v případě LTS2.

Metrika	LS	Huber ($\delta = 1.345$)	LAD	LTS ($\alpha = 0.9$)	LTS ($\alpha = 0.8$)	AWQ ($\alpha = 0.9$)	AWQ ($\alpha = 0.8$)	AWR ($\alpha = 0.9$)	AWR ($\alpha = 0.8$)
2000 pozorování, bez kontaminace:									
RMSE	0.054 (0.0002)	0.054 (0.0002)	0.057 (0.0002)	0.061 (0.0003)	0.068 (0.0004)	0.055 (0.0002)	0.056 (0.0002)	0.055 (0.0002)	0.057 (0.0002)
Sigma	0.057 (0.0002)	0.057 (0.0002)	0.044 (0.0004)	0.051 (0.0003)	0.047 (0.0003)	0.053 (0.0003)	0.049 (0.0003)	0.056 (0.0002)	0.055 (0.0002)
Přesnost	0.989 (0.0002)	0.989 (0.0003)	0.924 (0.0019)	0.930 (0.0009)	0.876 (0.0016)	0.970 (0.0008)	0.950 (0.0014)	0.985 (0.0004)	0.973 (0.0008)
F_1-skóre	—	—	—	—	—	—	—	—	—
2000 pozorování, 10 % odlehlých:									
RMSE	1.617 (0.0056)	1.407 (0.0076)	0.066 (0.0004)	0.056 (0.0002)	0.064 (0.0003)	0.056 (0.0002)	0.058 (0.0002)	0.056 (0.0002)	0.056 (0.0002)
Sigma	0.817 (0.0025)	0.567 (0.0026)	0.051 (0.0004)	0.064 (0.0003)	0.057 (0.0003)	0.061 (0.0003)	0.056 (0.0003)	0.064 (0.0003)	0.063 (0.0003)
Přesnost	0.915 (0.0003)	0.938 (0.0007)	0.954 (0.0014)	0.996 (0.0002)	0.948 (0.0009)	0.990 (0.0003)	0.972 (0.0009)	0.996 (0.0002)	0.994 (0.0003)
F_1-skóre	0.323 (0.0036)	0.629 (0.0048)	0.817 (0.0044)	0.983 (0.0008)	0.795 (0.0029)	0.951 (0.0014)	0.877 (0.0032)	0.983 (0.0008)	0.969 (0.0013)
2000 pozorování, 10 % vzdálených:									
RMSE	0.435 (0.0066)	0.406 (0.0066)	0.284 (0.0043)	0.056 (0.0002)	0.064 (0.0003)	0.056 (0.0002)	0.058 (0.0003)	0.056 (0.0002)	0.056 (0.0002)
Sigma	0.217 (0.0044)	0.106 (0.0020)	0.062 (0.0005)	0.064 (0.0003)	0.057 (0.0003)	0.061 (0.0003)	0.056 (0.0003)	0.064 (0.0003)	0.063 (0.0003)
Přesnost	0.911 (0.0005)	0.914 (0.0008)	0.907 (0.0011)	0.997 (0.0002)	0.948 (0.0009)	0.990 (0.0003)	0.972 (0.0008)	0.997 (0.0002)	0.994 (0.0003)
F_1-skóre	0.261 (0.0062)	0.423 (0.0067)	0.559 (0.0044)	0.983 (0.0008)	0.794 (0.0029)	0.951 (0.0016)	0.879 (0.0031)	0.983 (0.0008)	0.969 (0.0013)

Tabulka 5.3. Výsledky pro 10 nezávisle proměnných, $n = 2000$ pozorování a $m = 100$ MC iterací, včetně odhadů odpovídajících MC chyb. F_1-skóre pro data bez kontaminace nejsou definována. Hodnota α u adaptivních metod indikuje počáteční LTS odhad.

Závěr

V první kapitole jsme formulovali úlohu statistického učení a představili princip minimalizace rizika. Nastínili jsme, jaká úskalí s sebou nese minimalizace empirického rizika a diskutovali, jak pomocí křížové validace volit optimální hyperparametry. Nakonec jsme představili stochastický gradient, který se běžně používá při trénování neuronových sítí. Druhá kapitola se věnuje klasické metodě nejmenších čtverců a kvantilové regresi. Třetí kapitola představuje vybrané robustní odhady regresních koeficientů a diskutuje jejich vlastnosti, v některých případech za obecnějších předpokladů, kdy uvažujeme jistou formu závislosti v datech. Potřebné teoretické výsledky jsme krátce shrnuli v dodatku. Hlavními výsledky teoretické části práce je formulace a zdůvodnění asymptotické normality LTS odhadu pomocí výsledků pro LWS odhad. Podrobněji jsme rozepsali některé kroky důkazu věty 15, která se zabývá konzistentním odhadem asymptotické varianční matice LWS odhadu. S pomocí asymptotické normality jsme odvodili testy a konfidenční intervaly (množiny) pro regresní koeficienty založené na LWS metodě, které zůstávají platné i pro adaptivně vážené odhady. Nakonec jsme dokázali, že za normality mají odhady regresních koeficientů AW metodou a metodou nejmenších čtverců asymptoticky stejná rozdělení.

Ve čtvrté kapitole jsme definovali dopřednou neuronovou síť jako orientovaný acyklický graf a diskutovali, jak se při trénování počítá gradient ztrátové funkce pomocí zpětné propagace. Také jsme představili robustní neuronové sítě inspirované třetí kapitolou a diskutovali některé praktické problémy. Dalším vlastním přínosem je menší simulační studie, ve které jsme porovnali představené robustní sítě jak z hlediska predikčních schopností, tak jejich možné využití jako diagnostických nástrojů.

Simulační studie ukázala, že klasická metoda nejmenších čtverců je vhodnou volbou pro nekontaminovaná data, ale selhává v případě kontaminace. Značné problémy má také s detekováním kontaminovaných pozorování obsažených v trénovacích datech. Robustní alternativy dostupné v knihovnách jako `TensorFlow` a `PyTorch`, konkrétně Huberova a ℓ_1 ztrátová funkce, však řeší problém jen částečně, zejména kvůli jejich náchylnosti na vzdálená pozorování. V tomto případě se ukazuje síla odhadů s vysokým bodem selhání, známých z literatury o robustní regresi. Zejména nejmenší ořezané čtverce patřily mezi nejlepší, ale v praxi zůstává otázkou, jak volit optimální hodnotu parametru α. Robustní odhady jsou navíc často neefektivní, pokud trénovací data kontaminována nejsou. Jako možné řešení se ukázaly být adaptivně vážené robustní odhady, kterým prozatím v kontextu neuronových sítí nebyla věnována dostatečná pozornost. Pokud tedy nevíme, zda a případně jaký druh kontaminace naše trénovací data obsahují, můžeme na základě provedené simulační studie doporučit adaptivně vážené metody, které disponovaly vysokou robustností vzhledem k jak odlehlým, tak vzdáleným pozorováním a současně byly pro čistá data srovnatelné s metodou nejmenších čtverců. Jejich nevýhoda potom spočívá v nutnosti trénovat další neuronovou síť, zejména je potřeba pečlivě volit příslušné hyperparametry. Součástí práce je také implementace všech představených ztrátových a váhových funkcí, včetně příslušných metrik, které jsou připraveny pro trénování neuronových sítí s využitím knihovny `TensorFlow`.

Seznam použité literatury

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y. a Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. https://www.tensorflow.org/.

Andrews, D. W. K. (1988). Laws of large numbers for dependent non-identically distributed random variables. *Econometric Theory*, **4**(3), 458–467. ISSN 02664666, 14694360.

Bottou, L., Curtis, F. a Nocedal, J. (2018). Optimization methods for large-scale machine learning. *SIAM Review*, **60**(2), 223–311.

Čížek, P. (2004). Asymptotics of least trimmed squares regression. *Tilburg University, Center for Economic Research, Discussion Paper*. doi: 10.2139/ssrn.606982.

Čížek, P. (2007). Efficient robust estimation of regression models. *SSRN Electronic Journal*. doi: 10.2139/ssrn.888685.

Čížek, P. (2011). Semiparametrically weighted robust estimation of regression models. *Computational Statistics & Data Analysis*, **55**(1), 774–788.

Davidson, J. (1994). *Stochastic Limit Theory: An Introduction for Econometricians*. Advanced Texts in Econometrics. Oxford University Press. ISBN 9780191525049.

Gervini, D. a Yohai, V. J. (2002). A class of robust and fully efficient regression estimators. *The Annals of Statistics*, **30**(2), 583 – 616. doi: 10.1214/aos/1021379866.

Glorot, X. a Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR.

Goodfellow, I., Bengio, Y. a Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., Fernández del Río, J., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W.,

Abbasi, H., Gohlke, C. a Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, **585**(7825), 357–362. doi: 10.1038/s41586-020-2649-2.

Huber, P. J. a Ronchetti, E. M. (2009). *Robust Statistics*. Wiley Series in Probability and Statistics. Wiley. ISBN 9780470434680.

Hubert, M., Rousseeuw, P. J. a Van Aelst, S. (2008). High-breakdown robust multivariate methods. *Statistical Science*, **23**(1), 92 – 119. doi: 10.1214/088342307000000087.

Janáček, P. (2020). Optimalizační metody prvního řádu v úlohách strojového učení. Bakalářská práce, Univerzita Karlova, Praha.

Kalina, J. (2012). Implicitly weighted methods in robust image analysis. *Journal of Mathematical Imaging and Vision*, **44**(3), 449–462. ISSN 0924-9907. doi: 10.1007/s10851-012-0337-z.

Kalina, J. (2015). Three contributions to robust regression diagnostics. *Journal of Applied Mathematics, Statistics and Informatics*, **11**(2), 69–78. doi: 10.1515/jamsi-2015-0013.

Kalina, J. a Tichavský, J. (2020). On robust estimation of error variance in (highly) robust regression. *Measurement Science Review*, **20**, 6–14. doi: 10.2478/msr-2020-0002.

Kalina, J. a Vidnerová, P. (2020). Robust multilayer perceptrons: Robust loss functions and their derivatives. In *Proceedings of the 21st EANN (Engineering Applications of Neural Networks) 2020 Conference*, pages 546–557, Cham, 2020. Springer International Publishing. ISBN 978-3-030-48791-1.

Kidger, P. a Lyons, T. (2020). Universal approximation with deep narrow networks. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2306–2327. PMLR.

Koenker, R. (2005). *Quantile Regression*. Econometric Society Monographs. Cambridge University Press. doi: 10.1017/CBO9780511754098.

Komárek, A. (2021). NMSA 407 Linear regression, Course notes.

Kulich, M. (2021). NMST 432 Advanced regression models, Course notes.

Maronna, R. . A., Martin, D. R. a Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*. Wiley Series in Probability and Statistics. Wiley. ISBN 9780470010921.

Omelka, M. (2021). NMST 434 Modern statistical methods, Course notes.

Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, **7**(2), 186–199. doi: 10.1017/S0266466600004394.

Rousseeuw, P. J. a Driessen, K. V. (2006). Computing lts regression for large data sets. *Data Mining and Knowledge Discovery*, **12**, 29–45. doi: 10.1007/s10618-005-0024-4.

Rousseeuw, P. J. a Hubert, M. (2011). Robust statistics for outlier detection. *WIREs Data Mining Knowl. Discov.*, **1**(1), 73–79. doi: 10.1002/widm.2.

Rousseeuw, P. J. a Leroy, A. M. (1987). *Robust regression and outlier detection*. Wiley series in probability and mathematical statistics. Applied probability and statistics. Wiley. ISBN 0-471-85233-3.

Rusiecki, A. (2013). Robust learning algorithm based on lta estimator. *Neurocomputing*, **120**, 624–632. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2013.04.008. Image Feature Detection and Description.

Shalev-Shwartz, S. a Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 1 edition. ISBN 9781107057135.

Vapnik, V. (1991). Principles of risk minimization for learning theory. *NIPS*, **4**, 831–838.

Vapnik, V. (2000). *The Nature of Statistical Learning Theory*. Statistics for Engineering and Information Science. Springer New York. ISBN 0-387-98780-0.

Víšek, J. (2002). The least weighted squares II. consistency and asymptotic normality. *Bulletin of the Czech Econometric Society*, **9**(16).

Víšek, J. (2006a). The least trimmed squares. Part i: Consistency. *Kybernetika*, **42**(1), 1–36.

Víšek, J. (2006b). The least trimmed squares. Part iii: Asymptotic normality. *Kybernetika*, **42**(2), 203–224.

Víšek, J. (2010). Robust error-term-scale estimate. *Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in honor of Professor Jana Jurečková*, pages 254–267.

Víšek, J. (2011). Consistency of the least weighted squares under heteroscedasticity. *Kybernetika*, **47**(2), 179–206.

Yan, X. a Su, X. (2009). *Linear Regression Analysis: Theory And Computing*. World Scientific. ISBN 9789814470087.

A. Přílohy

A.1 Ekvivarianční vlastnosti regresních odhadů

Definice 27. *Odhad regresních koeficientů* $\widehat{\boldsymbol{\beta}}_n(\boldsymbol{Y}, \mathbb{X})$ *je* **ekvivariantní vzhledem k regresi**, *pokud pro libovolný vektor* $\boldsymbol{a} \in \mathbb{R}^k$ *platí*

$$\widehat{\boldsymbol{\beta}}_n(\boldsymbol{Y} + \mathbb{X}\boldsymbol{a}, \mathbb{X}) = \widehat{\boldsymbol{\beta}}_n(\boldsymbol{Y}, \mathbb{X}) + \boldsymbol{a}.$$

Definice 28. *Řekneme, že odhad regresních koeficientů* $\widehat{\boldsymbol{\beta}}_n(\boldsymbol{Y}, \mathbb{X})$ *je* **ekvivariantní vzhledem k měřítku**, *pokud pro každý skalár* $a \in \mathbb{R}$ *platí*

$$\widehat{\boldsymbol{\beta}}_n(a\boldsymbol{Y}, \mathbb{X}) = a\widehat{\boldsymbol{\beta}}_n(\boldsymbol{Y}, \mathbb{X}).$$

Definice 29. *Buď* $\mathbb{A} \in \mathbb{R}^{k\times k}$ *libovolná regulární matice. Odhad* $\widehat{\boldsymbol{\beta}}_n(\boldsymbol{Y}, \mathbb{X})$ *nazveme* **ekvivariantní vzhledem k afinní transformaci**, *pokud*

$$\widehat{\boldsymbol{\beta}}_n(\boldsymbol{Y}, \mathbb{X}\mathbb{A}) = \mathbb{A}^{-1}\widehat{\boldsymbol{\beta}}_n(\boldsymbol{Y}, \mathbb{X}).$$

A.2 Výsledky z matematické analýzy

Definice 30. *Nechť* $f : \mathbb{R}^n \to \mathbb{R}$ *je diferencovatelná funkce. Potom definujeme její* **gradient** *v bodě* $\boldsymbol{w} \in \mathbb{R}^n$ *jako*

$$\nabla f(\boldsymbol{w}) = \left[\frac{\partial f(\boldsymbol{w})}{\partial w_1}, \ldots, \frac{\partial f(\boldsymbol{w})}{\partial w_n}\right]^\top.$$

Definice 31. *Buď* $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ *diferencovatelná funkce. Potom definujeme* **Jakobiho matici** $\boldsymbol{f}$ *v bodě* $\boldsymbol{w} \in \mathbb{R}^n$ *jako* $m \times n$ *matici*

$$\mathbb{J}_{\boldsymbol{f}}(\boldsymbol{w}) = \begin{bmatrix} \frac{\partial f_1(\boldsymbol{w})}{w_1} & \cdots & \frac{\partial f_1(\boldsymbol{w})}{w_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m(\boldsymbol{w})}{w_1} & \cdots & \frac{\partial f_m(\boldsymbol{w})}{w_n} \end{bmatrix}.$$

Příklad. Uvažujme funkci $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ definovanou jako $\boldsymbol{f}(\boldsymbol{w}) = \mathbb{A}\boldsymbol{w}$ pro $\mathbb{A} \in \mathbb{R}^{m\times n}$. Potom $\mathbb{J}_{\boldsymbol{f}}(\boldsymbol{w}) = \mathbb{A}$.

Příklad. Mějme funkci $\boldsymbol{\sigma} : \mathbb{R}^n \to \mathbb{R}^n$, kterou dostaneme aplikací skalární funkce σ po složkách. Potom $\mathbb{J}_{\boldsymbol{\sigma}}(\boldsymbol{w})$ je diagonální matice s prvky $\sigma'(w_i)$ na diagonále. Budeme používat značení $\mathbb{J}_{\boldsymbol{\sigma}}(\boldsymbol{w}) = \mathsf{diag}(\boldsymbol{\sigma}'(\boldsymbol{w}))$.

Věta 23 (Řetízkové pravidlo)**.** *Nechť* $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ *je diferencovatelná v bodě* $\boldsymbol{w} \in \mathbb{R}^n$ *a* $\boldsymbol{g} : \mathbb{R}^m \to \mathbb{R}^k$ *je diferencovatelná v bodě* $\boldsymbol{f}(\boldsymbol{w}) \in \mathbb{R}^m$. *Potom* $\boldsymbol{g} \circ \boldsymbol{f}$ *je diferencovatelná v bodě* $\boldsymbol{w}$ *a platí*

$$\mathbb{J}_{\boldsymbol{g}\circ\boldsymbol{f}}(\boldsymbol{w}) = \mathbb{J}_{\boldsymbol{g}}(\boldsymbol{f}(\boldsymbol{w}))\mathbb{J}_{\boldsymbol{f}}(\boldsymbol{w}).$$

A.3 Závislá pozorování

Mějme posloupnost náhodných vektorů $\{\boldsymbol{Z}_i : i \in \mathbb{N}\}$ definovaných na pravděpodobnostním prostoru $(\Omega, \mathcal{F}, \mathsf{P})$ s hodnotami v měřitelném prostoru $(\mathbb{R}^d, \mathcal{B}_0^d)$.

Definice 32. *Posloupnost náhodných vektorů* $\{\boldsymbol{Z}_i : i \in \mathbb{N}\}$ *nazveme* **slabě stacionární**, *pokud*

- $\mathsf{E}\,\boldsymbol{Z}_i = \boldsymbol{\mu}$ *pro všechna* $i \in \mathbb{N}$,
- $\mathsf{cov}(\boldsymbol{Z}_i, \boldsymbol{Z}_k) = \mathsf{E}(\boldsymbol{Z}_i - \boldsymbol{\mu})(\boldsymbol{Z}_k - \boldsymbol{\mu})^\top = \mathsf{cov}(\boldsymbol{Z}_{i+h}, \boldsymbol{Z}_{k+h})$ *pro všechna* $i, k \in \mathbb{N}$ *a libovolné* h *takové, že* $i + h, k + h \in \mathbb{N}$.

Poznámka. Slabě stacionární posloupnost má tedy konstantní střední hodnotu a kovarianční matice vektorů $\boldsymbol{Z}_i$ a $\boldsymbol{Z}_k$ závisí pouze na rozdílu $i - k$.

Terminologie. V této práci budeme pod pojmem stacionarita vždy rozumět slabou stacionaritu.

Uvažujme dvě σ-algebry $\mathcal{A}, \mathcal{B} \subseteq \mathcal{F}$. Na součinové σ-algebře $\mathcal{A} \otimes \mathcal{B}$ definujme následující míry charakterizující závislost náhodných jevů ze σ-algebry $\mathcal{A}$ na náhodných jevech ze σ-algebry $\mathcal{B}$:

- $\alpha(\mathcal{A}, \mathcal{B}) := \sup_{A \in \mathcal{A}, B \in \mathcal{B}} |\,\mathsf{P}(A \cap B) - \mathsf{P}(A)\,\mathsf{P}(B)|$,
- $\beta(\mathcal{A}, \mathcal{B}) := \mathsf{E} \sup_{B \in \mathcal{B}} |\,\mathsf{P}(B|\mathcal{A}) - \mathsf{P}(B)|$.

Pozorování 24. *$\mathcal{A}$ a $\mathcal{B}$ jsou nezávislé právě tehdy, když* $\alpha(\mathcal{A}, \mathcal{B}) = 0$.[1]

Pro $1 \le a \le b \le \infty$ nechť $\mathcal{F}_a^b$ značí σ-algebru událostí generovanou náhodnými vektory $\boldsymbol{Z}_a, \ldots, \boldsymbol{Z}_b$.

Definice 33. *Pro* $n \in \mathbb{N}$ *definujme následující koeficienty závislosti*

- $\alpha(n) := \sup_{j \in \mathbb{N}} \alpha(\mathcal{F}_1^j, \mathcal{F}_{j+n}^\infty)$,
- $\beta(n) := \sup_{j \in \mathbb{N}} \beta(\mathcal{F}_1^j, \mathcal{F}_{j+n}^\infty)$.

Potom posloupnost $\{\boldsymbol{Z}_i : i \in \mathbb{N}\}$ *nazveme*

- **α-mixing** *(strongly mixing), pokud* $\alpha(n) \to 0$ *pro* $n \to \infty$,
- **β-mixing** *(absolutně regulární), pokud* $\beta(n) \to 0$ *pro* $n \to \infty$.

Poznámka. Absolutní regularita je jednou z nejslabších mixing podmínek, avšak silnější než α-mixing, neboť se dá ukázat, že platí $2\alpha(\mathcal{A}, \mathcal{B}) \le \beta(\mathcal{A}, \mathcal{B})$.

[1] Podobně pro koeficient β.

Jelikož mixing není tolik vlastností posloupnosti $\{\boldsymbol{Z}_i\}$ jako spíše posloupnosti σ-algeber generovaných $\{\boldsymbol{Z}_i\}$, platí i pro měřitelné transformace $\{\boldsymbol{Z}_i\}$, jak říká následující tvrzení.

Tvrzení 25. *Nechť $\boldsymbol{V}_i := \boldsymbol{f}(\boldsymbol{Z}_i)$ pro měřitelnou $\boldsymbol{f}$. Pokud $\{\boldsymbol{Z}_i : i \in \mathbb{N}\}$ je α-mixing, potom $\{\boldsymbol{V}_i : i \in \mathbb{N}\}$ je také α-mixing.*

Důkaz. Ukážeme podobně jako Davidson (1994) ve větě 14.1. Označme

$$\mathcal{G}_1^j := \sigma\{\boldsymbol{V}_1, \dots, \boldsymbol{V}_j\} \text{ a podobně } \mathcal{G}_{j+n}^{\infty} := \sigma\{\boldsymbol{V}_{j+n}, \boldsymbol{V}_{j+n+1}, \dots\},$$

Nyní, náhodný vektor $\boldsymbol{V}_i$ je měřitelný na každé σ-algebře, na které je vektor $\boldsymbol{Z}_i$ měřitelný. Odtud dostáváme, že $\mathcal{G}_1^j \subseteq \mathcal{F}_1^j$ a $\mathcal{G}_{j+n}^{\infty} \subseteq \mathcal{F}_{j+n}^{\infty}$. Položme

$$\alpha_{\boldsymbol{V}}(n) := \sup_{j \in \mathbb{N}} \alpha(\mathcal{G}_1^j, \mathcal{G}_{j+n}^{\infty}),$$

potom $\alpha_{\boldsymbol{V}}(n) \leq \alpha_{\boldsymbol{Z}}(n)$, jelikož $\alpha(\mathcal{A}', \mathcal{B}') \leq \alpha(\mathcal{A}, \mathcal{B})$ pro $\mathcal{A}' \subseteq \mathcal{A}$ a $\mathcal{B}' \subseteq \mathcal{B}$. Odtud již plyne tvrzení.

□

Příklad. Striktně stacionární Markovův řetězec s nejvýše spočetnou množinou stavů je absolutně regulární právě tehdy, když je nerozložitelný a neperiodický.

A.4 Zákon velkých čísel pro $\mathcal{L}^1$-mixingaly

Mějme posloupnost náhodných veličin $\{Z_i : i \in \mathbb{N}\}$ definovaných na pravděpodobnostním prostoru $(\Omega, \mathcal{F}, \mathsf{P})$ a neklesající posloupnost σ-algeber $\{\mathcal{F}_i : i \in \mathbb{Z}\}$, $\mathcal{F}_i \subseteq \mathcal{F}$. Často volíme $\mathcal{F}_i = \sigma\{Z_1, \dots, Z_i\}$ pro $i \geq 1$ a $\mathcal{F}_i = \{\emptyset, \Omega\}$ pro $i \leq 0$.

Definice 34. *Posloupnost $\{Z_i, \mathcal{F}_i\}$ nazveme $\mathcal{L}^1$-mixingalem, pokud existují nezáporné posloupnosti konstant $\{\lambda_i : i \in \mathbb{N}\}$ a $\{\psi_m : m \in \mathbb{N}_0\}$ takových, že $\psi_m \to 0$ pro $m \to \infty$ a pro všechna $i \in \mathbb{N}$ a $m \in \mathbb{N}_0$ platí*

- $\| \mathsf{E}[Z_i | \mathcal{F}_{i-m}] \|_1 \leq \lambda_i \psi_m$,
- $\| Z_i - \mathsf{E}[Z_i | \mathcal{F}_{i+m}] \|_1 \leq \lambda_i \psi_{m+1}$.

Poznámka. Druhá podmínka z předchozí definice většinou platí triviálně, neboť jsou-li veličiny Z_i $\mathcal{F}_i$-měřitelné, potom $\mathsf{E}[Z_i | \mathcal{F}_{i+m}] = Z_i$ skoro jistě.

Poznámka. $\mathcal{L}^1$-mixingaly jsou centrované, tedy pro obecnou posloupnost náhodných veličin $\{Z_i : i \in \mathbb{N}\}$ musíme uvažovat $\{Z_i - \mathsf{E}\, Z_i : i \in \mathbb{N}\}$.

Tvrzení 26. *Buď $\{Z_i, \mathcal{F}_i\}$ stejnoměrně integrovatelný $\mathcal{L}^1$-mixingal. Nechť navíc $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n \lambda_i < \infty$, nebo $\{\lambda_i\} = \{\|Z_i\|_1\}$. Potom $\frac{1}{n} \sum_{i=1}^n Z_i \xrightarrow{\mathsf{P}} 0$ pro $n \to \infty$.*

Důkaz. Viz Andrews (1988), věta 1.

□

Poznámka. Varianta definice 34, resp. tvrzení 26 pro trojúhelníkové schéma náhodných veličin se formuluje obdobně, viz Andrews (1988), definice 2, resp. tvrzení 2.

Příklad. Mějme $\{Z_i : i \in \mathbb{N}\}$ posloupnost centrovaných α-mixing náhodných veličin, které jsou $\mathcal{L}^p$ omezené pro nějaké $p > 1$. Položme $\mathcal{F}_i = \sigma\{Z_1, \dots, Z_i\}$ pro $i \in \mathbb{N}$ a $\mathcal{F}_i = \{\emptyset, \Omega\}$ pro $i \leq 0$.

$\{Z_i\}$ jsou stejnoměrně integrovatelné, neboť se jedná o posloupnost $\mathcal{L}^p$ omezených náhodných veličin, $p > 1$. Nyní ve větě 14.2 (Davidson, 1994) uvažujme $p = 1$, $r = p$ a připomeňme, že se jedná o centrované veličiny. Dostáváme

$$\| \mathsf{E}[Z_{i+n}|\mathcal{F}_i]\|_1 \leq 6\alpha(n)^{1-1/p}\|Z_{i+n}\|_p.$$

Tedy $\{Z_i, \mathcal{F}_i\}$ je stejnoměrně integrovatelný $\mathcal{L}^1$-mixingal s $\{\lambda_i\} = \{\|Z_i\|_p\}$ splňující tvrzení 26.

Zákon velkých čísel pro $\mathcal{L}^1$-mixingaly tedy můžeme použít i v případě absolutně regulární posloupnosti, neboť β-mixing implikuje α-mixing.

Příklad. Také posloupnost martingalových diferencí $\{Z_i, \mathcal{F}_i\}$ je $\mathcal{L}^1$-mixingal s $\psi_m = 0$ pro $m \geq 1$ a $\lambda_i = \|Z_i\|_1$, pokud položíme $\mathcal{F}_i = \{\emptyset, \Omega\}$ pro $i \leq 0$ a $\mathcal{F}_i = \sigma\{Z_1, \dots, Z_i\}$ pro $i \in \mathbb{N}$. Tudíž, jsou-li $\{Z_i\}$ stejnoměrně integrovatelné, platí tvrzení 26.

A.5 Klasifikační metriky

Uvažujme úlohu binární klasifikace, kdy chceme rozpoznat kontaminovaná pozorování od těch běžných.

Jako **skutečně pozitivní** (TP) budeme nazývat počet správně rozpoznaných odlehlých pozorování a jako **falešně pozitivní** (FP) počet odlehlých pozorování, která jsme nesprávně označili za běžná. Podobně, **skutečně negativní** (TN) značí počet správně rozpoznaných běžných pozorování a **falešně negativní** (FN) počet běžných pozorování označených za odlehlá.

Definice 35. *Označme*

$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}, \quad \mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \quad \mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}.$$

Potom ACC *nazýváme* **přesnost**, PPV **preciznost** *a* TPR **paměť**.

Přesnost je vhodná metrika, pokud jsou všechny případy stejně důležité. V našem případě, kdy je počet běžných a odlehlých pozorování nevyvážený je vhodnější metrikou F_1-skóre.

Definice 36. F_1-skóre *definujeme jako harmonický průměr preciznosti a paměti,*

$$\mathrm{F}_1 = \frac{2}{\mathrm{PPV}^{-1}+\mathrm{TPR}^{-1}} = \frac{2\cdot\mathrm{PPV}\cdot\mathrm{TPR}}{\mathrm{PPV}+\mathrm{TPR}}.$$