Draft poznámok k predmetu

Moderná Aplikovaná regresia 2

Tento dokument je predbežný a nedokončený, nešírte ho prosím. Budem ho priebežne upravovať. Jeho ambíciou je podať stručný prehľad a okrem toho doplniť a dovysvetliť niektoré časti v [Far14], sám o sebe je preto len ťažko čitateľný. Ďakujem Michaele Mihokovej za mnoho pripomienok.

Lukáš Lafférs*

4. mája 2021

1 Binomické dáta

Majme výstup, ktorý je typu 0 alebo 1, teda napríklad akási udalosť nastala alebo nenastala. V dátach pozorujeme počet "úspechov"(udalosť nastala) a počet "neúspechov"(udalosť nenastala), pre rôzne situácie, napríklad počet 50-ročných pacientov - mužov, ktorí dostali alebo nedostali chorobu. Zaujímalo by nás, ako dobre vysvetliť a predikovať pravdepodobnosť tejto udalosti.

Klasický lineárny model $Y_i = X_i^T\beta + \epsilon_i$ na toto nie je vhodný, nakoľko môže predikovať hodnoty mimo $[0,1]$ intervalu. Aj ak by sme hodnoty mimo tohoto intervalu posunuli na hranicu $[0,1]$, pre nejaké X_i (riadkový náhodný vektor) by sme predikovali pravdepodobnosť rovnú 1 alebo 0, a to nie je žiadúce.

Uvažujme, že vysvetľovaná premenná Y_i, ktorú pozorujeme, je pre fixné regresory $X_i = x_i$ dobre popísateľná binomickým rozdelením. Náhodná premenná Y_i má binomické rozdelenie, označujeme $Y_i \sim B(n_i, p_i)$ ak $P(Y_i = y_i) = \binom{n_i}{y_i} p_i^{y_i}(1-p_i)^{n_i - y_i}$. Pre takúto náhodnú premennú variancia $Var(Y_i) = n_i p_i(1-p_i)$ závisí od pravdpedpodobnosti p_i, čo je ďalší dôvod, prečo ju nemodelovať klasickým regresným modelom s normálne rozdelenými a homoskedastickými chybami. Takéto rozdelenie modeluje počet úspešných pokusov, ak máme n_i **nezávislých** pokusov, každý s **fixnou** pravdepodobnosťou úspechu p_i. "Úspech" v našom prípade znamená, že akási udalosť nastala. Môže ísť o skrachovanie firmy, výbuch elektrárne či víťazstvo vo voľbách.

Nakoľko potrebujeme, aby náš model predikoval hodnoty v $[0,1]$ intervale, zaujíma nás vhodná transformácia $g(p)$ (ktorú budeme nazývať **link** a $\eta_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_q x_{iq}$ budeme nazývať **lineárny prediktor**), taká aby jej definičný obor bola reálna os a obor hodnôt bol interval $[0,1]$. Príklady takejto transformácie sú

- $g(p) = \text{logit}(p) \equiv \log\left(\frac{p}{1-p}\right)$,
- $g(p) = \text{probit}(p) = \Phi^{-1}(p)$, kde $\Phi(u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right) dx$.

Takéto modely patria do skupiny **zovšeobecnených lineárnych modelov** a budeme sa im venovať podrobne neskôr.

Vezmime si teda model

$$g(p_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_q x_{iq}, \tag{1.1}$$

kde predpokladáme, že $Y_i \sim B(n_i, p_i)$. V rovnici (1.1) predpokladáme, že neznámy parameter p_i je deterministickou funkciou regresorov $(x_{i1}, ..., x_{iq})$. Teda prepojíme **náhodný komponent** Y_i a **systematický komponent** η_i pomocou **linkovej funkcie**.

*E-mail: lukas.laffers@umb.sk

Ak sú všetky n_i dostatočne veľké, vďaka Centrálnej limitnej vete vieme aproximovať $B(n_i, p_i)$ normálnym rozdelením $N(n_i p_i, n_i p_i(1-p_i))$, a preto aj ϵ_i bude dobre aproximované normálnym rozdelením.

V prípade, že sa na Y_i nemôžeme pozerať ako na výsledok nezávislých udalostí s fixnou pravdepodobnosťou, nie je tento model rozumný. Linearita prediktora v parametre β je menej obmedzujúca, než by sa na prvý pohľad zdalo, nakoľko sa tým dajú modelovať aj nelineárne vplyvy prediktorov, aj kategorické premenné (cez dummy premenné).

Voľba linkovej funkcie pri binomických dátach nie je až tak dôležitá, pokiaľ nemodelujeme príliš malé alebo príliš veľké pravdepodobnosti, tam totiž môžu byť rozdiely výrazné. Niekedy však povaha problému pomôže s výberom modelu. Uvažujme nasledovný prípad: nech hmyz umrie pokiaľ je dávka insekticídu x vyššia ako jeho tolerancia T_i. Pravdepodobnosti úhynu hmyzu i je $p = P(Y_i = 1|X_{i1} = x_{i1}) = P(T \leq x_{i1})$. Ak aproximujeme distribúciu T normálnym rozdelením s parametrami (μ, σ^2), potom

$$\Phi^{-1}(p) = (x_{i1} - \mu)/\sigma = \beta_0 + \beta_1 x_{i1},$$

teda poskytuje nám to zdôvodnenie výberu probitu. Ak by bolo rozdelenie tolerancie logistické (výraz pre hustotu logistickej distribúcie vynechávame), zdôvodnili by sme takto logit ako linkovú funkciu.

Ak v rámci nejakého modelu vieme data generujúci proces úplne popísať konečnorozmerným parametrom, v našom prípade je to parameter β, vieme odhadnúť tento parameter metódou **maximálnej vierohodnosti** (maximum likelihood). Táto metóda je optimálna v zmysle asymptoticky (teda limitne, pre veľkú dátovú vzorku) najmenšej variancie (odhad je efektívny), odhady MLE sú konzistentné a asymptoticky normálne (viacej v časti A.4). V prípade lineárneho modelu s predpokladom normálnych chýb sme mali analytický tvar pre odhad MLE ($\hat{\beta}_{MLE} = (X^T X)^{-1} X^T y$, kde X je matica $n \times p$, kde p je počet parametrov v modeli). Teraz tomu tak nie je a odhad musíme získať numericky. Log-likelihood pre náš model vyzerá nasledovne:

$$\begin{aligned} l(\beta) &= \log \prod_{i=1}^{n} P(Y_i = y_i) = \sum_{i=1}^{n} \log \left(\binom{n_i}{y_i} p_i^{y_i} (1-p_i)^{n_i - y_i} \right) \\ &= \sum_{i=1}^{n} \left[\log \binom{n_i}{y_i} + y_i \log p_i + (n_i - y_i) \log(1-p_i) \right] \\ &= \sum_{i=1}^{n} \left[\log \binom{n_i}{y_i} + y_i \log \frac{p_i}{1-p_i} + n_i \log(1-p_i) \right] \\ &= \sum_{i=1}^{n} \left[\log \binom{n_i}{y_i} + y_i \eta_i - n_i \log(1 + \exp(\eta_i)) \right], \end{aligned} \tag{1.2}$$

kde η_i je funkciou β. Odhad MLE, teda $\hat{\beta}_{MLE} = \arg\max L(\beta)$, odpovedá na otázku: ktorý model (teda ktorý konkrétny vektor parametrov β) mal najväčšiu šancu vygenerovať naše dáta?

1.1 Štatistická inferencia

Majme 2 modely, jeden veľký model s l parametrami a ďalší malý model s s parametrami ($s < l$). Nech menší model je špeciálny prípad väčšieho modelu pri $l - s$ lineárnych reštrikciách (vtedy hovoríme, že malý model je **vnorený** (nested) vo veľkom modeli). Nasledujúca štatistika sa nazýva **pomer vierohodností** (likelihood ratio)

$$LR = 2 \log \frac{L_L}{L_S},$$

kde L_S (L_L) označuje vierohodnosť malého (veľkého) modelu. Za predpokladu správnosti (korektnej špecifikácie) malého modelu, má LR asymptoticky rozdelenie χ^2_{l-s}, teda chí-kvadrát s $l - s$ stupňami voľnosti.

Uvažujme teraz **nasýtený model**, to je taký, že má toľko parametrov ako pozorovaní a vie predikovať vysvetľované premenné v našej dátovej vzorke úplne presne. Pre tento model platí $p_i = y_i/n_i$. Pre zjednodušenie si môžeme tento model predstaviť tak, že regresory pozostávajú z n rôznych faktorov,

kde n je počet pozorovaní. Teraz zavedieme mieru **vhodnosti fitu** (goodness-of-fit) nasledovným spôsobom. Pre model s s parametrami uvažujme test pomocou pomeru vierohodností, kde $H_0 : L_L = L_S$, oproti alternatíve $H_0 : L_L > L_S$, kde veľký model je nasýtený model. Testovaciu štatistiku takého testu nazveme **deviancia** (deviance) a má nasledovnú formu.

$$\begin{aligned} D \equiv 2\log\frac{L_L}{L_S} &= 2\sum_{i=1}^{n}\left[\log\binom{n_i}{y_i} + y_i\log p_i + (n_i - y_i)\log(1-p_i)\right] \\ &- 2\sum_{i=1}^{n}\left[\log\binom{n_i}{\hat{y}_i} + y_i\log\hat{p}_i + (n_i - y_i)\log(1-\hat{p}_i)\right] \\ &= 2\sum_{i=1}^{n}\left[y_i\log\frac{y_i}{\hat{y}_i} + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right)\right], \end{aligned} \tag{1.3}$$

kde $\hat{y}_i$ sú hodnoty predikované menším modelom (poznamenávame, že saturovaný model predikuje y_i pre $\forall i$). V prípade, že je deviance nie príliš veľká, malý model má dosť dobrý fit. Za predpokladu správnosti malého modelu, má totiž deviance rozdelenie χ^2_{n-s}, teda rozdelenie, ktoré má strednú hodnotu $n - s$ a smerodajnú odchýlku $\sqrt{2(n-s)}$.

Všetky tieto testy sú založené na approximácii binomického rozdelenia normálnym rozdelením. Binomicky rozdelená náhodná premenná $Y \sim B(n,p)$ má $E(Y) = np$ a $Var(Y) = np(1-p)$. Keďže $Y = X_1 + \cdots + X_n$ je súčtom rovnako rozdelených náhodných premenných, podľa Centrálnej limitnej vety vieme, že pre veľké n, bude Y rozdelené približne ako $N(np, np(1-p))$. Pre malú dátovú vzorku n máme problém, lebo normálna aproximácia nebude fungovať, preto to, čo predtým bolo rozdelené ako χ^2_{l-s} (súčet nezávislých $N(0,1)$ premenných na druhú), už teraz nebude. Potom na štatistickú inferenciu môžeme použiť *bootstrap*, o ktorom budeme hovoriť neskôr.

Špeciálnym prípadom je, keď sú $n_i = 1$ pre všetky i. Vtedy deviance vôbec nezáleží od y_i a preto logicky nemôže hovoriť nič o vhodnosti fitu. Je to preto, lebo likelihood saturovaného modelu v tomto prípade je vždy 1 a s takýmto dobrým fitom sa ťažko súperí. V tomto modeli je likelihood

$$\prod_{i=1}^{n}\hat{p}_i^{y_i}(1-\hat{p}_i)^{1-y_i} = \prod_{i=1}^{n}1 = 1,$$

pretože platí $y_i = \hat{p}_i$. V takomto prípade potrebujeme čosi úplne iné (napr. Hosmer-Lemeshow test). Ale čo je to "dosť veľké" n_i? Niektorí tvrdia, že by malo byť $n_i \geq 5$ alebo pomer celkového počtu pozorovaní a parametrov v modeli musí byť aspoň 15.

Porovnať dva vnorené modely sa dá tiež porovnaním deviancií, je to v podstate to isté ako likelihood ratio test.

$$2\log\frac{L_{M_1}}{L_{M_2}} = 2\log\frac{L_{saturated}}{L_{M_2}} - 2\log\frac{L_{saturated}}{L_{M_1}} = D_{M_2} - D_{M_1},$$

kde model M_2 je špeciálnym prípadom modelu M_1 (model M_2 je vnorený do modelu M_1, teda nutne dáva väčšiu devianciu).

Waldove Konfidenčné intervaly pre jednotlivé β_j sú $[\hat{\beta}_j - z^{\alpha/2}se(\hat{\beta}_j), \hat{\beta}_j + z^{\alpha/2}se(\hat{\beta}_j)]$, kde $z^{\alpha/2}$ je $100(\alpha/2)$-%ný kvantil $N(0,1)$ rozdelenia. Tieto konfidenčné intervaly sú tak dobré, ako je normálna aproximácia $\hat{\beta}$ normálnym rozdelením.

Alternatívou je použiť **profile-likelihood**, ktorý funguje lepšie pre malé dátové vzorky ako spomínaný Waldov konfidenčný interval. Predstavme si, že chceme konfidenčný interval pre β_1 v logite s jedným regresorom a konštantou

$$L_1(\beta_1) = \max_{\beta_0} L(\beta_0, \beta_1).$$

Nulovú hypotézu $H_0 : \beta_1 = \theta$ nezamietame práve vtedy a len vtedy, ak

$$2[\log L(\hat{\beta}_0, \hat{\beta}_1) - \log L_1(\theta)] < \chi^2_1(1-\alpha),$$

preto množina všetkých θ, ktoré spĺňajú túto nerovnosť, je konfidenčný interval založený na profilovom likelihood-e.

Profile likelihood funguje lepšie, teda dáva užšie konfidenčné intervaly kvôli tomu, že používa viacej informácie, používa celý likelihood, kým Waldov konfidenčný interval neberie do úvahy neistotu v iných regresoroch ako v samotnom j-tom regresore. Teda rozdiel medzi týmito dvomi konfidenčnými intervalmi bude výraznejší, ak sú odhady parametrov veľmi závislé.

1.2 Interpretácia

V obyčajnom lineárnom regresnom modeli mali parametre β jednoduchú interpretáciu, to bola aj jedna z výhod tohoto modelu. Ako je to v tomto prípade?

Model s logit linkovou funkciou vyzerá nasledovne

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_{i1} + ... + \beta_q x_{iq}$$

$\text{logit}(p) = \log \frac{p}{1-p}$ a ide teda o logaritmus **pomeru šancí** (odds ratio). Pomer šancí hovorí, koľkokrát je pravdepodobnosť nastatia udalosti väčšia ako pravdepodobnosť nenastatia udalosti. Ak teda zvýšime x_j o jednu jednotku *a zároveň sa všetky ostatné premenné nezmenia*, pomer šancí sa ponásobí faktorom $\exp(\beta_j)$. Toto je výhoda oproti probitu, pretože preň neexistuje žiadna podobná interpretácia. Táto interpretácia je možno pre nás menej prirodzená, lebo nie sme zvyknutý premýšľať v *pomeroch šancí.* V prípade, že je β_j malá (napr. $-0.25 \leq \beta_j \leq 0.25$), potom môžeme použiť nasledovnú aproximáciu:

$$\log\left(\frac{p}{1-p}(1+\beta_j)\right) = \log\frac{p}{1-p} + \log(1+\beta_j) \approx \log\frac{p}{1-p} + \beta_j.$$

Ak teda $\beta_j = 0.05$, potom predikujeme, že pomer pravdepodobností, či udalosť nastala voči tomu, že nenastala, vzrastie o 5%, ak x_j stúpne o jednotku a zároveň sa ostatné premenné nezmenia. Poznamenávame, že na to, aby sme mohli koeficienty interpretovať separátne, potrebujeme aby v modeli neboli interakčné členy. **Dôrazne sa stránime kauzálnej interpretácie**, pokiaľ nemáme zaručené, že zber dát nesledoval experimentálny protokol.

1.3 Výber dátovej vzorky

Ako boli zozbierané naše dáta? Ak sme si fixovali vzorku a potom sme sledovali výstupy, tak sa tomu hovorí prospektívny výber vzorky (*cohort study*). Alebo výstup bol fixovaný a my sme pozbierali dáta, napríklad sme sa osobitne pozreli na chorých a potom na zdravých pacientov, toto je retrospektívny výber dát (*case-control study*).

Predstavme si, že chceme vedieť, ako nám vedomosť o spôsobe príjmu potravy novorodencov-chlapcov pomôže predikovať respiračné choroby. V skupine, kde sú dojčené deti, malo respiračné problémy 47 chlapcov oproti 447, ktorí problémy nemali. V skupine detí kŕmených z fľašky je to 77 oproti 381.

Ak máme prospektívnu štúdiu, tak máme vzorku chlapcov, o ktorých vieme, ako boli krmení, a potom sa dozvieme, či mali alebo nemali respiračné problémy. Rozdiel v log-odds pre respiračnú chorobu (logaritmus podielu pravdepodobností, že dieťa bude mať problémy voči tomu, že nebude mať problémy) pre dojčené a fľaškou kŕmené deti je

$$\log\frac{77}{381} - \log\frac{47}{447} = 0.65.$$

Ak by sme tie isté dáta zozbierali retrospektívnou štúdiou, ktorá je lacnejšia a jednoduchšia, pozreli by sme sa na chlapcov, ktorí mali respiračné problémy a ktorí nemali respiračné problémy separátne. Rozdiel v log-odds pre tieto skupiny je

$$\log\frac{77}{47} - \log\frac{381}{447} = 0.65,$$

teda prišli sme k tomu istému výsledku, oveľa lacnejšou cestou. Toto neplatí pre probit alebo pre iné linkové funkcie.

V praxi však musíme zobrať do úvahy aj efekt iných premenných a máme problém, pretože pravdepodobnosť toho, či sa jednotlivé deti ocitnú v našej vzorke, závisí od toho, či majú alebo nemajú respiračné problémy. Typicky by sme čakali, že pravdpodobnosť, že dieťa je v našej vzorke ak má respiračné problémy (označíme ako π_1), je oveľa väčšia ako pravdepodobnosť, že dieťa je v našej vzorke, ak nemá respiračné problémy (označíme ako π_0). Pri prospektívnej štúdii platí približne $\pi_0 = \pi_1$. Podmienená pravdepodobnosť, že pre dané $X = x$ bude mať dieťa respiračnú chorobu, ak bolo v našej retrospektívnej vzorke, je daná vzťahom

$$p^*(x) = \frac{\pi_1 p(x)}{\pi_1 p(x) + \pi_0(1 - p(x))},$$

podľa Bayesovej vety, kde $p(x)$ je nepodmienená pravdepodobnosť choroby. Toto však nevadí, lebo logity sa budú líšiť len o konštantu, a tá nás aj tak väčšinou nezaujíma:

$$\text{logit}(p^*(x)) = \log \frac{\pi_1}{\pi_0} + \text{logit}(p(x)),$$

teda relatívne efekty premenných odhadneme správne aj napriek tomu, že π_0 a π_1 nepoznáme(!) Toto znova nefunguje pre iné linkové funkcie ako logit.

Takže dokopy máme už 3 výhody logitu oproti probitu: (1) má matematicky jednoduchšiu formu (mnoho vecí vieme analyticky vyjadriť), (2) koeficienty vieme interpretovať a (3) namiesto prospektívnych štúdií nám stačia retrospektívne.

1.4 Problémy s odhadovaním

Môže sa nám stať, že program pri maximalizovaní likelihood funkcie nenájde optimum. Ak sme v situácii, že sa snažíme vysvetľovať binárne dáta ($n_i = 1$), môže to byť kvôli tomu, že dáta sú lineárne separovateľné a model poskytuje perfektný fit. Vtedy je likelihood funkcia plochá a solver si nevie na tejto plochej časti vybrať optimum. Zbadáme to tak, že nás program upozorní na zlyhanie konvergencie a zároveň zvyknú byť odhady koeficientov veľmi nepresné, s obrovskými štandardnými chybami. Vtedy môžeme použiť `brglm` funkciu z rovnomennej knižnice (Bias Reduction in Binomial-Response Generalized Linear Models).

1.5 Vhodnosti fitu

Okrem deviancie môžeme použiť na overenie vhodnosti fitu aj pomer vysvetlenej deviancie.

$$R^2 = \frac{1 - (\hat{L}_0/\hat{L})^{2/n}}{1 - \hat{L}_0^{2/n}} = \frac{1 - \exp((D - D_0)/n)}{1 - \exp(-D_0/n)},$$

kde $\hat{L}_0$ je odhad maximum likelihood (null) modelu len s interceptom a D_0 je deviancia tohoto modelu.

1.6 Efektívna dávka

Môžeme sa opýtať otázku: aká má byť dávka insekticídu tak, aby bola pravdepodobnosť vyhubenia hmyzu (napr.) 50%? Nastavíme $p = 0.5$ a zistíme, že je to $-\beta_0/\beta_1$. Toto je pre nás neznáma kvantita, ktorú však vieme odhadnúť ako $-\hat{\beta}_0/\hat{\beta}_1$. Ak chceme vyjadriť neistotu v tomto odhade pomocou konfidenčného regiónu, môžeme použiť delta-metódu. Pre $g(.)$ transformáciu odhadu viacrozmerného vektora parametrov $\hat{\theta}$ vieme, že jej variancia je približne:

$$Var(g(\hat{\beta})) \approx g'(\hat{\beta})^T Var(\hat{\beta}) g'(\hat{\beta}).$$

V našom prípade $g(\hat{\beta}) = -\hat{\beta}_0/\hat{\beta}_1$ a $g'(\hat{\beta}) = \left(-\frac{1}{\hat{\beta}_1}, \frac{\hat{\beta}_0}{\hat{\beta}_1^2}\right)^T$. Približný 95%ný konfidenčný interval pre neznámu efektívnu dávku je $\left(g(\hat{\beta}) - 1.96\sqrt{Var(g(\hat{\beta}))}, g(\hat{\beta}) + 1.96\sqrt{Var(g(\hat{\beta}))}\right)$. Táto aproximácia je taká dobrá, ako veľmi je funkcia g aproximovateľná lineárnou funkciou a aká veľká je dátová vzorka.

1.7 Overdispersion

Ak sme v situácii, že deviancia je veľmi veľká rozmýšlame, čo v našom modeli je nesprávne. Môže to byť tým, že sme do regresie nedali správne prediktory alebo sme zvolili nevhodnú transformáciu regresorov. Podobne to môže byť kvôli outlierom, v prípade, že ich je veľa, tak niečo nie je v poriadku s distribúciou chýb. Podobne to môže byť pre veľmi malé n_i, lebo vtedy normálna aproximácia nie je vhodná.

V prípade, že sme vylúčili všetky tieto možnosti, mohlo nastať, že v rámci skupiny veľkosti m, výsledky jednotlivých experimentov nie sú nezávislé alebo nie sú rovnako rozdelené, teda binomický model nie je rozumnou aproximáciou. V prípade skúmania okrúhlych tesnení na raketopláne Challenger mohli byť udalosti zlyhania tesnenia pozitívne korelované, vtedy je variancia celkového počtu chýb väčšia ako $mp(1-p)$ a tento problém nazývame **overdispersion**. Alebo mohla byť pravdepodobnosť pre tesnenia v rôznych častiach raketoplánu rôzna. Vo vzácnych prípadoch dochádza aj k opačnému problému - underdispersion. Ako príklad môžeme uvažovať situáciu, keď úmrtie jedného zvieraťa má pozitívny efekt na prežitie ostatných, lebo potravy bude dostatok.

Nech je k veľkosť klastra (skupiny), l je počet klastrov a $m = kl$ je počet pozorovaní. Nech je počet úspechov v klastri i rozdelený ako $Z_i \sim B(k, p_i)$ a nech p_i je náhodná premenná s vlastnosťami $E(p_i) = p$ a $Var(p_i) = \tau^2 p(1-p)$. Potom pre $Y = Z_1 + \cdots + Z_l$ platí

$$E(Y) = mp, \qquad Var(Y) = [1 + (k-1)\tau^2]mp(1-p).$$

Ako modelovať overdispersion? Disperzný parameter $\sigma^2 = 1$ v prípade, že predpoklady binomického modelu platia. Inak ho môžeme odhadnúť nasledovne

$$\hat{\sigma}^2 = \frac{X^2}{n-p} = \frac{1}{n-p}\sum_{i=1}^{n}\frac{(y_i - n_i\hat{p}_i)^2}{n_i\hat{p}_i(1-\hat{p}_i)},$$

kde X^2 nazývame **Pearsonove** X^2, ktoré sa dá, podobne ako deviancia, použiť ako miera vhodnosti fitu modelu. Pearsonove X^2 má tvar $X^2 = \sum_{i=1}^{n}\frac{(O_i - E_i)^2}{E_i}$, kde O_i je pozorovaný počet udalostí a E_i je priemerný počet udalostí za predpokladu modelu, ktorý testujeme.

Naše odhady β sa nemenia, pretože σ^2 nemení strednú hodnotu Y_i ale zmení sa kovariančná matica β, ktorá teraz bude mať tvar $\hat{Var}(\hat{\beta}) = \hat{\sigma}^2(X^T\hat{W}X)^{-1}$, takže štandardné chyby musíme ponásobiť $\hat{\sigma}^2$.[1] Na test podmodelu musíme narozdiel od rozdielu deviancií použiť

$$F = \frac{(D_S - D_L)/(df_S - df_L)}{\hat{\sigma}^2},$$

kde táto štatistika je len približne rozdelená ako F (S (small) označuje malý model, L (large) označuje veľký model). Toto sa však dá použiť len ak sú veľkosti skupín n_i približne podobne veľké. Ak nie sú, potrebujeme použiť čosi rozumnejšie.

1.8 Matching

Keď máme case-control štúdiu (teda retrospektívnu vzorku, nemáme experiment), častokrát chceme brať do úvahy nejaké **risk faktory**. Medzi ľudmi, ktorí podstúpili liečbu môže byť napríklad viacej starších ľudí. Môžeme risk faktory pridať do regresie, ale akú transofrmáciu máme zvoliť? Tohoto problému sa môžeme zbaviť pomocou **matchingu**. Základom matchingu je, že porovnávame podobné objekty.

Každému case-u (napr. človek, ktorý podstúpil liečbu) sa snažíme nájsť control (napr. človek, ktorý nepodstúpil liečbu) s podobnými risk faktormi. Ak máme 56-ročného hispánca, ktorý napr. dostal nejakú chorobu (case), ideálne sa snažíme nájsť tiež 56-ročného hispánca, ktorý však túto chorobu nedostal (control). Toto sa nedá vždy nájsť presne a preto musíme poľaviť z našich požiadaviek na matching, napríklad budeme uvažovať skupinu 50-59 ročných hispáncov. Výhodou je, že môžeme brať do úvahy nám neznáme faktory, ktoré napr. súvisia s geografickou polohou. Ako urobiť matching? To je náročná

[1] Matica $\hat{W}$ označuje maticu váh. K vysvetleniu súvisu zovšeobecneného lineárneho modelu a váhovanej regresie sa dostaneme v piatej kapitole.

otázka a je veľa rôznych spôsobov, ktoré sú vhodné pre rôzne situácie. Je totiž veľa rôznych spôsobov ako merať, že ktoré objekty sú podobné. Nevýhodou matchingu je, že efekt risk faktorov, na základe ktorých je urobený matching nevieme skúmať.

Častokrát máme oveľa viacej control-ov ako case-ov. Ak máme 1:M dizajn, tak na 1 case máme M control-ov. Zvyčajne $M = 5$ stačí a viacej control=ov nám spresní odhad iba minimálne. Pre pozorovania $i = 0, 1, ..., M$ v skupine j uvažujme

$$\text{logit}(p_j(x_{ij})) = \alpha_j + x_{ij}^T \beta,$$

kde α_j je efekt risk faktorov v skupine j. V rámci skupiny j je pravdepodobnosť case-u, teda toho, že $i = 0$

$$\frac{\exp x_{0j}^T \beta}{\sum_{i=0}^{M} \exp x_{ij}^T \beta},$$

teda efekt skupiny α_j nám vypadne a nebudeme ho vedieť odhadnúť. Likelihood funkcia potom vyzerá nasledovne

$$L(\beta) = \prod_{j=1}^{n} \left\{ 1 + \sum_{i=1}^{M} \exp\left[(x_{ij} - x_{0j})^T \beta \right] \right\}^{-1}.$$

Toto je rovnaký likelihood ako pri tzv. *conditional hazard* modeloch, a to je výhodné, lebo sú na to hotové knižnice (`survival`).

1.9 Literatúra

Tieto poznámky sú založené predovšetkým na [Far05], ktorá pokračuje prístupnú expozíciu lineárnych modelov v [Far14]. Na úvod si môžete prečítať jednoduché čítanie [A+07] a doplniť v oveľa podrobnejšej knihe [AK11]. [Agr15] poskytuje dobrý mix prístupnosti a detailov.

2 Dáta typu počet udalostí - Count Data

Nech naša závislá premenná má tvar počtu nastatia udalostí, teda $Y_i \in \mathbb{N}$. Obyčajný lineárny model môže predikovať záporné hodnoty. Ak sa to snažíme vyriešiť tým, že použijeme lineárny regresný model s $\log Y_i$ na ľavej strane, máme problém s $Y_i = 0$.

Uvažujeme, že Y_i má **Poissonovo** rozdelenie s parametrom μ_i, $Y_i \sim Pois(\mu_i)$, teda $P(Y_i = y_i) = \exp(-\mu_i)\frac{\mu_i^{y_i}}{y_i!}$ s vlastnosťami $E(Y_i) = Var(Y_i) = \mu_i$. Zároveň uvažujeme, že

$$\log(\mu_i) = \eta_i = x_i^T\beta.$$

Používame logaritmickú linkovú funkciu, aby sme dostali kladné hodnoty priemerného počtu udalostí μ_i.

Príklady:

- Počet rastlinných druhov na ostrove. Prediktory môžu byť plocha alebo nadmorská výška.
- Počet ľudí v rade pred Vami v obchode. Prediktory môžu byť počet produktov v zľave alebo či je vonku slnečné počasie.
- Počet ocenení študentov na strednej škole. Prediktory môžu byť typ školy alebo skóre z matematickej skúšky.

Prečo Poissonovo rozdelenie? Poissonovo rozdelenie má tú vlastnosť, že súčet nezávislých $Pois(\mu_1)$ a $Pois(\mu_2)$ rozdelených náhodných premenných nám dá $Pois(\mu_1+\mu_2)$. Toto je výhodné, lebo poissonovská aproximácia bude fungovať nezávisle od veľkosti skupín (či už sa pozeráme na individuálov, pre ktorých platí $Y_{ij} \sim Pois(\mu_i)$ pre $j = 1, 2, ..., n_i$ alebo na celú skupinu, kde $Y_i \sim Pois(n_i\mu_i)$).

Okrem toho je toto rozdelenie ideálne napríklad na modelovanie počtov nezávislých udalostí v čase, skrátka na dáta, ktoré sú realizáciou Poissonovho procesu, alebo kde je Poissonov proces rozumnou aproximáciou. Na to potrebujeme: (1) pravdepodobnosť, že nastane aspoň jedna udalosť je proporciálna času, (2) pravdepodobnosť nastatia dvoch udalostí na veľmi malom intervale je zanedbateľná, (3) počty udalostí v neprekrývajúcich sa časových intervaloch sú nezávislé. Vieme, že čas medzi dvoma udalosťami Poissonovho procesu má exponenciálne rozdelenie a to, ako jediné má vlastnosť, že si nič nepamätá (memoryless property). My môžeme modelovať, ako sa mení intenzita udalostí, tá môže záležať nielen od času ale aj od iných parametrov, presne na to slúžia regresory X_i. Týmto modelom môžeme teda popísať aj intenzitu nastávania javu za nejaký čas:

$$\log(\mu_i/t) = \eta_i = x_i^T\beta \quad \sim \quad \log(\mu_i) = \eta_i = x_i^T\beta + \log(t),$$

kde $\log t$ nazývame **offset**, je to vlastne regresor s fixným parametrom. Offsetom nemusí byť len čas ale čokoľvek, čo rastie alebo klesá proporciálne s Y_i. (Offset je výborne vysvetlený v [Cra12] od strany 518.)

Poissonovo rozdelenie môže byť užitočné ako aproximácia binomického rozdelenia pri veľkom n_i a malom p_i, ale fixnom $\mu_i = n_ip_i$. Ak pri 5000 predaných telefónov sa pokazí 0.1% z nich, potom počet reklamovaných telefónov je približne rozdelený ako $Pois(5)$. Podobne, pri skúmaní vzácneho typu rakoviny je počet postihnutých jedincov malý a Poissonov model s logaritmickou linkovou funkciou bude podobný ako binomický model s logit linkovou funkciou.

Log-Likelihood vyzerá nasledovne

$$\log L(\beta) = \sum_{i=1}^{n}(y_ix_i^T\beta - \exp(x_i^T\beta) - \log y!),$$

po zdiferencovaní podľa β dostávame $X^Ty = X^T\hat{\mu}$. Toto je rovnaké ako pri Gaussovskom modeli, kedy $\hat{\mu} = X\hat{\beta}$, no neplatí to pre hocijakú linkovú funkciu. Taká, pre ktorú to platí, sa nazýva **kanonická**.

2.1 Štatistická inferencia

Štatistická inferencia vyzerá rovnako ako pri binomických dátach, môžeme používať devianciu ako mieru vhodnosti fitu (deviancia nepovie, *prečo* model nefituje dáta dobre), pomer vierohodností (likelihood ratio) na testovanie lineárnych hypotéz vnorených modelov, používať Waldove konfidenčné intervaly alebo (lepšie je použiť) profile likelihood konfidenčné intervaly.

Deviancia a Pearsonove X^2 majú nasledovnú podobu

$$D = 2\sum_{i=1}^{n}\left\{y_i \log\frac{y_i}{\hat{\mu}_i} - (y_i - \hat{\mu}_i)\right\}, \qquad X^2 = \sum_{i=1}^{n}\frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}.$$

2.2 Interpretácia

Interpretácia parametra β_j je, že zmena regresora x_j o jednotku ponásobí priemerný počet udalostí μ_i faktorom $\exp(\beta_j)$, pokiaľ ostatné prediktory zostanú zafixované.

2.3 Overdispersion

Overdispersion je tiež problém, lebo máme jeden parameter μ_i, ktorý je zároveň aj stredná hodnota, aj variancia. Zvyčajne v dátach vidíme, že variancia Y_i je väčšia ako stredná hodnota, čo môže byť spôsobené tým, že udalosti nie sú nezávislé alebo tým, že nedostatočne vysvetlíme parameter μ_i, pomocou X_i, a to napríklad preto, že nepozorujeme nejaké dôležité prediktory. Teda istú skupinu považujeme za homogénnu, aj keď v skutočnosti je to mix nejakých iných skupín, napríklad aj poissonovsky rozdelených. Pri lineárnom modeli toto nebol problém, lebo sme mali separátny parameter pre strednú hodnotu aj pre varianciu Y_i. Disperzný parameter odhadneme z dát a ponásobíme ním štandardné chyby (smerodajné odchýlky) estimátorov, ktoré sa, mimochodom, **nezmenia** (!) Takže overdispersion je problém len pre štandardné chyby estimátorov, takže je to zlý dôvod, prečo úplne zavrhnúť tento model, lebo štandardné chyby vieme odhadnúť robustnejšími metódami. Alebo použijeme iné flexibilnejšie rozdelenie pre Y_i, negatívne binomické.

V prípade, že **nepoznáme mechanizmus**, ktorý mohol spôsobiť príliš veľkú varianciu (overdispersion), môžeme to modelovať nasledovne: $Var(Y) = \phi\mu$ a odhadnúť ϕ pomocou

$$\hat{\phi} = \frac{X^2}{n-p},$$

následne musíme týmto faktorom ponásobiť smerodajné chyby. Keď budeme medzi sebou porovnávať Poissonove modely s overdisperziou, budeme používať F-štatistiku namiesto χ^2. Tento model sa volá **quasi-poissonovský** a metóda sa volá **quasi-maximum likelihood** alebo **pseudo-maximum likelihood**, nakoľko pri odhade sa používa MLE, ale pri štatickej inferencii (napr. pri odhade kovariančnej matice $\hat{\beta}$) sa predpokladá, že skutočná funkcia hustoty Y_i je iná. V tomto prípade môžeme použiť **robustnú kovariančnú maticu** estimátora $\hat{\beta}_{ML}$, ktorá namiesto

$$Var(\hat{\beta}_{ML}) \to_P I^{-1}(\beta_0),$$

predpokladá

$$Var(\hat{\beta}_{ML}) \to_P H(\beta_*)^{-1} I(\beta_*) H(\beta_*)^{-1},$$

kde I je informačná matica, H je matica druhých derivácií a β_* je minimizátor KL divergencie (viď časť (A.4)). Túto metódu na výpočet lepšieho odhadu $Var(\hat{\beta}_{ML})$ nájdete pod názvom **sandwich** covariance matrix v rovnomennej R knižnici založenej na [Zei06].

2.4 Negatívne binomické rozdelenie

Negatívne binomické rozdelenie je bohatšie rozdelenie, ktorého limitným prípadom (pre $Var(Y) \to \mu$ a $D = \frac{1}{k} \to 0$) je Poissonove rozdelenie

$$E(Y) = \mu, \quad Var(Y) = \mu + D\mu^2,$$

kde D je disperzný parameter.

Ako sme prišli k tomuto rozdeleniu? Majme skupinu, ktorá je mix poissonovsky rozdelených náhodných premenných, a nech je to **Gamma** mix týchto distribúcií. Teda náhodná premenná má rozdelenie $Pois(\lambda)$ ale λ nebude fixný parameter ale náhodná premenná, ktorá má $Gamma(\theta, k)$ rozdelenie. Ale v rámci veľkej skupiny sú podskupiny, ktoré majú iné väčšie alebo menšie priemery - väčšie alebo menšie λ. Pre túto náhodnú premennú λ platí: $E(\lambda) = \theta k = \mu$, $Var(\lambda) = \theta^2 k = \mu^2/k$ (táto parametrizácia sa volá NB2) a teda $E(Y) = E(E(Y|\lambda)) = \mu$ a $Var(Y) = E(Var(Y|\lambda)) + Var(E(y|\lambda)) = \mu + Var(\lambda) > \mu$. Potom

$$P(Y = y) = \binom{y + k - 1}{k - 1} \left(\frac{\mu}{\mu + k}\right)^y \left(\frac{k}{\mu + k}\right)^k.$$

Potom $E(Y) = \mu$, $\quad Var(Y) = \mu + \frac{1}{k}\mu^2 = \mu + D\mu^2$, pre $D = 1/k$. Negatívne binomické rozdelenie (nech Z je rozdelené NB) popisuje počet nezávislých pokusov s fixnou pravdpodobnosťou $\frac{\mu}{\mu+k}$, kým nastane k-ty "úspech", v tomto prípade Y je, o koľko viacej ako k musíme mať pokusov, aby nastal k-ty "úspech"($Y = Z - k$). Negatívne binomické rozdelenie teda dobre modeluje situácie, kde napr. systém zvládne k zlyhaní. V prípade NB rozdelenia $\eta_i = \log \frac{\mu}{\mu+k}$.

Porovnanie Poissonovho a Negatívne binomického GLM modelu. Neformálne môžeme porovnať tieto modely na základe AIC. Ďalej, nakoľko Poissonov model je špeciálnym prípadom pre $D = 0$, avšak testujeme parameter D na okraji parametrického priestoru, lebo $D \geq 0$, preto likelihood ratio **nemá** asymptoticky χ^2 rozdelenie. Ale je to mix dvoch distribúcií: $0.5\delta_0 + 0.5\chi^2_1$, kde δ_0 je Diracova miera v bode 0, preto správna p-hodnota pre test $H_0 : D = 0$ je polovica z tej z χ^2_1 testu ([SL87]).

Iný test na overdispersion: Chceme otestovať $H_0 : D = 0$ voči alternatíve $H_1 : D > 0$ (alebo $D < 0$ pri underdispersion), vieme, že $Var(Y) = \mu + D\mu^2$, teda pozrieme sa na parameter D v nasledovnej lineárnej regresii bez interceptu

$$\frac{(y_i - \hat{\mu}_i)^2 - y_i}{\hat{\mu}_i} = D\frac{\hat{\mu}_i^2}{\hat{\mu}_i} + u_i,$$

detaily môžete nájsť v [CT05] a [CT10].

Existuje aj **iná parametrizácia** (NB1) negatívne binomického modelu ($E(\lambda) = \mu$ a $Var(\lambda) = \mu + \frac{\mu}{k}$), ktorá uvažuje, že pre veľké k ide variancia Y k poissonovskej variancii μ. Táto parametrizácia vedie k $E(Y) = \mu$ a $Var(Y) = \mu(1 + k)/k$, teda variancia je lineárna v μ (narozdiel od kvadratickej ako v prípade pôvodnej parametrizácie, ktorá nesie názov NB2). Narozdiel od NB2, pri NB1 modeli β a k **nie sú** ortogonálne parametre a teda $\hat{\beta}$ nie je konzistentý odhad, keď model pre strednú hodnotu platí, ale skutočné rozdelenie Y nie je negatívne binomické.

2.5 Príliš veľa alebo príliš málo núl

Môže sa nám stať, že v dátach pozorujeme oveľa viacej núl, ako by predikoval Poissonov model, toto je **iný** problém ako overdispersion. Ako sa s tým vysporiadať? Sú na to dva základné modely.

Hurdle Model uvažuje, že nuly sú generované iným procesom ako čísla väčšie ako nula, teda predpokladamé, že nuly pochádzajú z distribúcie f_{zero}, teda $f_{zero}(0) = P(Y = 0)$ a kladné hodnoty pochádzajú z distribúcie $f_{count}(y|y > 0) = f_{count}(y)/(1 - f_{count}(0))$. Preto pozorované Y pochádzajú z distribúcie

$$f(y) = \begin{cases} f_{zero}(0), & \text{if } y = 0, \\ \frac{1-f_{zero}(0)}{1-f_{count}(0)} f_{count}(y), & \text{if } y \geq 1. \end{cases} \tag{2.1}$$

distribúcie f_{count} a f_{zero}, môžu byť napríklad poisonovské alebo NB a sú odhadnuté separátne pomocou ML. Poznamenajme, že rôzne regresory môžu byť použité na odhadnutie f_{count} a f_{zero}. Aká je motivácia tohoto modelu? Keď pacient už raz prišiel k doktorovi, jeho ďalšia návšteva sa bude riadiť iným procesom. Ak $f_{count} = f_{zero}$, tak sme v obyčajnom Poissonovom alebo NB GLM modeli, takže tento model je bohatší. Na druhej strane musíme odhadovať dvakrát toľko koeficientov.

Zero-inflated Model Podobne ako v predošlom modeli, len tu je f_{zero} binárne rozdelené, teda s pravdepodobnosťou $f_{zero}(0)$ je $y = 0$ a s pravdepodobnosťou $1 - f_{zero}(0)$ pochádza y z f_{count}. Model

teda uvažuje, že pozorované Y pochádzajú z distribúcie

$$f(y) = \begin{cases} f_{zero}(0) + (1 - f_{zero}(0)) f_{count}(0), & \text{if } y = 0, \\ (1 - f_{zero}(0)) f_{count}(y), & \text{if } y \geq 1. \end{cases} \tag{2.2}$$

Regresný model pre f_{zero} je logit a pre f_{count} je to Poissonova alebo NB distribúcia.

2.6 Literatúra

Okrem hlavnej čítanky of Juliana Farawaya [Far05] a kníh Alana Agrestiho [A+07], [AK11], [Agr15] sú to knihy od Camerona a Trivediho [CT05],[CT10] a Rodriguezove poznámky ku kurzu o GLM na Princetone [Rod15]. O priveľa nulách sa dá dočítať napr. tu [ZKJ07].

3 Kontingenčné tabuľky

Budeme uvažovať situáciu, keď máme 2 alebo viac kategorických dát - tieto môžu byť **nominálne** (bez poradia - žltý, zelený, modrý) alebo **ordinálne** (silno nesúhlasím - nesúhlasím - súhlasím - silno súhlasím).

3.1 [2x2] tabuľky

Chceme vedieť, či má vlastnosť B vplyv na kvalitu súčiastky oproti vlastnosti A. V dátach pozorujeme počty súčiastok kvalitné-nekvalitné s vlastnosťou A alebo B.

3.1.1 Výberové schémy

Sú rôzne spôsoby akými môžeme tieto počty merať. Môžeme odmerať všetky počty za určitý čas, táto situácia bude modelovaná **Poissonovským** rozdelením. Alebo môžeme náhodne vybrať fixný počet súčiastok a potom spočítať počty v rôznych kategóriách, čo je rozumné modelovať **multinomickým** rozdelením.

	A	B	
✓	y_{11}	y_{12}	$n_{1.}$
✗	y_{12}	y_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	n

Ak samplujeme s fixovaným počtom súčiastok s vlastnosťami A a B, potom tomu zodpovedá **binomické** rozdelenie. V prípade, že nadizajnujeme experiment tak, aby sme dostali požadovaný počet dobrých a zlých súčiastok a zároveň fixovaný počet s vlastnosťami A a B, potom sa pravdepodobnostná distribúcia počtov súčiastok v jednotlivých kategóriách dá popísať **hypergeometrickým** rozdelením.

- **Poissonove** rozdelenie
  Log-likelihood (bez členov nezahrňujúcich odhadované parametre)

$$\log L = \sum_i y_i \log(\mu_i),$$

  Linková funkcia

$$\log \mu = \gamma + \alpha_i + \beta_j,$$

  pozor i a j teraz definujú kategórie dvojsmernej tabuľky, predtým i je index jednotlivých pozorovaní. Nakoľko by po pridaní interakčného člena bol model saturovaný (mal by devianciu 0), deviancia D_P z tohoto modelu sa dá použiť ako testovacia χ^2_1 štatistika na testovanie signifikantnosti interakcie.
- **Multinomické** rozdelenie
  Log-likelihood (bez členov nezahrňujúcich odhadované parametre)

$$\log L = \sum_{i,j} y_{ij} \log p_{ij},$$

  ak by sme chceli otestovať hypotézu nezávislosti dvoch kategórií, teda $\forall i = j : p_{ij} = p_i p_j$, maximum likelihood odhad pre proporcie p_{ij} za predpokladu nulovej hypotézy je $\hat{\mu}_{ij} = n\hat{p}_i\hat{p}_j = \sum_i y_{ij} \sum_j y_{ij}/n$. Pre saturovaný model platí $\hat{\mu}_{ij} = y_{ij}$, preto je deviancia rovná

$$D_M = 2\sum_{i,j} y_{ij} \log(y_{ij}/\mu_{ij})$$

  a ide o presne tú istú hodnotu ako deviancia z modelu z Poissonovského rozdelenia. Toto však nie je prekvapivé, pretože ak $Y_1, \cdots, Y_k$ sú nezávislé s rozdelením $Pois(\lambda_k)$, potom distribúcia $Y_1, \cdots, Y_k$ za podmienky $\sum_i Y_i = n$ je multinomické rozdelenie s parametrom $p_j = \lambda_j / \sum_i \lambda_i$.
- **Binomické** rozdelenie
  Log-likelihood a deviancia binomického regresného modelu bola popísaná v predošlej časti textu. Deviancia D_B modelu iba s interceptom, teda za platnosti hypotézy homogeneity (teda platí model len s interceptom), je rovnaká ako D_M a D_P:

$$D_P = D_M = D_B$$

- **Hypergeometrické** rozdelenie - máme zafixované marginálne súčty a chceme vedieť ako bude vyzerať tabuľka. Za predpokladu nezávislosti vyzerá Likelihood nasledovne

$$L = \frac{(y_{11}+y_{12})!(y_{11}+y_{21})!(y_{12}+y_{22})!(y_{21}+y_{22})!}{y_{11}!y_{12}!y_{21}!y_{22}!n!}.$$

  Pri zafixovanom y_{11} vieme ostatné členy dopočítať, nakoľko čiastočné súčty sú fixované. Preto môžeme sčítať pravdepodobnosť všetkých možných situácií extrémnejších ako tie $\{y_{ij}\}$, ktoré pozoriujeme v dátach. Toto sa nazýva **Fischerov exaktný test**. Výhodou je, že nie je založený na žiadnej aproximácii, preto funguje aj pri ľubovoľne malej dátovej vzorke. Problémom môže byť ho vyrátať pri väčšom n.

Tabuľka 1: Rôzne modely na výber dát (sampling schemes). Symboly ✓a ✗ označujú či je subjekt s vlasnosťami A alebo B v poriadku. Modrou sú označené kvantity, ktoré sú zafixované predtým ako pozorujeme počty subjektov v rôznych kategóriách $\{y_{ij}\}$.

(a) Truth

	A	B	
✓	p_{11}	p_{12}	$p_{1.}$
✗	p_{12}	p_{22}	$p_{2.}$
	$p_{.1}$	$p_{.2}$	1

(b) Poisson

	A	B	
✓	y_{11}	y_{12}	?
✗	y_{12}	y_{22}	?
	?	?	?

(c) Multinomial

	A	B	
✓	y_{11}	y_{12}	?
✗	y_{12}	y_{22}	?
	?	?	n

(d) Binomial

	A	B	
✓	y_{11}	y_{12}	?
✗	y_{12}	y_{22}	?
	$n_{.1}$	$n_{.2}$	n

(e) Hypergeometric

	A	B	
✓	y_{11}	y_{12}	$n_{1.}$
✗	y_{12}	y_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	n

3.2 [IxJ] tabuľky

Teraz nás zaujíma vzťah medzi dvomi kategorickými premennými, kde počet kategórií je I a J. Môžeme **testovať nezávislosť** týchto dvoch faktorov. Nulovou hypotézou je $\forall i,j : p_{ij} = p_{i.}p_{.j}$ a na jej otesovanie môžeme použiť Pearsonove X^2, ktoré má tvar $X^2 = \sum_{i,j} \frac{(O_{ij}-E_{ij})^2}{E_{ij}} = \frac{(y_{ij}-np_{ij})^2}{np_{ij}}$, kde $O_{ij} = y_{ij}$ sú pozorované počty a $E_{ij} = np_{ij}$ sú očakávané počty za predpokladu nulovej hypotézy v kategórii (i,j). Počet stupňov voľnosti je rozdiel počtu stupňov voľnosti saturovaného modelu $IJ-1$ a počtu stupňov voľnosti modelu za predpokladu nezávislosti $(I-1)+(J-1)$, takže $IJ-I-J+1 = (I-1)(J-1)$. Príklad: oči $\in$ {zelené, orieškové, modré, hnedé} a vlasy $\in$ {čierne, hnedé, červené, blond} a $I = J = 4$ a preto je počet stupňov voľnosti $(4-1)\cdot(4-1) = 9$.

Je nesmierne užitočné dáta vizualizovať. Okrem tradičných zobrazovacích techník je užitočným nástrojom aj **korešpondenčná analýza**.

Začneme tým, že sa v poissonovskej regresii bez interakčných členov snažíme vysvetliť počet pozorovaní v jednotlivých kategóriách pomocou rôznych kategórií. Koeficienty v tejto regresii nám však nič nehovoria o vzťahu týchto dvoch kategorických premenných, naopak, táto regresia uvažuje nezávislosť týchto dvoch premenných. Ak tento model nepopisuje dáta dobre (príliš veľká deviancia), predpoklad nezávislosti zamietame a potom má zmysel pozrieť sa bližšie na nevysvetlenú variáciu v počtoch pozorovaní v jednotlivých kategóriách (za predpokladu nezávislosti), napríklad cez Pearsonove reziduá.

Ak by bol model nezávislosti vhodný, v reziduách by sme nič neuvidieli, teraz sa však môžeme pozrieť na **štruktúru reziduí**. Matica reziuduí R je $I \times J$ matica. Tú vieme rozložiť pomocou **singular value dekompozície** (SVD)

$$R_{I\times J} = U_{I\times W} D_{W\times W} V^T_{W\times J},$$

kde D je diagonálna matica a $W = \min\{I, J\}$. Geometria SVD je názorne vysvetlená v [Aus16]. Nech je podstatná časť variácia reziduí vysvetlená pomocou dvoch najväčších singulárnych hodnôt.

$$R_{ij} \approx U_{i1}d_1V_{j1} + U_{i2}d_2V_{j2} = (U_{i1}\sqrt{d_1})(V_{j1}\sqrt{d_1}) + (U_{i2}\sqrt{d_2})(V_{j2}\sqrt{d_2}) \equiv U^*_{i1}V^*_{j1} + U^*_{i2}V^*_{j2}$$

Potom vieme reziduá zobraziť (v jednom grafe) v priestore (U^*_{i1}, U^*_{i2}) a (V^*_{j1}, V^*_{j2}), čo je výhodné.

- Kategórie ďaleko od $(0,0)$ sú netypické.
- Ak sú kategórie i (zobrazené v priestore (U^*_{i1}, U^*_{i2})) a j (zobrazené v priestore (V^*_{j1}, V^*_{j2})) blízko seba, sú pozitívne asociované.
- Ak sú stredovo súmerné so stredom súmernosti v bode $(0,0)$, potom sú negatívne asociované.
- Ak sú dve i kategórie blízko seba (alebo dve j kategórie blízko seba), môžme uvažovať o ich spojení do jednej kategórie.

Zhrnutie: Z regresie predpokladajúcej nezávislosť premietneme reziduá do dvojrozmerného ortonormálneho priestoru, kde lepšie vidíme ich asociáciu.

3.3 Matched pairs

Doteraz sme sa zaoberali situáciu, že máme dve kategorické premenné toho istého subjektu. Teraz budeme pozorovať takú istú kategorickú premennú pre dva spárované objekty. Napríklad kvalita zraku pravého a ľavého oka, kde spárovanie je prirodzené podľa ľudí (teda majiteľov daných očí).

Zaujímavé hypotézy, ktoré môžeme testovať

- (I) **Nezávislosť**: $\forall i,j : p_{ij} = p_{i.}p_{.j}$ toto môžeme testovať pomocou Pearsonovho X^2.
- (S) **Symetria**: $\forall i,j : p_{ij} = p_{ji}$. Testujeme to tak, že vytvoríme nové dvojité faktory u ktorých nezáleží na poradí. Napr. $(\text{dobrý}, \text{uspokojivý})$ a $(\text{uspokojivý}, \text{dobrý})$ bude jeden a ten istý faktor a použijeme napríklad poissonovskú regresiu (pozrieme sa na devianciu).
- (QS) **Kvázi-symetria**:$\forall i,j : \frac{p_{ij}}{p_{i.}p_{.j}} = \frac{p_{ji}}{p_{i.}p_{.j}}$, je to isté ako nasledovný model $p_{ij} = \alpha_i\beta_j\gamma_{ij}$ takže môžeme použiť poissonovskú regresiu na $\log y_{ij} = \log np_{ij} = \log n + \log\alpha_i + \log\beta_j + \log\gamma_{ij}$ a pozrieť sa na devianciu, či tento model dobre popisuje dáta.
- (MH) **Marginálna homogenita**: $\forall k : p_{k.} = p_{.k}$. Nakoľko (MH) + (QS) $\implies$ (S), môžeme porovnať modely založené na (QS) a (S). Ak je fit týchto modelov výrazne rôzny, potom zamietame hypotézu o platnosti (MH).
- (QI) **Kvázi-nezávislosť**: $\forall i \neq j : p_{ij} = p_{i.}p_{.j}$ otestujeme tak, že vyhodíme dáta na diagonále a otestujeme nezávislosť ako predtým (cez Pearsonove X^2 alebo pomocou poissonovskej regresie).

3.4 Trojsmerové tabuľky a Simpsonov paradox - Three-way tables

Predstavme si, že nás zaujíma ako vplýva fajčenie na dĺžku života. Pozrieme sa na istú skupinu ľudí či fajčia alebo nie a o 20 neskôr sa pozrieme, či umreli alebo ešte žijú. Keď sa pozrieme do dát, tak medzi fajčiarmi je 76% živých avšak medzi nefajčiarmi je ich len 68%. Toto sa môže zdať na prvý pohľad prekvapivé a jeden by sa mohol nazdávať, že fajčenie predlžuje život (!) Keď sa však pozrieme na podskupiny v rôznom veku, napr. na 35-44 ročných, vidíme, že naopak, fajčiari sa dožívajú viacej. A to dokonca v v každej vekovej kategórií. Tento fenomén je pre mnohých ľudí natoľko prekvapivý, že má svoj vlastný názov **Simpsonov paradox**. V prípade fajčiarov je dôvod ten, že mladší ľudia, ktorí sa prirodzene dožívajú vyššieho veku fajčia oveľa viacej. Známy je príklad z prijímacích skúšok z roku 1973 z University of California Berkeley, ktorá bola súdená kvôli tomu, že kým 46% mužov bolo prijatých, pri ženách toto čislo bolo len 30%. Keď sa však bližšie pozrieme na dáta zistíme, že ženy boli úspešnejšie na každom departmente (a to dokonca štatisticky významne). Znalivý paradox má jednoduché vysvetlenie: ženy sa v priemere viacej hlásili na prestížnejšie departmenty, kde podiel prijatých uchádzačov alebo uchádzačiek je nižší. Vizualizácia tohoto problému aj s užitočnými odkazmi a vysvetleniami nájdete tu [LP].

Pre danú konkrétnu vekovú podskupinu môžeme otestovať **nezávislosť** napríklad pomocou Fisherovho exaktného testu, kde nulová hypotéza je, že pomer šancí (odds ratio) je rovný jeden. Ak však chceme testovať združenú hypotézu nezávislosti pre všetky vekové podskupiny naraz, potrebujeme čosi rozumnejšie. Pre tabuľky typu 2x2xK, kde vzťah je podobný pre každú z K podskupín máme Mantel-Haenszel test a jeho testovacia štatistika má tvar

$$\frac{(\sum_k y_{11k} - \sum_k E(y_{11k}) - 1/2)^2}{\sum_k var(y_{11k})} \sim_{n\to\infty} \chi^2_1$$

Pri trojsmerných tabuľkách je kompletná pravdepodobnostná distribúcia popísaná čislami $\{p_{ijk}\}$, označme merginálne distribúcie pre jednotlivé kategórie ako p_i, p_j, p_k. Môžeme sa venovať nasledovným zaujímavým otázkam/hypotézam:

- **Vzájomná nezávislosť**

$$\forall i,j,k: \; p_{ijk} = p_i p_j p_k$$

$$\log E(Y_{ijk}) = \log n + \log p_i + \log p_j + \log p_k$$

  Testovať môžeme pomocou Pearsonovho X^2 alebo poissonovskej regresie.
- **Združená nezávislosť**

$$\forall i,j,k: \; p_{ijk} = p_{ij} p_k$$

$$\log E(Y_{ijk}) = \log n + \log p_{ij} + \log p_k$$

  Testovať môžeme pomocou Pearsonovho X^2 alebo poissonovskej regresie, kde kategórie i a j sú aj s interakčným členom.
- **Podmienená nezávislosť**

$$\forall i,j,k: \; p_{ij|k} = p_{i|k} p_{j|k},$$

  čo nie je nič iné ako $p_{ijk} = p_{ik}p_{jk}/p_k$.

$$\log E(Y_{ijk}) = \log n + \log p_{ij} + \log p_{jk} - \log p_k$$

  Testovať môžeme pomocou Pearsonovho X^2 alebo poissonovskej regresie, kde kategórie i a j a kategórie j a k sú aj s interakčným členom.
- **Uniformná asociácia**

$$\log E(Y_{ijk}) = \log n + \log p_i + \log p_j + \log p_k + \log p_{ij} + \log p_{ik} + \log p_{jk}$$

  Tento model predikuje pre každú skupinu k rovnaké log-odds ratio, a to nasledovné:

$$(EY_{11k}EY_{22k})/(EY_{12k}EY_{21k}).$$

Pri trojsmerných tabuľkách sa dá pozerať na jednu premennú ako na vysvetľovanú a na ďalšie dve ako vysvetľujúce a pozerať sa na problém ako na model **binomických dát**. V prípade príkladu s fajčením by sme vysvetľovali to či pacienti prežili 20 rokov pomocou toho či fajčili a veku. Vlastnosti takéhoto modelu

- binomický model s interakčným prvkom je nasýtený
- binomický model model bez interakcií je ekvivalentný poissonovskému modelu za predpokladu uniformnej asociácie
- binomický model len s interceptom je ekvivalentný poissonovskému modelu za predpokladu združenej nezávislosti vysvetľovanej premennej a dvoch vysvetľujúcich premenných

Binomický model preferujeme keď sme v situácii keď je jedna z premenných jasne identifikovateľná ako vysvetľovaná. Ak je vzťah medzi premennými symetrickejší preferujeme Poissonovský model.

Korešpondenčná analýza je možná pokiaľ spojíme dve faktory do jedného, teda uvažujeme všetky možné kombinácie.

3.5 Usporiadané premenné

Niektoré premenné sú kategorické ale usporiadané. Môžeme

- túto informáciu zahodiť a tváriť sa, že žiadne usporiadanie neexistuje.
- – použiť metódy pre usporiadané multinomické dáta (o tomto viacej nabudúce)

– očíslovať naše usporiadané kategorické premenné. Netreba zabudnúť vyskúšať aj iné očíslovanie a zistiť ako veľmi sú naše výsledky citlivé na toto očíslovanie. Rôzne skóre môže viesť k rôznym výsledkom a to nám môže pomôcť lepšie spoznať problém.

Teraz sa budeme venovať tejto poslednej možnosti. Ak sú dáta intervalové, môžeme použiť stred intervalu. Uvažujme nasledovný model

$$\log EY_{ij} = \log np_{ij} = \log n + \alpha_i + \beta_j + \gamma u_i v_j,$$

kde nezávislosť testuje hypotéza $\gamma = 0$. Alebo môžeme byť v situácii, kde len jednu premennú budeme považovať za ordinálnu, potom náš model bude vyzerať nasledovne

$$\log EY_{ij} = \log np_{ij} = \log n + \alpha_i + \beta_j + u_i \gamma_j.$$

3.6 Literatúra

Oproti našej čítanke [Far05] ešte napríklad kurz z Penn State University [Uni16].

4 Multinomické dáta

Sme v situácii, že vysvetľovaná premenná je kategorická. Môže (usporiadaná premenná) alebo nemusí (nominálna premenná) byť pre ňu definované usporiadanie. Majme individuála i a náhodnú premennú, ktorá môže nadobúdať J rôznych hodnôt. Označme $p_{ij} = P(Y_i = j)$ pre ktoré musí platiť $\sum_{j=1}^{J} p_{ij} = 1$.

V dátach pozorujeme Y_{ij} - teda počet pozorovaní $Y_i = j$ a označíme $n_i = \sum_j Y_{ij}$. Pravdepodobnostná distribúcia vektora Y_{ij} (Jx1 vektor pre konkrétne i) za podmienky, že poznáme n_i je

$$P(Y_{i1} = y_{i1}, \cdots, Y_{iJ} = y_{iJ}) = \frac{n_i!}{y_{ij}! \cdots y_{iJ}!} p_{i1}^{y_{i1}} \cdots p_{iJ}^{y_{iJ}}.$$

Hovoríme jej **multinomické rozdelenie**. (Pravdepodobosť jednej konkrétnej realizácie je $p_{i1}^{y_{i1}} \cdots p_{iJ}^{y_{iJ}}$ keď záleži na poradí. Musíme to teda ponásobiť počtom všetkých kombinácií s opakovaním $\frac{n_i!}{y_{ij}! \cdots y_{iJ}!}$).

Špeciálnym prípadom je keď $n_i = 1$, potom má náhodná premenná Y_{ij} **kategorické** rozdelenie. Použili sme označenie individuál pre i avšak to môže byť aj skupina. Ďalším špeciálnym prípadom je binomické rozdelenie, a to pre $J = 2$.

4.1 Multinomický logit

Multinomický logit model je prirodzeným rozšírením binárneho logitu. Kým pri binárnych dátach sme mali udalosti typu úspech-neúspech, takže sme mali dve kategórie, teraz ich je viacej, konkrétne J.

Chceme urobiť model, preto je našou ambíciou prepojiť parametre pozorovaného multinomického Y_{ij} s ostatnými vysvetľujúcimi X_i. Urobíme to rovnako ako pri binomickom modeli, teda keď $J = 2$. Teraz si zvolíme jeden výstup ako bázický - zafixovaný a všetky pravdepodobnosti budeme modelovať relatívne k tomuto bázickému výstupu. Môže to byť napríklad $j = 1$ alebo $j = J$, vôbec na tom nezáleží a vedie to k úplne totožným výsledkom.

$$\forall j = 2, \cdots, J: \quad \log \frac{p_{ij}}{p_{i1}} = x_i^T \beta_j = \eta_{ij}$$

Dá sa ľahko ukázať, že $p_{ij} = \frac{\exp(\eta_{ij})}{1+\sum_{j=2}^{J} \exp(\eta_{ij})}$.

Teraz však máme $J - 1$ rovníc a pre každý log-odds máme iný vektor parametrov (ak máme K regresorov, potom je β_j taktiež vektor dĺžky K). Parametre odhadujeme metódou maximum likelihood. V programe R na odhadovanie máme funkciu `multinom` z knižnice `nnet` (neexistuje žiadne mystické prepojenie multinomického rozdelenia a neurónových sietí, ide len o to, že sa dá na odhadovanie oboch použiť podobná optimalizačná metóda.)

Výhodou multinomického rozdelenia je, že ak spojíme viacero kategórií do jednej, potom výsledné rozdelenie je taktiež multinomické. Preto môžeme výstup typu (silno nesúhlasím - skôr nesúhlasím - skôr súhlasím - silno súhlasím) modelovať aj jednoduchšie, napríklad takto (nesúhlasím - súhlasím).

Interpretácia parametrov je rovnaká ako v prípade binomického modelu, tu však relatívne k bázickej kategórii. Na porovnanie vnorených modelov môžeme použiť rozdiel deviancií, ktorý má χ^2 rozdelenie.

4.1.1 Predpoklad nezávislosti irelevantných alternatív (Independence of Irrelevant Alternatives)

Dôležitým dôsledkom multinomického logit modelu je to, že pomer pravdepodobností dvoch alternatív (napríklad vlak, autobus) nezávisí od vlastností iných alternatív (auto). Toto nie je splnené ak sú niektoré alternatívy substitúty. Klasický príbeh na porušenie tohoto predpokladu je nasledovný: modelujete akým spôsobom sa ľudia presúvajú z miesta A do miesta B, nech sú nasledovné 4 alternatívy rovnako pravdepodobné (červený autobus, modrý autobus, auto, vlak), teda proporcie ľudí využívajúcich tieto alternatívy sú $(0.25, 0.25, 0.25, 0.25)$. Čo sa stane ak premaľujeme modrý autobus na červeno? Podľa nášho modelu by mali byť pomery medzi kategóriami zachované, takže by sme mali pozorovať $(0.33, 0.33, 0.33)$, avšak v praxi je zrejmé, že na farbe až tak nezáleží a ľudia využívajúci modrý bus budú radšej využívať červený autobus ako auto alebo vlak, takže proporcie budú $(0.5, 0.25, 0.25)$. Toto

je však v rozpore s multinomickým logitom. Testy na IIA sú popísané v [LF06] v kapitole 6, ale nestoja za veľa.

4.1.2 Multinomický logit ako GLM

Nakoľko Poissonove rozdelenie podmienené počtom nastaných udalostí je multinomické rozdelenie, môžeme odhadnúť multinomický model pomocou poissonovej regresie. Musíme si však prerobiť dáta. Pre každé pozorovanie v našej vzorke si dodefinujeme **response faktor**, ak máme 413 pozorovaní, tak si vytvoríme faktor, ktorý nadobúda 413 rôznych hodnôt. V prípade, že máme individuálne dáta, tak každé 1 pozorovanie prerobíme na J pozorovaní. Ak napríklad máme individuála, ktorý volil demokratov, tak sa na toto porozovanie pozeráme ako na 3 pozorovania: že 1 krát volil demokratov, 0 krát volil indififerentne a 0 krát volil republikánov ($J = 3$ v tomto prípade). Pre každé z týchto J porozovaní máme aj **kategorický faktor** (demokrat, indiferentný, republikán). Takto sme si vlastne **preorganizovali dáta** do kontingenčnej tabuľky (viď kapitola o kontingenčných tabuľkách).

Null model, teda model kde počet udalostí vysvetľujeme response faktorom a kategorickým faktorom bez interakčného členu zodpovedá hypotéze, že všetky výstupy (demokrat, indiferentný, republikán) majú fixnú pravdepodobnosť pre všetkých individuálov i, inými slovami: je to multinomický model len s interceptom, bez žiadnych vysvetľovaných premenných.

Null model - kategórie majú fixnú pravdepodobnosť

- Multinomický model len s interceptom na pôvodných dátach (veľkosti N)

$$Y_{ij} \sim Multinom(p_{i1}, p_{i2}, \cdots, p_{iJ}), \quad \forall i = 1 \ldots N, \forall j = 2, \cdots, J: \quad \log \frac{p_{ij}}{p_{i1}} = c_j$$

- Poissonovský model bez interakčného člena na reorganizovaných dátach (veľkosti $J \times N$)

$$Y_{ij} \sim Pois(\mu_{ij}), \quad \log(\mu_{ij}) = \gamma + rf_i + cf_j$$

  kde rf_i je response faktor a cf_j je kategorický faktor.[2]

Tieto dva modely vedú k úplne totožnej deviancii, sú ekvivalentné.

Teraz ak chceme napr. modelovať efekt mzdy na pravdedpodobnosť voľby jednotlivých strán, pridáme do poissonovskej regresie interakčný člen s kategorickým faktorom.

Model s regresorom - pravdepodobnosť kategórie závisí od regresora x

- Multinomický model s regresorom na pôvodných dátach (veľkosti N)

$$Y_{ij} \sim Multinom(p_{i1}, p_{i2}, \cdots, p_{iJ}), \quad \forall i = 1 \ldots N, \forall j = 2, \cdots, J: \quad \log \frac{p_{ij}}{p_{i1}} = c_j + \beta_j x_i$$

- Poissonovský model s interakčným členom na reorganizovaných dátach (veľkosti $J \times N$)

$$Y_{ij} \sim Pois(\mu_{ij}), \quad \log(\mu_{ij}) = \gamma + rf_i + cf_j + cf_j * x_j$$

  kde rf_i je response faktor a cf_j je kategorický faktor.

Tieto dva modely vedú k úplne totožnej deviancii, sú ekvivalentné.

Musíme si však dať pozor na to, ktorá je bázická (referenčná) kategória.

[2] Poznamenajme, že $\log(\mu_{ij}) = \gamma + rf_i + cf_j$ je schematický zápis. Pri 413 pozorovaniach máme 413 response faktorov a efektívne 412 dummy premenných v regresii so 412timi prislúchajúcimi regresnými koeficientami.

4.2 Lineárna diskriminačná analýza

Majme $p \times 1$ vektor prediktorov x_i a zoberme si odhad jej kovariančnej matice $(p \times p)$

$$S = \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T$$

Táto matica sa dá rozložiť ako $S = W + G$, teda within group kovariančná matica

$$W = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (x_{gi} - \bar{x}_g)(x_{gi} - \bar{x}_g)^T$$

a between group kovariančná matica

$$B = \sum_{g=1}^{G} (\bar{x}_g - \bar{x})(\bar{x}_g - \bar{x})^T.$$

Tu nejde o nič iné ako aplikáciu $Var(X) = E_G Var(X|G) + Var(E_G(X|G))$

Teraz našou úlohou je nájsť takú lineárnu kombináciu vektora $a^T x$, že between variancia v smere a, teda $a^T Ba$ bude čo najväčšia oproti within variancii. Teda vyberieme a, ktoré maximalizuje

$$\frac{a^T Ba}{a^T Wa}$$

4.3 Hierarchická vysvetľovaná premenná

Predstavme si výstup typu (žiadna choroba, choroba1, choroba2, choroba3). V dátach pozorujeme situáciu, že najviac individuálov je v prvej kategórii, teda nemajú žiadnu chorobu. Na problém sa môžeme pozerať dvoma spôsobmi:

- Multinomický model z poslednej podsekcie:

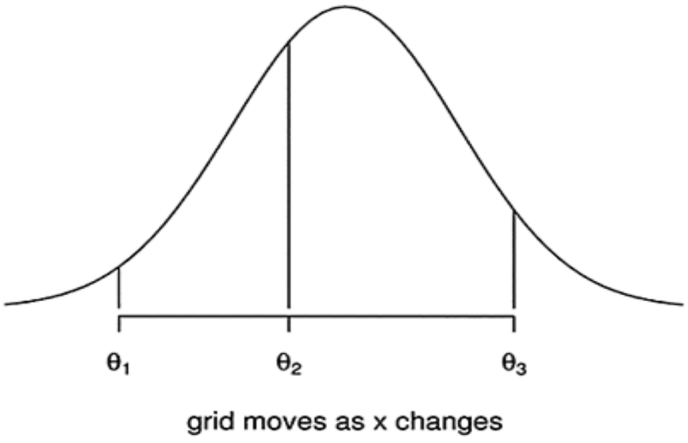
$$P(Y_i = j) \sim p_{i1}^{y_{i1}} p_{i2}^{y_{i2}} p_{i3}^{y_{i3}} p_{i4}^{y_{i4}}$$

- Hierarchický model: pravdepodobnosť prepíšeme iným spôsobom

$$P(Y_i = j) \sim p_{i1}^{y_{i1}} (p_{i2}+p_{i3}+p_{i4})^{y_{i2}+y_{i3}+y_{i4}} \times \left(\frac{p_{i2}}{p_{i2}+p_{i3}+p_{i4}}\right)^{y_{i2}} \left(\frac{p_{i3}}{p_{i2}+p_{i3}+p_{i4}}\right)^{y_{i3}} \left(\frac{p_{i4}}{p_{i2}+p_{i3}+p_{i4}}\right)^{y_{i4}}$$

  - Má chorobu? $p_{i1}^{y_{i1}} (p_{i2} + p_{i3} + p_{i4})^{y_{i2}+y_{i3}+y_{i4}}$
    Môžeme použiť napríklad binárny logit
  - Ak áno, akú chorobu? $\left(\frac{p_{i2}}{p_{i2}+p_{i3}+p_{i4}}\right)^{y_{i2}} \left(\frac{p_{i3}}{p_{i2}+p_{i3}+p_{i4}}\right)^{y_{i3}} \left(\frac{p_{i4}}{p_{i2}+p_{i3}+p_{i4}}\right)^{y_{i4}}$
    Môžeme použiť napríklad multinomický logit.

Na čo to vlastne je? Vďaka takému prístupu môžeme lepšie porozumieť akým spôsobom sú vysvetľujúce premenné ovplyvňujú pravdpodobnosť choroby. Môžeme byť napríklad v situácii, že nejaký regresor predikuje pravdepodobnosť nejakej choroby ale na druhej strane nevie medzi nimi rozlíšiť, ktorá to je. Tento hierarchický trik môže byť výhodný aj v situácii, keď máme usporiadanú premennú (viď ďalšia podsekcia) ale máme tam jednu kategóriu typu 'neviem' alebo 'neodpovedal', ktorá nám kazí to usporiadanie.



Obr. 1: Zdroj: Faraway (2014)

4.4 Usporiadaná kategorická vysvetľovaná premenná

Majme teraz kategorickú premennú na ktorej je definované akési prirodzené usporiadanie. Môžeme použiť skóre a priradiť jej hodnoty ale tým pádom prinesieme do modelu novú informáciu o kvantitatívnych rozdieloch medzi skupinami.

Budeme pracovať s kumulatívnymi distribučnými funkciami. Prečo? Má to výhodu, v tom, že poľahky spojíme dve susedné kategórie do jednej a nič sa nám nezmení. $\gamma_{ij} = P(Y_i \leq j)$ a $\gamma_{iJ} = 1$.

Linkovou funkciou prepojíme regresory s γ_{ij}:

$$g(\gamma_{ij}) = \theta_j - x_i^T \beta,$$

ako funkciu $g(.)$ budeme uvažovať logit, probit a complementary log-log funkciu. Všimnime si, že teraz máme len jeden vektor β (a nie J vektorov ako v prípade neusporiadanej vysvetľovanej premennej).

Ako sme k tomuto došli? Majme premennú Z_i, pre ktorú platí, že $Y_i = j$ vtedy keď $\theta_{j-1} < Z_i \leq \theta_j$. Nech má posunutá Z_i, teda $Z_i - x_i^T\beta$ rozdelenie F, potom:

$$P(Y_i \leq j) = P(Z_i \leq \theta_j) = P(Z_i - x_i^T\beta \leq \theta_j - x_i^T\beta) = F(\theta_j - x_i^T\beta).$$

Ak je F cdf logistickej náhodnej premennej potom $\gamma_{ij} = \frac{\exp(\theta_j - x_i^T\beta)}{1+\exp(\theta_j - x_i^T\beta)}$ a teda máme logit pre kumulatívne pravdepodobnosti γ_{ij}. Ak je F cdf normálne rozdelenej náhodnej premennej tak dostávame probit model, ak je F cdf náhodnej premennej, ktorá má extreme value distribution, potom dostávame komplementárny log-log model.

Voľba linkovej funkcie môže viesť k týmto 3 rôznym modelom:

- Proportional Odds Model
- Usporiadaný Probit Model
- Proportional Hazard Model

Nech teraz $\gamma_j(X_i) = P(Y_i \leq j|X_i)$.

4.4.1 Proportional Odds Model

$$\log \frac{\gamma_j(x_i)}{1-\gamma_j(x_i)} = \theta_j - x_i^T\beta, \quad j = 1, \cdots, J-1$$

Tento model má vlastnosť, že relatívne odds nezávisia od skupiny j. Teda ak sa x_1 zmení na x_2, odds sa zmenia rovnako pre všetky kategórie j úplne rovnako

$$\left(\frac{\gamma_j(x_1)}{1-\gamma_j(x_1)}\right) / \left(\frac{\gamma_j(x_2)}{1-\gamma_j(x_2)}\right) = \exp(-(x_1 - x_2)^T\beta).$$

4.4.2 Usporiadaný Probit Model

$$\Phi^{-1}(\gamma_j(x_i)) = \theta_j - x_i^T\beta, \quad j = 1, \cdots, J-1$$

4.4.3 Proportional Hazard Model

$$\log(-\log(1-\gamma_j(x_i))) = \theta_j + x_i^T\beta, \quad j = 1, \cdots, J-1$$

$$\text{Hazard}(j) = P(Y_i = j | Y_i \geq j) = \frac{P(Y_i = j)}{P(Y_i \geq j)} = \frac{p_j}{1-\gamma_{i,j-1}} = \frac{\gamma_{ij} - \gamma_{i,j-1}}{1-\gamma_{i,j-1}}$$

5 GLM (Generalized Linear Model) všeobecne

Teraz si ideme zovšeobecniť to, čo sme už viackrát videli v predošlých kapitolách. A upraceme to všetko pod jednu strechu.

GLM je definovaný dvoma vecami,

- distribúciou Y
- linkovou funkciou, ktorá prepája $E(Y)$ a lineárny prediktor $\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$.

5.1 Distribúcia

GLM uvažuje, že dáta Y prichádzajú z nasledovnej pravdepodobnostnej distribúcie:

$$f(y|\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right],$$

kde

- θ - parameter polohy (*location parameter*),
- ϕ - parameter disperzie (*dispersion parameter*).

Pre rôzne $\theta, a(\theta), \phi, a(\phi), c(y,\phi)$ dostávame Normálne rozdelenie, Poissonove rozdelenie, binomické rozdelenie alebo mnoho ďalších iných rozdelení. Ide o **exponenciálnu** triedu rozdelení. Pre niektoré rozdelenia je disperzný parameter zafixovaný ako $\phi = 1$, ako napríklad Poissonove alebo binomické. (Negatívne binomické rozdelenie napríklad *nepatrí* do triedy exponenciálnych rozdelení.)

Normálne rozdelenie

$$\theta = \mu, \quad b(\theta) = \frac{\theta^2}{2}, \quad \phi = \sigma^2, \quad a(\phi) = \phi, \quad c(y,\phi) = -\frac{1}{2}\left(\frac{y^2}{\phi} + \log(2\pi\phi)\right)$$

$$f(y|\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right] = \exp\left[\frac{y\mu - \frac{\mu^2}{2}}{\sigma^2} - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right)\right]$$
$$= \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right]$$

Poissonove rozdelenie

$$\theta = \log(\mu), \quad b(\theta) = \exp(\theta), \quad \phi = 1, \quad a(\phi) = 1, \quad c(y,\phi) = -\log(y!)$$

$$f(y|\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right] = \exp\left[\frac{y\log(\mu) - \mu}{1} - \log(y!)\right] = \exp(-\mu)\frac{\mu^y}{y!}$$

Binomické rozdelenie

$$\theta = \log\left(\frac{\mu}{1-\mu}\right), \quad b(\theta) = n\log(1+\exp(\theta)), \quad \phi = 1, \quad a(\phi) = 1, \quad c(y,\phi) = \log\binom{n}{y}$$

$$f(y|\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right] = \exp\left[\frac{y\log\left(\frac{\mu}{1-\mu}\right) - n\log(1+\exp(\log\left(\frac{\mu}{1-\mu}\right)))}{1} + \log\binom{n}{y}\right] =$$
$$= \binom{n}{y}\mu^y(1-\mu)^{n-y}$$

Pre exponenciálnu triedu rozdelení platí

$$E(Y) = \mu = b'(\theta)$$

$$Var(Y) = b''(\theta)a(\phi)$$

Pre normálne rozdelenie $b''(\theta) = 1$, preto $Var(Y)$ nezávisí od $E(Y)$.

Funkciu $b''(\theta)$ nazývame aj variančnou funkciou (*variance function*) a budeme označovať $V(\mu)$.

Tu bolo teraz zadefinovaných veľmi veľa nových symbolov, funkcií ale netreba sa toho báť. Ide o to všimnúť si, že v čom sú všetky tieto rozdelenia a rovnako aj prístupy k modelovaniu dát podobné.

5.2 Linková funkcia

Linková funkcia je spôsob akým prepojíme regresory a priemernú hodnotu Y. Uvažujme lineárnu kombináciu regresorov, ktorú budeme nazývať **lineárny prediktor**

$$\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = x^T\beta$$

Linková funkcia g je funkcia o ktorej GLM predpokladá, že $g(\mu) = g(E(Y)) = \eta$. Toto platí pre každé jedno pozorovanie. V skutočnosti je správnejšie zapisovať $g(E(Y_i|X_i = x_i)) = \eta_i$, pre pozorovanie i, nesmieme totiž zabúdať, že modelujeme podmienené rozdelenie Y za podmienky $X = x$.

V Gaussovskom modeli je linková funkcia jednoduchá, a to konkrétne $\eta = \mu$. Ak by sme zvolili inú linkovú funkciu, dostali by sme $y = g^{-1}(x^T\beta) + \epsilon$. Pozor, toto nie je to isté ako $g(y) = x^T\beta + \epsilon$, teda ako lineárna regresia na transformovaných odozvách (ypsilonoch).

Voľba linkovej funkcie g musí rešpektovať definičný obor parametra. Linkovú funkciu nazývame **kanonická** ak platí $g(\mu) = \theta$, väčšinou používame túto linkovú funkciu.

Distribúcia	Link	Var. funkcia	Deviancia
Normálna	$\eta = \mu$	1	$\sum_i (y_i - \hat{\mu}_i)^2$
Poissonovská	$\eta = \log \mu$	μ	$\sum_i [y_i \log \frac{y_i}{\hat{\mu}_i} - (y_i - \hat{\mu}_i)]$
Binomická	$\eta = \log \frac{\mu}{1-\mu}$	$\mu(1-\mu)/n$	$\sum_i [y_i \log \frac{y_i}{\hat{\mu}_i} - (n_i - \hat{\mu}_i) \log((n_i - y_i)/(n_i - \hat{\mu}_i))]$
Gamma[NEW!]	$\frac{1}{\mu}$	μ^2	$\sum_i [-\log \frac{y_i}{\hat{\mu}_i} - (y_i - \hat{\mu}_i)/\hat{\mu}_i]$
Inverzná Gauss[NEW!]	$\frac{1}{\mu^2}$	μ^3	$\sum_i (y_i - \hat{\mu}_i)^2/(\hat{\mu}_i^2 y_i)$

5.3 Odhad parametrov

Odhad parametrov budeme robiť cez maximum likelihood, log-likelihood vyzerá nasledovne,

$$\sum_i \log L(\theta_i, \phi; y_i) = \sum_i \left(\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} \right] + c(y_i, \phi) \right).$$

Všimnime si, že disperzný parameter ϕ nezávisí od pozorovaní, nemá preto index i. To je tým, že distribúcia regresorov ovplyvňuje len parameter strednej hodnoty a tým aj parameter θ. Analytické riešenie máme len pre Gaussovský model, inak riešenie musíme hľadať numericky. Taktiež poznamenajme, že $\beta \to \eta_i \to \mu_i \to \theta_i \to L_i$ Teraz popíšeme ako jeden konkrétny algoritmus na nájdenie odhadu maximum likelihood vyzerá. Volá sa **Iteratively Reweighted Least Squares**.[3]

Pre jednoduchosť začnime lineárnym modelom $Y = X\beta + \epsilon$, v ktorom je však heteroskedasticita a teda variancia Y nie je konštantá ale je proporciálna nejakej funkcii f lineárneho prediktora ($Var(Y) \propto f(\hat{\eta})$, kde $\hat{y} = \hat{\eta} = x^T\hat{\beta}$). V prípade heteroskedasticity váhujeme pozorovania inverzne variancii, preto $\frac{1}{w} \propto f(\hat{\eta})$. Začneme s rovnakými váhami, odhadneme β, prepočítame váhy až pokým sa parametre β ani váhy príliš nemenia.

V prípade GLM modelu budeme postupovať rovnako, len s jednou modifikáciou. Chceli by sme urobiť regresiu $g(y)$ na X avšak s váhami inverzne proporciálnymi $Var(g(y))$. Pre niektoré prípady však priamo regresovať priamo $g(y)$ nedáva zmysel (pre binomický model máme logit(0) alebo logit(1)). Preto namiesto $g(y)$ budeme budeme na X regresovať jeho lineárnu aproximáciu okolo μ: $g(y) \approx g(\mu) + (y - \mu)g'(\mu) = \eta + (y - \mu)\frac{d\eta}{d\mu} = z$, s tým, že variancia tejto lineárnej aproximácie bude $Var(z) = \left(\frac{d\eta}{d\mu}\right)^2 V(\hat{\mu})$ a váhy nastavíme ako $\frac{1}{w} = Var(z)$.

Algoritmus bude vyzerať nasledovne

[3] Toto nám dá rovnaký výsledok ako keby sme použili známy Newton-Raphsonov algoritmus na počítanie optima.

(1) Nastavíme úvodné hodnoty pre $\hat{\eta}_0$ a $\hat{\mu}_0$
(2) Skonštruujeme $z_0 = \hat{\eta}_0 + (y - \hat{\mu}_0)\frac{d\eta}{d\mu}|_{\hat{\eta}_0}$
(3) Skonštruujeme váhy[4] $\frac{1}{w_0} = \left(\frac{d\eta}{d\mu}\right)^2 |_{\hat{\mu}_0} V(\hat{\mu}_0)$
(4) Regresiou z_0 na X odhadneme β a pomocou toho dostaneme $\hat{\eta}_1$ a z toho potom aj $\hat{\mu}_1$
(5) Opakujeme kroky (2)-(4) dokým parametre neskonvergujú.

Odhady variancie majú nasledovnú formu $\hat{Var}(\hat{\beta}) = (X^T W X)^{-1}\hat{\phi}$, kde W je matica, ktorá má na diagonále výsledné váhy.

V prípade binomických dát vyzerajú členy vstupujúce do algoritmu nasledovne:

$$\eta = \text{logit}(\mu), \qquad \frac{d\eta}{d\mu} = \frac{1}{\mu(1-\mu)}, \qquad V(\mu) = \mu(1-\mu)/n, \qquad w = n\mu(1-\mu),$$

ako úvodné hodnoty môžeme nastaviť $\hat{\mu}_0 = y$ a $\eta_0 = \text{logit}(\hat{\mu}_0)$. Toto môžeme urobiť ak y pre žiadne pozorovanie nie je nula.

Nakoľko je disperzný parameter $\phi_{bin} = 1$ pre binomický model a pre gaussovský je to $\phi_{lin} = \sigma^2$, musíme výsledné smerodajné odchýlky podeliť odhadom variancie reziduí z lineárneho modelu.

5.4 Iteratively Reweighted Least Squares

Teraz si ukážeme, že prečo je hore uvedený IRWLS algoritmus len to isté ako Fisher-scoring algoritmus na maximalizovanie likelihoodu.

Uvažujme len 1 porozovanie. Našou úlohou je odhadnúť β pomocou metódy maximálnej vierohodnosti. Uvažujme škálovací parameter ϕ fixný.

Parciálna derivácia log-likelihoodu jedného pozorovania je

$$\frac{\partial l_i}{\partial \beta_j} = \left(\frac{\partial l_i}{\partial \theta_i}\right)\left(\frac{\partial \theta_i}{\partial \mu_i}\right)\left(\frac{\partial \mu_i}{\partial \eta_i}\right)\left(\frac{\partial \eta_i}{\partial \beta_j}\right)$$

$$\frac{\partial l_i}{\partial \beta_j} = \left(\frac{y_i - b'(\theta_i)}{a(\phi)}\right)\left(\frac{1}{b''(\theta_i)}\right)\left(\frac{\partial \mu_i}{\partial \eta_i}\right)(x_{ij})$$

$$\frac{\partial l_i}{\partial \beta_j} = \left(\frac{y_i - \mu_i}{a(\phi)}\right)\left(\frac{a(\phi)}{Var(y_i)}\right)\left(\frac{\partial \mu_i}{\partial \eta_i}\right)(x_{ij}) = \frac{y_i - \mu_i}{Var(y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)x_{ij}$$

Označme $u_{ij}(\beta) = \frac{\partial l_i}{\partial \beta_j}(\beta)$, teda ide o skóre funkciu. Pre maximum likelihood estimátor poznáme vzťah medzi Fisherovou informačnou maticou $I(\beta)$ a priemernou hodnotou kvadrátu skóre funkcie.

$$I_i^{(k,l)} = -E\left(\frac{\partial^2 l_i}{\partial \beta_j \partial \beta_k}\right) = E\left(\frac{\partial l_i}{\partial \beta_l}\frac{\partial l_i}{\partial \beta_k}\right) = E\left(u_{ik}u_{il}\right)$$

V našom prípade preto máme $I_i^{(k,l)} = E\left(\frac{(y_i-\mu_i)^2}{(Var(y_i))^2}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 x_{ij}x_{ik}\right) = \frac{1}{Var(y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 x_{ij}x_{ik}$

Pre n pozorovaní máme

$$u = \frac{\partial l}{\partial \beta} = X^T A(y - \mu)$$

pre vhodnú voľbu matice A.

Podobne Fisherova informačná matica má tvar

$$I = -E\left(\frac{\partial^2 l}{\partial \beta \partial \beta^T}\right) = X^T W X$$

pre vhodnú voľbu matice W. Naviac platí $A = W\frac{\partial \eta}{\partial \mu}$.

[4] Aby sme sa vyhli nekonečným váham tak sa váhy regularizujú, teda $w_0 = \frac{1}{\max\{\delta, \left(\frac{d\eta}{d\mu}\right)|_{\hat{\mu}_0} V(\hat{\mu}_0)\}}$, kde δ je nejaké malé číslo.

Z Taylorovho rozvoja vieme, že

$$u(\hat{\beta}) \approx u(\beta_0) + \frac{\partial u(\beta_0)}{\partial \beta}(\hat{\beta} - \beta_0).$$

Označme maticu $H(\beta) = \frac{\partial u(\beta)}{\partial \beta}$.

V optime má platiť $u(\hat{\beta}) = 0$, preto nám ľavá strana rovnice dáva

$$\hat{\beta} = \beta_0 - H^{-1}(\beta_0)u(\beta_0).$$

Iteratívna metóda založená na využití tohto vzťahu sa nazýva Newtonov-Raphsonov algoritmus. Ak $-H(\beta_0)$ nahradíme jej strednou hodnotou, teda $I(\beta)$, dostávame Fisher-scoring algoritmus

$$\hat{\beta} = \beta_0 + I^{-1}(\beta_0)u(\beta_0).$$

V našom prípade teda bude mať iteratívna schéma nasledovný tvar

$$\beta^{(t+1)} = \beta^{(t)} + (X^T W X)^{-1} X^T A(y - \mu)$$

Preusporiadaním dostávame

$$\beta^{(t+1)} = (X^T W X)^{-1} \left(X^T W X \beta^{(t)} + X^T A(y - \mu) \right)$$

$$\beta^{(t+1)} = (X^T W X)^{-1} X^T W z,$$

kde $z = \eta + \left(\frac{\partial \eta}{\partial \mu}\right)(y - \mu)$ a $\eta = X\beta$. Takže maximalizovanie likelihoodu pri GLM pomocou metódy Fisher-scoring je to isté ako iteratívne aplikovanie metódy vážených štvorcov na z, teda na linearizovanú verziu odhadu odozvy.

5.5 Testovanie hypotéz

Vhodnosť fitu

Deviancia, teda log likelihood ratio štatistika na porovnanie skúmeného modelu voči saturovanému modelu, teda takému, čo dáta fituje úplne presne

$$2(\log L(y, \phi; y) - \log L(\hat{\mu}, \phi; y))$$

vyzerá pre nezávislé pozorovania nasledovne

$$\sum_i 2w_i(y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i))/\phi,$$

kde $\tilde{\theta}_i$ sú odhady v saturovanom modeli. Devianciu používame na to testovanie vhodnosti fitu.

Pearsonove X^2 vyzerá nasledovne

$$X^2 = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

a dá sa použiť na testovanie vhodnosti fitu podobne ako deviancia.

Obe štatistiky, aj deviancia aj Pearsonov X^2 sú asymptoticky rozdelené ako χ^2 s počtom stupňov voľnosti ako rozdielom v počtoch identifikovateľných parametrov (pre saturovaný model je to n, pre menší model je to p, takže máme $n - p$ stupňov voľnosti).

Toto je neužitočné pre Gaussovský lineárny model, lebo nepoznáme disperzný parameter narozdiel od binomického a Poissonovho modelu kde je $\phi = 1$.

Porovnávanie vnorených modelov

Rozdiel deviancií $D_{small} - D_{large}$ má asymptoticky rozdelenie ako χ^2 s počtom stupňov voľností $df_{large} - df_{small}$. Pre modely kde nepoznáme disperzný parameter ϕ, sa však dá odhadnúť pomocou $X^2/(n-p)$. Potom má štatistika

$$\frac{(D_{small} - D_{large})/(df_{small} - df_{large})}{\hat{\phi}}$$

približne F-rozdelenie. Pre Gaussovský lineárny model má táto štatistika presne F-rozdelenie.

5.6 Diagnostika

Dôležitou súčasťou každej dátovej analýzy je kritické zhodnotenie prepodkladov modelu. Toto môžeme robiť obrázkami alebo štatistickými testami.

Teraz nejaké názvoslovie:

Reziduá

- Response reziduá: $r = y - \hat{\mu}$
- Pearsonove reziduá: $r_P = \frac{y-\hat{\mu}}{\sqrt{V(\hat{\mu})}}$
- Deviance reziduá: $r_D = sign(y - \hat{\mu})\sqrt{d_i}$, kde $D = \sum_i d_i$

Vplyvné pozorovania

Diagonálny člen h_i matice H

$$H = W^{1/2}X(X^TWX)^{-1}X^TW^{1/2}$$

nám hovorí ako veľmi vie pozorovanie i *potenciálne* (napr. ak by sme pozorovali inú hodnotu y_i) ovplyvniť fit (tomuto sa hovorí **leverage**).

Okrem toho nás môže zaujímať ako priamo by ovplyvnilo vynechanie i-teho pozorovania fit (tomuto sa hovorí **influence**). Na toto slúži Cookova štatistika

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})^T(X^TWX)(\hat{\beta}_{(i)} - \hat{\beta})}{\hat{\phi}},$$

kde $\hat{\beta}_{(i)}$ je odhad parametra pri vynechanom i-tom pozorovaní. Jednoduché pravdilo je, že ak je $D_i > 1$ (iní používajú $D_i > 4/n$) dané pozorovanie je vplyvné. Nakoľko je Cookova štatistika asymptoticky rozdelená ako $F_{p,n-p}$, môžeme ako rozhodovacie pravidlo na určenie vplyvnosti daného pozorovania použiť $(1-\alpha)$ kvantil tohoto rozdelenia.

Diagnostika modelu

Reziduá

Pozeráme sa na reziduá voči fitovanej hodnote $\hat{\mu}$ alebo $\hat{\eta}$. Na týchto obrázkoch pozeráme či neuvidíme nejakú nelineárnu závislosť. Ak ju vidíme tak čo s tým? Môžeme

- pretransformovať závislú premennú (pozor, lebo tým zmeníme distribúciu Y)
- porozmýšľame, či nezmeniť linkovú funkciu (ale väčšinou tie kanonické linky sú najrozumnejšie)
- pridať alebo pretransformovať prediktory (asi najlepšia rada)

Teraz sa pozrieme na varianciu reziduí. Záleží aj na aké reziduá sa pozeráme: deviance reziduá sú normalizované varianciou ale response reziduá nie sú, takže napríklad pri Poissonovom GLM budeme vidieť rastúcu varianciu s $\hat{\eta}$ pri response reziduách ale pri deviance reziduách nie. Čo s tým? Napríklad môžeme zmeniť variančnú funkciu $V(\mu)$ ale potom už môžeme byť mimo GLM špecifikácie.

Pre niektoré modely môžu byť diagnostické obrázky s reziduami neužitočné, napríklad pri binárnom modeli s individuálnymi dátami ($n_i = 1$) môžu reziduá nadobúdať len dve hodnoty.

Špecifikácia vzťahu medzi premennými

Môžeme sa pozrieť na grafu vzťahu medzi závislá premenná voči prediktorom a z toho niekedy vidno, že prediktor treba pretransformovať. Podobne sa môžeme pozrieť na vzťah prediktora a lineárnej aproximácie vysvetľovanej premennej ($z = \eta + (y - \mu)\frac{d\eta}{d\mu}$, môžeme použiť $\hat{\eta}$ a $\hat{\mu}$ z nášho odhadnutého modelu). Tieto obrázky však neberú do úvahy vplyv ostatných prediktorov. Na to je lepšie sa pozerať **parciálne reziduá**: teda zobrazíme vzťah $z - (\hat{\mu} - \beta_j x_j)$ voči x_j.

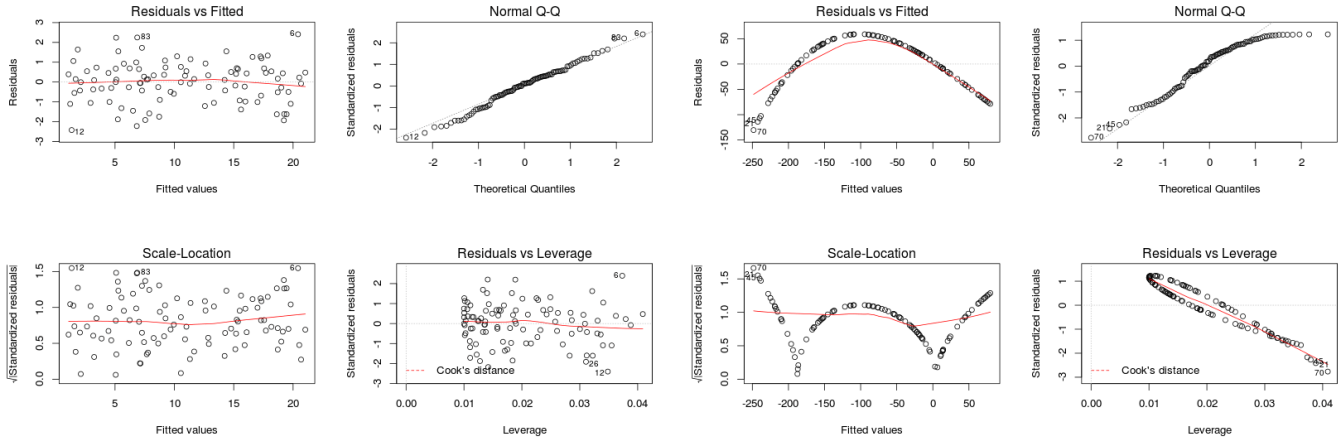
Nezvyčajné porozovania

Užitočný grafický nástroj na detekciu nezvyčajných pozorovaní je Q-Q plot. Ide o vzťah usporiadaných

- absolútnych hodnôt reziduí
- vplyvnosti pozorovaní - leverage
- Cookových štatistík
- alebo hocijakých iných kvantít, ktoré hovoria čosi o vplyvnosti na fit modelu

voči kvantilom normálneho rozdelenia $\Phi^{-1}\left(\frac{n+i}{2n+1}\right)$. Reziduá nemusia byť normálne pri všeobecnom GLM, preto sa pozeráme na body mimo trendu.

Viac o diagnostike veľmi prístupnou formou tu [glmb] a tu [glma].




```r
# correct model
x <- runif(100, 0, 10)
y <- 1 + 2 * x + rnorm(100, 0, 1)
m <- lm(y ~ x)
par(mfrow = c(2, 2))
plot(m)
```

```r
# some wrong model
y <- 1 + 2 * x + 1 * x^2 - 0.5 * x^3
m <- lm(y ~ x)
par(mfrow = c(2, 2))
plot(m)
```

Obr. 2: Diagnostika: vľavo korektne špecifikovaný model, vpravo nie, zdroj: [glmb].

5.7 Ďalšie GLM - Gamma regresia a Inverzná Gaussovská regresia

Sú vhodné na modelovanie spojitej náhodnej premennej s nahnutím (*skewness*). Niekedy totiž chceme modelovať aj strednú hodnotu aj varianciu naraz. Ďalej kvázi-GLM model bude vhodný v situácii, keď si nie sme istý distribúciou Y ale (z povahy problému) máme dosť dobrú predstavu o linkovej a variančnej funkcii.

Gamma regresia a Inverzná Gaussovská regresia

Uvažujme nasledovnú parametrizáciu Gamma rozdelenia pre náhodnú premennú Y

$$f(y) = \frac{1}{\Gamma(\nu)} \left(\frac{\nu}{\mu}\right)^{\nu} y^{\nu-1} \exp\left(-\frac{y\nu}{\mu}\right),$$

kde $E(Y) = \mu$ a $Var(Y) = \frac{\mu^2}{\nu} = \phi\mu^2$. Gamma rozdelenie je rozdelenie sumy ν nezávislých exponenciálne rozdelených náhodných premenných s parametrom $\lambda = \nu/\mu$, špeciálnym prípadom je aj χ^2, kde $\lambda = 1/2$ a $\nu = df/2$.

Kanonický parameter $\theta = -\frac{1}{\mu}$ a kanonická link funkcia je $\eta = -\frac{1}{\mu}$, mínuska sa zbavíme aby bola notácia jednoduchšia (ale dobre si to zapamätáme). Variančná funkcia je $b''(\theta) = \mu^2$ (lebo $b(\theta) = \log(1/\mu) = -\log(-\theta)$).

Ak máme rozumné dôvody sa domnievať, že Y má skutočne Gamma rozdelenie, tak potom je Gamma regresia samozrejme veľmi vhodný model. Avšak aj ak nie, a domnievame sa, že $Var(Y) \propto (E(Y))^2$, tak to môže byť rozumná voľba. Ak $Var(Y) \propto (E(Y))^2$, tak môžeme zvážiť aj Gaussovský model s log transformovanou vysvetľovanou premennou, toto však nemusí byť žiadúce, nakoľko transformovaním môžeme dostať premennú, ktorá je ťažšie interpretovateľná, kdežto pri Gamma regresii modelujeme Y priamo!

Používajú sa tri linkové funkcie

- $\eta = \frac{1}{\mu}$, tu však nemáme garantované $\mu > 0$ a preto sa môže urobiť nasledovná modifikácia $E(Y) = \mu = \frac{\alpha_0 x}{1+\alpha_1 x} = \frac{1}{\alpha_1/\alpha_0 + 1/(\alpha_0 x)} = \frac{1}{\eta}$, ktorá zabezpečí, že stredná hodnota μ je ohraničená. V tomto prípade sme predefinovtali $\eta = \alpha_1/\alpha_0 + 1/(\alpha_0 x)$.
- $\eta = \log \mu$, používame ak sa domnievame, že prediktory majú na vysvetľovanú premennú multiplikatívny efekt.
- $\eta = \mu$, je vhodný na modelovanie súčtu štvorcov, ktoré majú rozdelenie χ^2

Inverzná Gaussovská regresia sa oproti Gamma regresii líši v tom, že variančná funkcia rastie rýchlejšie (ako funkcia strednej hodnoty).

5.7.1 Spoločné modelovanie strednej hodnoty a variancie

Kým pri GLM sme mali disperzný parameter ϕ_i fixný pre každé pozorovanie (v prípade Poissonovej alebo Binomickej regresie platilo dokonca $\forall i : \phi_i = 1$), teraz budeme považovať za náhodnú premennú a modelovať ju cez (napríklad) Gamma GLM.

Máme štandardné GLM pre strednú hodnotu

$$E(Y_i) = \mu_i, \quad \eta_i = g(\mu) = x_i^T\beta, \quad Var(Y_i) = \phi_i V(\mu_i), \quad w_i = 1/\phi_i$$

a Gamma GLM, ktorý použijeme na nejaký odhad disperzie d_i (štvorce deviance reziduálov napr.)

$$E(d_i) = \phi_i, \quad \zeta_i = \log(\phi_i) = z_i^T\gamma, \quad Var(d_i) = \tau\phi_i^2.$$

Prediktory z_i vysvetľujúce strednú hodnotu ϕ_i náhodnej variancie d_i môžu byť napríklad podmnožinou prediktorov v x_i.

Prirodzene ide o bohatší model nakoľko uvažujeme štruktúru aj vo variancii.

5.7.2 Kvázi GLM

Môžeme byť v situácii, že máme dosť dobrú predstavu o linkovej a variančnej funkcii ale netušíme, aké by bolo vhodné pravdepodobnostné rozdelenie pre Y. Pre výpočet maximum likelihood odhadu pre GLM nám stačí linková a variančná funkcia. Ak by náhodou Y patrilo do rodiny exponenciálnych rozdelení, potom odhad, ktorý dostaneme pomocou nich bude maximum likelihood. Takže či patrí alebo nepatrí, odhadovať parametre môžeme nezávisle od toho. Čo však so štatistickou inferenciou? Ak Y nie je z rodiny exponenciálnych rozdelení, tak na to, aby sme dostali štandardné chyby estimátorov (a tým pádom aj konfidenčné intervaly), potrebujeme likelihood. Ale ten nemáme, takže sa ideme uspokojiť aspoň s čímsi podobným ako likelihood.

Majme nezávislé Y_i so strednou hodnotou μ_i a varianciou $\phi V(\mu_i)$. Pre túto náhodnú premennú $U_i = \frac{Y_i - \mu_i}{\phi V(\mu_i)}$ platia nasledovné tri vlastnosti:

$$E(U_i) = 0, \quad Var(U_i) = \frac{1}{\phi V(\phi)}, \quad -E\left(\frac{\partial U_i}{\partial \mu_i}\right) = -E\left(\frac{-\pi V(\mu_i) - (Y_i - \mu_i)\phi V(\mu_i)}{[\phi V(\mu_i)]^2}\right) = \frac{1}{\phi V(\phi_i)}.$$

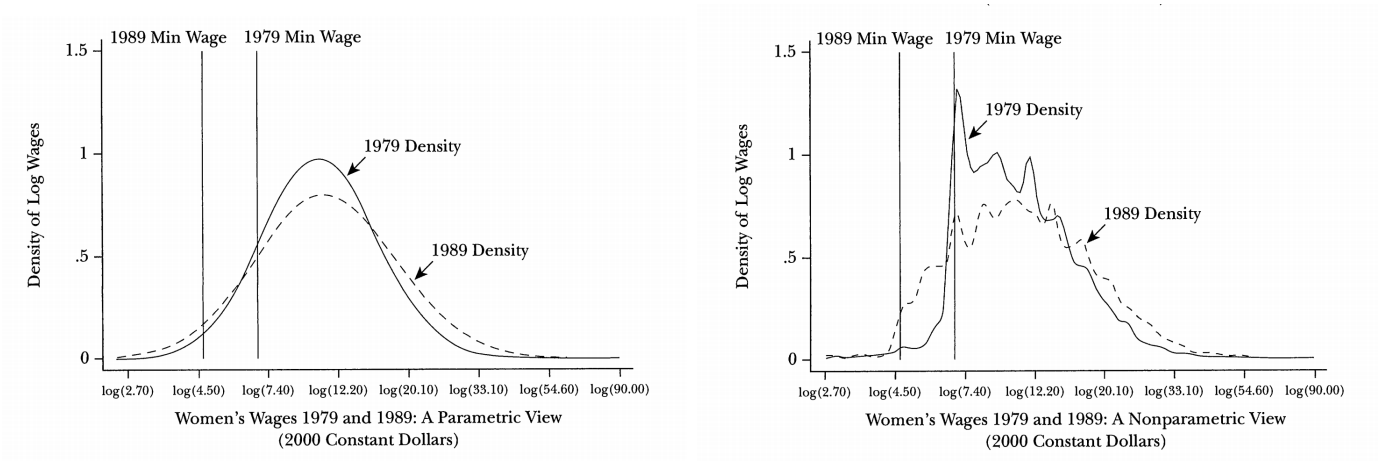
Tieto tri vlastnosti však zdieľa aj prvá derivácia log-likelihoodu, preto sa budeme tváriť, že U_i je čosi podobné. Zadefinujeme *kvázi-likelihood* ako

$$Q = \sum_i Q_i = \sum_i \int_{y_i}^{\mu_i} \frac{y_i - t}{\phi V(t)} dt.$$

Pracovať s ním budeme ako s normálnymi likelihoodom, len disperzný parameter odhadneme ako $\hat{\phi} = \frac{X^2}{n-p}$. Poznamenajme, že kvázi-likelihood záleží len na variančnej funkcii. Kvázi-deviancia je $-2\phi Q$ a porovnávať modely môžeme pomocou rozdielu kvázi-deviancií.

6 Neparametrická regresia

Parametrický model je taký, ktorý má konečne veľa parametrov. **Neparametrický model** má nekonečne veľa parametrov.



Obr. 3: Parametrický vs. neparametrický odhad hustoty rozdelenia miezd [DT01]

Parametrický model

- jednoduchý
- ľahšie interpretovateľný
- ak korektne špecifikovaný, tak efektívnejší
- na predikciu si nám stačí pamätať menej informácií
- vhodný ak vieme čosi viac o skúmanom probléme

Neparametrický model

- flexibilný
- výsledky sú ťažšie interpretovateľné
- robustnejší voči zlej špecifikácii modelu
- vhodný ak o skúmanom probléme vieme len málo

Ak má model aj parametrickú aj neparametrickú časť, volá sa **semiparametrický**.
Začnime s jednorozmerným prípadom. Majme nasledujúci model

$$Y_i = f(X_i) + \epsilon_i, \quad \{\epsilon_i\}_{i=1}^n \text{ i.i.d.}, \quad E(\epsilon_i) = 0, \quad Var(\epsilon_i) = \sigma^2.$$

Našou úlohou je odhadnúť funkciu f. Parametrický model napríklad uvažuje $f(X_i) = \beta_0 + \beta_1 X_i$. Vtedy je model jednoznačne popísaný dvomi číslami: (β_0, β_1). Neparametrický model uvažuje, že f patrí do akejsi rodiny funkcií, najčastejšie (dostatočne) hladkých funkcií.

Na neznámu funkciu f sa môžeme pozerať aj nasledovne

$$f(x_i) = E(Y_i | X_i = x_i) = \frac{\int y_i g(x_i, y_i) dy_i}{\int g(x_i, y_i) dy_i},$$

kde $g(y_i, x_i)$ je združená funkcia hustoty vektora (Y_i, X_i).

6.1 Jadrové odhady (Kernel Estimators)

Najjednoduchším nápadom ako vyhladiť oblak bodov je zobrať akýsi priemer pozorovaní. Chceme odhadnúť funkciu f v nejakom konkrétnom bode x. Pozrieme sa do nejakého okolia tohoto bodu (**bandwidth**) a preváhujeme všetky pozorovania v tomto okolí. Porozovaniam, ktoré sú blízko x dáme väčšiu váhu ako pozorovaniam, ktoré sú ďaleko. Tie váhy urobíme pomocou funkcie, ktorá sa volá jadrová funkcia **kernel**. Toto je funkcia pre ktorú platí $\int K(x)dx = 1$, $\int xK(x)dx = 0$ a $\int x^2 K(x)dx < \infty$.

Označme $\hat{f}_\lambda(x)$ odhadcu neznámej funkcie f, λ označuje bandwith a K označuje kernel.

$$\hat{f}_\lambda(x) = \frac{1}{n\lambda} \sum_{i=1}^n K\left(\frac{x - x_i}{\lambda}\right) Y_i = \frac{1}{n} \sum_{i=1}^n w_i Y_i,$$

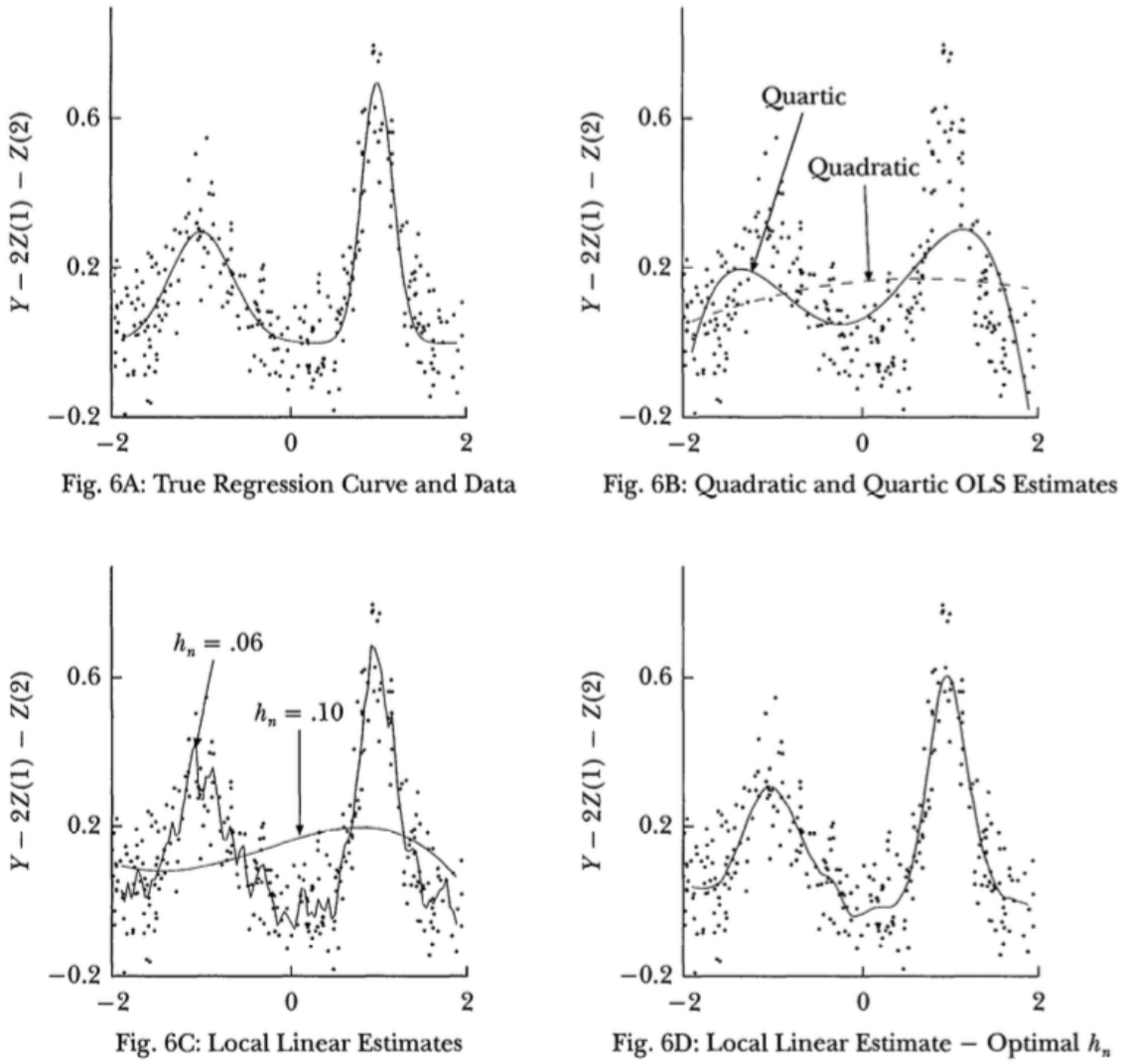


Fig. 6A: True Regression Curve and Data

Fig. 6B: Quadratic and Quartic OLS Estimates

Fig. 6C: Local Linear Estimates

Fig. 6D: Local Linear Estimate − Optimal h_n

Obr. 4: Parametrická vs Neparametrická regresia [DT01].

je veľa rôznych kernelov, aj celá teória o tom ako zvoliť dobrý kernel. Aké je však kritérium podľa ktorého budeme posudzovať daný kernel a bandwith ako dobrý alebo nie? Dobrým kandidátom je stredná štvorcová chyba ďalej označovaná ako MSE = mean squared error.

$$\begin{aligned} MSE(x) = \; & E\left[(f(x) - \hat{f}_\lambda(x))^2\right] = E\left[(\hat{f}_\lambda(x) - E(\hat{f}_\lambda(x))^2\right] + (E(\hat{f}_\lambda(x)) - f(x))^2 \\ = \; & Var(\hat{f}_\lambda(x)) + Bias(\hat{f}_\lambda(x), f(x))^2. \end{aligned}$$

MSE teda berie do úvahy aj bias aj varianciu.

Motivačný príklad: pre dané x vezmime len obyčajný priemer pre všetky x_i , ktoré sú dostatočne blízko.

$$\hat{f}_\lambda(x) = \frac{\sum_{i=1}^n I(|x_i - x| \leq \lambda).y_i}{\sum_{j=1}^n I(|x_j - x| \leq \lambda)} = \frac{\sum_{i=1}^n K_u\left(\frac{x_i - x}{h}\right).y_i}{\sum_{j=1}^n K_u\left(\frac{x_j - x}{h}\right)},$$

kde $K_u(u) = \frac{1}{2} I(|u| \leq 1)$ je uniformný kernel.

Estimátor tohoto typu (pre všeobecný kernel K)

$$\hat{f}_\lambda(x) = \frac{\sum_{i=1}^n K\left(\frac{x_i - x}{\lambda}\right).y_i}{\sum_{j=1}^n K\left(\frac{x_j - x}{\lambda}\right)} = \frac{\sum_{i=1}^n w_i y_i}{\sum_{j=1}^n w_j}.$$

sa nazýva **Nadaraya-Watson** estimátor.

Kernel	Formula
Uniformný	$K_U(u) = \frac{1}{2} I(\|u\| \leq 1)$
Epanechnikov	$K_E(u) = \frac{3}{4}(1-u^2) I(\|u\| \leq 1)$
Gaussovský	$K_G(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$

Nejaké príklady kernelov.

Pre širokú triedu funkcií je Epanechnikov kernel optimálny v zmysle minimalizácie asymptotickej integrovanej MSE (teda $AMISE = \int_{-\infty}^{\infty} AMSE(\hat{f}_\lambda(x))dx$).

V praxi je však dôležitejší výber bandwithu ako kernelu. Optimálny bandwith záleží od funkcie f, ktorú mi nepoznáme. Dá sa aspoň odvodiť ako rýchlo má klesať optimálny bandwidth (AMISE minimalizujúci) spolu s rastúcou vzorkou. Existuje však aj jednoduchšie riešenie, a to pomocou krosvalidácie.

Môžeme vybrať bandwidth λ, ktorý minimalizuje

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{f}_{\lambda,j}(x_j))^2,$$

index j znamená, že j-te pozorovanie je vynechané. Toto je automatická selekcia λ a musí byť používaná opatrne.

Na kraji aproximovaného územia máme problém, lebo máme len polovicu bodov, z ktorých robíme priemer, preto tam je problém veľkého biasu. Existuje teória, ktorá hovorí akým spôsobom tento bias čiastočne korigovať [Han09].

Jadrový odhad je vlastne výsledkom nasledovnej minimalizácie

$$\min_\theta \sum_i K\left(\frac{x_i - x}{h}\right)(y_i - \theta)^2$$

teda fitujeme konštantu na kernelom váhované dáta.

6.2 Splajny

Splajn (rádu k) je po čiastkach polynomiálna funkcia, ktorá je spojitá a má spojité všetky derivácie až do rádu $k-1$. Body, medzi ktorými sa mení polynomiálna funkcia sa nazývajú **uzly.**

Aká je naša motivácia aby funkcia bola spojitá (aj spojite diferencovateľná) všade, teda aj v uzloch? Znižujeme tým varianciu.

Zaujímavosťou je, že rýchlosť klesania MSE je rovnakého rádu ako pri jadrovom odhade.

Majme uzly fixované. Každý splajn rádu k sa dá vyjadriť ako lineárna kombinácia **bázických funkcií.** Ako tieto vyzerajú? Majme uzly $t_1 < t_2 < ... < t_m$, bázických funkcii je $m + k + 1$, a sú to $1, x, x^2, ...$ okrem toho $(x - t_j)_+^k$. (Ak by sme tam nemali kladnú časť, tak sú tam zbytočne.) Tieto funkcie spravia uzly (nod producing function), práve oni tam dodajú to "lokálno". Táto sada bázických funkcií sa nazýva **truncation power basis** a teraz vieme problém nájdenia optimálneho MSE minimalizujúceho splajnu formulovať ako úlohu lineárnej regresie. Nech $G_{ij} = g_j(x_i)$, kde g_j je j-ta bázická funkcia. Nakoľko naša funkcia $f(x) = \sum_{j=1}^{m+k+1} \beta_j g_j(x)$ a teda koeficienty β nájdeme ako $\arg\min_\beta |y - G\beta|^2$ a preto ich voláme aj **regresné splajny**.

Mimochodom, truncation power basis sa príliš nepoužívajú v praxi, kvôli tomu, že sú numericky nestabilné. Na počítanie výhodnejšie bázické splajny, ktoré pokrývajú ten istý priestor sa nazývajú **B-splajny**.

Pre kubický splajn naše oko nevie rozlíšiť kde sú uzly (čo je fajn vlastnosť).

Na kraji aproximovaného územia máme opačný problém ako pri jadrových odhadoch, máme tu priveľkú varianciu. Riešenie je používať **prirodzené splajny** (*natural splines*), teda polynómy rádu $(k-1)/2$. Pre prirodzené splajny máme len m bázických funkcií, teda to vôbec nezávisí od stupňa polynómu! Tento zdanlivo protiintuitívny fakt sa dá vysvetliť nasledovne. Máme $m-1$ "vnútorných"regiónov, kde sú popísané $k+1$ parametrami. Na kraji máme $2((k-1)/2+1)$ parametrov. Dokopy máme mk reštrikcií, teda [počet voľných parametrov] - [počet reštrikcií] $= (m-1)(k+1) + 2((k-1)/2+1) - mk = m$.

Ako vybrať uzly? Na to existuje veľa spôsobov. Použijeme na to **regularizačnú** metódu. Uvažujme nasledovnú úlohu

$$\min_f \sum_i (y_i - f(x_i))^2 + \lambda \int (f''(x))^2 dx,$$

kde minimalizujeme cez priestor funkcií. Riešenie sa nazýva **vyhladzovací splajn** (*smoothing spline*). Člen s λ zabezpečuje, aby sme nemali príliš veľkú varianciu. Pre $\lambda = 0$ sme v situácii, že splajn úplne fituje dáta (nízky bias, vysoká variancia). Z teórie vieme, že minimizátor je jednoznačný a je to prirodzený kubický splajn s uzlami v bodoch $x_1, ..., x_n$. To znamená, že ide o po čiastkach kubickú funkciu, ktorá je spojitá a má spojité prvé dve derivácie.

Môžeme uvažovať aj robustnú verziu vyhladzovacieho splajnu, kde sumu štvorcov nahradíme inou funkciou, takou ktorá nepenalizuje vzdialenejšie body tak prísne

$$\min_f \sum_i \rho(y_i - f(x_i)) + \lambda \int (f''(x))^2 dx,$$

kde napríklad $\rho(x) = |x|$.

6.3 Lokálne polynómy

Jadrové vyhladzovače aj splajny sú zraniteľné voči outlierom. Tieto môžeme vyhodiť alebo zväčšiť vyhladzovanie. Lokálne polynómy majú dve dobré dôležité vlastnosti: sú robustné ako lineárna regresia a poskytujú lokálny fit podobne ako jadrové odhady.

Fungujú nasledovne: zvolíme si určité dátové okno. Pre dáta v tomto okne fitujeme polynóm robustnými metódami. Za odhad považujeme polynómom predikovanú hodnotu v strede okna. Populárna implmentácia tejto metódy sa nazýva **loess**. V praxi treba zvoliť stupeň polynómu a veľkosť okna.

V podstate ide o prirodzené rozšírenie jadrového odhadu na regresiu

$$\min_{(\beta_0,\beta_1)} \sum_i K\left(\frac{x_i - x}{h}\right)(y_i - \beta_0 - \beta_1 x_i)^2$$

teda namiesto konštanty fitujeme lineárnu funkciu na kernelom váhované dáta (ak každému pozorovaniu v okne dáme rovnakú váhu, tak tomu zodpovedá uniformný kernel). Toto sa zvykne tiež volať **lokálna lineárna regresia**. Tu je matematická formulácia problému.

$$B = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdots & \\ 1 & x_n \end{bmatrix}, \quad \Omega = \text{diag}\left(K\left(\frac{x_1 - x}{h}\right), \cdots, K\left(\frac{x_n - x}{h}\right)\right), \quad \min_{\beta=(\beta_0,\beta_1)} ||\Omega^{1/2}(y - B\beta)||_2^2,$$

$$\hat{\beta} = (B^T \Omega B)^{-1} B^T \Omega y, \quad \hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x.$$

6.4 Wavelety

Majme fixovaný ekvidištančný grid veľkosti 2^b.

Ideme skonštruovať bázické funkcie $\phi_1, ..., \phi_n$, ktoré budú ortogonálne ($W^T W = I$) a zároveň budú mať kompaktný suport, takže budú vhodné na modelovanie "lokálnosti". $W_{ij} = \phi_i(x_j)$.

- Hard thresholding

$$\min_\theta ||y - W\theta|| + \lambda ||\theta||_0$$

$$\text{threshold}_{hard}(\theta_i) = \begin{cases} \theta_i, & \text{if } \theta_i > \lambda \\ 0 & \text{if } \theta_i \in [-\lambda, \lambda] \\ \theta_i, & \text{if } \theta_i < \lambda \end{cases}$$

- Soft thresholding

$$\min_{\theta} ||y - W\theta|| + \lambda||\theta||_1$$

$$\text{threshold}_{soft}(\theta_i) = \begin{cases} \theta_i - \lambda, & \text{if } \theta_i > \lambda \\ 0 & \text{if } \theta_i \in [-\lambda, \lambda] \\ \theta_i + \lambda, & \text{if } \theta_i < \lambda \end{cases}$$

a je to isté ako

$$\theta = W^T y$$
$$\hat{\theta} = \text{threshold}(\theta)$$
$$\hat{y} = W\hat{\theta}$$

Počítanie je extrémne rýchle.

Sú rôzne wavelet bázické funkcie. Napríklad *Haar* wavelet. Na intervale $[0, 1)$ definujeme *mother wavelet* ako

$$w(x) = \begin{cases} 1, & \text{if } x \leq 1/2 \\ -1 & \text{if } x > 1/2 \end{cases},$$

a ostatné wavelety sú len posunuté a natiahnuté verzie mother waveletu $h_n(x) = 2^{j/2} w(2^j - k)$, pre $n = 2^j + k$ a $0 \leq k \leq 2^j$.

6.5 Viacrozmerné prediktory

Vizualizácia je náročná v tomto prípade.

Nadaraya-Watson estimátor vo viacrozmernom prípade vyzerá

$$\hat{f}_\lambda(x) = \frac{\sum_{i=1}^n K\left(\frac{X_i - X}{\lambda}\right).Y_i}{\sum_{j=1}^n K\left(\frac{X_j - X}{\lambda}\right)}$$

a kernel zväčša býva sféricky symmetrická funkcia.

Je tu veľký problém s presnosťou fitu pri vyšších dimenziách. Pozrime sa napríklad na MSE pre lokálne polynómy. $MSE = E(\hat{f}(x) - f(x))^2 = O\left(\frac{1}{n}^{\frac{2k}{2k+d}}\right)$, takže s rastúcou dimenziou d sa rýchlosť poklesu prudko zmenší. Na to aby sme mali MSE menšie ako nejaké ϵ potrebujeme $n \geq \left(\frac{1}{\epsilon}\right)^{\frac{2k+d}{2k}}$, teda exponenciálne rýchlo. (Na druhej strane čím hladšia funkcia, tým bližšie sme pri $O(1/n)$ teda pri rýchlosti parametrického modelu). Pre porovnanie: ak je skutočná f lineárna, a $\hat{f}$ je odhad lineárnej regresie potom $MSE = O\left(\frac{d}{n}\right)$. Teda tiež závisí negatívne od d ale oveľa menej tragicky ako pri neparametrických metódach.

6.6 Výber modelu? Receptár

- Málo šumu, tak je vhodné použiť interpoláciu.
- Stredne veľa šumu, neparametrické metódy sú vhodné.
- Veľa šumu, parametrické metódy sú vhodné na zrekonštuovanie základných vzťahov. Koniecconcov, ak je veľa šumu, tak dáta nám neprezradia viacej ako jednoduchý model.
- Treba dávať pozor na automatické vyhladzovanie, vždy je lepšie sa na to pozrieť. Ak to nie je možné, treba používať robustnejšie metódy.
- Podľa autora [Far05], je loess vyhladzovač dobrý všestranný prvý nástrel.

Výborné video prednášky môžete nájsť tu [not].

7 Bootstrap

Nasledujúca pasáž o výberovom rozdelení je inšpirovaná [Ken01], kde autor zdôrazňuje, prečo je dôležité aby študenti poriadne pochopili koncept výberovej štatistiky a ako nám k tomu môže dopomôcť bootstrap.

7.1 Výberové rozdelenie - Sampling distribution

Predstavme si situáciu, že si vyberáme z viacerých estimátorov parametra β, napríklad $\hat{\beta}^*, \hat{\beta}^{**}$.

Čo znamená vybrať si estimátor $\hat{\beta}^*$? Znamená to zavrieť si oči a vybrať si náhodne jeden prvok z výberového rozdelenia (*sampling distribution*).

Výber medzi $\hat{\beta}^*$ a $\hat{\beta}^{**}$ zodpovedá rozhodnutie, či si chceme náhodne vybrať estimátor z jedného alebo z druhého výberového rozdelenia. Tieto estimátory majú rôzne kvality, jeden má napríklad malý bias ale na druhej strane veľkú varianciu.

Vlastnosti výberového rozdelenia závisia od procesu, ktorý generuje dáta, preto neexistujú univerzálne najlepšie estimátory, ktoré fungujú na všetky prípady.

Testovacie štatistiky majú tiež výberové rozdelenie, ak je nulová hypotéza pravdivá potom vieme ako toto výberové rozdelenie vyzerá (napr. t-rozdelenie, χ^2).

7.2 Náš Problém

Uvažujme, že sme v jednej z nasledovných situácií.

- poznáme asymptotické rozdelenie estimátora ale máme len dátovú vzorku veľkosti napr. $n = 15$, štatistické testy založené na takejto aproximácii budú nedôveryhodné. Príklad: lineárna regresia s malým počtom pozorovaní.
- náš estimátor je výsledkom komplexného procesu a nevieme odvodiť ani len asymptotické rozdelenie estimátora.
- rozdelenie nášho estimátora je založené na predpokladoch, ktoré sú spochybniteľné. Príklad: výnosy akcií majú častokrát rozdelenie, ktoré nie je dobre aproximovateľné normálnym rozdelením, kvôli príliš ťažkým chvostom.
- asymptotická distribúcia estimátora závisí od skutočného a nám neznámeho dáta generujúceho procesu. Príklad: $X_1, X_2, ..., X_n \sim f(.)$ je výberový medián $\hat{m}$ asymptoticky rozdelený ako $N\left(m, \frac{1}{4nf(m)^2}\right)$.

7.3 Dáta sú všetko čo máme

Budeme považovať distribúciu z našich dát za skutočnú distribúciu a pomocou nej simulovať datasety. Tieto umelo vytvorené datasety budú mať rovnakú veľkosť ako náš pôvodný dataset a budeme ich vytvárať výberom s opakovaním.

Hádzanie kockou

Hádžeme kockou a chceli by sme vedieť akú priemernú hodnotu môžeme očakávať. V 25 nezávislých hodnoch pozorujeme nasledovné proporcie: $\left(\frac{2}{25}, \frac{1}{25}, \frac{8}{25}, \frac{7}{25}, \frac{3}{25}, \frac{4}{25}\right)$. Pýtame sa: padá na tejto kocke v priemere hodnota 3.5? V našich dátach je to 3.8.

Test hypotézy v lineárnom regresnom modeli

Predstavte si, že sme v nasledovnej situácii. Máme 25 pozorovaní a veríme, že náš model je korektný $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$. O chybách nepredpokladáme, že sú normálne. Ďalej uvažujme, že minimalizovaním štvorcov reziduí dostávame $(\hat{\beta}_0, \hat{\beta}_1) = (0.05, 2.25)$ a $\hat{\epsilon}$ nech označuje odhad reziduí. Uvažujme nasledujúci počítačový program.

(1) Vyberieme náhodne 25 hodnôt z $\hat{\epsilon}$ s opakovaním. (To znamená, že vyberieme jedno $\hat{\epsilon}_i$, pozrieme sa naň a vrátime ho späť.) Tento vektor 25 hodnôt označíme ako e.
(2) Vypočítame 25 hodnôt Y^* ako $0.5 + 2X + \left(\frac{25}{24}\right) e$
(3) Urobíme regresiu, kde Y^* vysvetľujeme pomocou X a dostaneme odhad $\hat{\beta}_1$ a jeho smerodajnú odchýlku se.
(4) Vypočítame $t = \frac{\hat{\beta}_1 - 2}{se}$ a uložíme ho.
(5) Zopakujeme body (1)-(4) veľakrát napr. 10000 a získame 10000 hodnôt t.
(6) Usporiadame t hodnoty od najväčšej po najmenšiu a vypíšeme 250-tu hodnotu a 9750-tu hodnotu.

Týmto spôsobom otestujeme, či je hodnota 2.5 štatisticky významne inšia ako hodnota 2, ktorú používame pri simulácii.

Notácia a teória

- $\{X_i, i = 1, \cdots, n\}$ dáta pochádzajúce z nám neznámej $F_0 \in \mathcal{I}$
- Niekedy uvažujeme akúsi parametrizáciu: $F_0(x, \theta_0) = P(X \leq x)$
- Testovacia štatistika $\hat{T}_n = T_n(X_1, ..., X_n)$
- $G_n(\tau, F_0) = P(\hat{T}_n \leq \tau)$ označuje skutočnú CDF testovacej štatistiky $\hat{T}_n$
- $\hat{T}_n$ je *pivotálna* ak $G_n(\tau, F)$ nezávisí na F
- $\hat{T}_n$ is *asymptoticky pivotálna* ak $G_\infty(\tau, F)$ nezávisí od F
- ako môžeme odhadnúť $G_n(., F_0)$?
  - napríklad pomocou G_∞ - teda asymptotickou aproximáciou (potrebujeme veľké n)
  - nahradením F_0 nejakým estimátorom - **bootstrap**
- nech $\hat{F}_n$ označuje estimátor neznámej F_0
  - ECDF (empirical cumulative distribution function) - $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n I(X_i \leq x)) \rightarrow_{a.s.} F_0(x)$
  - from a parametric family: $F_0(.) = F(., \theta_0)$

Procedúra na aproximáciu $G_n(\tau, F_0)$

Krok 1 Vygenerujeme náhodnú vzorku veľkosti n z $\hat{F}_n$: $\{X_i^* : i = 1, ..., n\}$
Krok 2 Vypočítame $\hat{T}_n^* = T_n(X_1^*, ..., X_n^*)$
Krok 3 Zopakujeme (1) a (2) veľakrát na to aby sme dostali empirickú distribúciu $(\hat{T}_n^* \leq \tau)$

Poznámka, zvyšovaním počtu simulovaných bootstrapových datasetov B zlepšíme odhad neznámej $G_n(\tau, \hat{F}_n)$. Teda simulovaním dostávame len $\hat{G}_n(\tau, \hat{F}_n)$, ak však máme dostatok trpezlivosti a výpočtovej sily, vieme urobiť tento odhad ľubovoľne dobrý, teda $\hat{G}_n(\tau, \hat{F}_n) \rightarrow G_n(\tau, \hat{F}_n)$ pre $B \rightarrow \infty$.

Čo to znamená, že bootstrap "funguje"? Znamená to, že $G_n(., \hat{F}_n) \rightarrow G_n(., F_0)$ Prinajhoršom by sme očakávali, že aproximácia bude správne keď veľkosť dátovej vzorky porastie do nekonečna. Táto vlastnosť sa nazýva konzistencia.

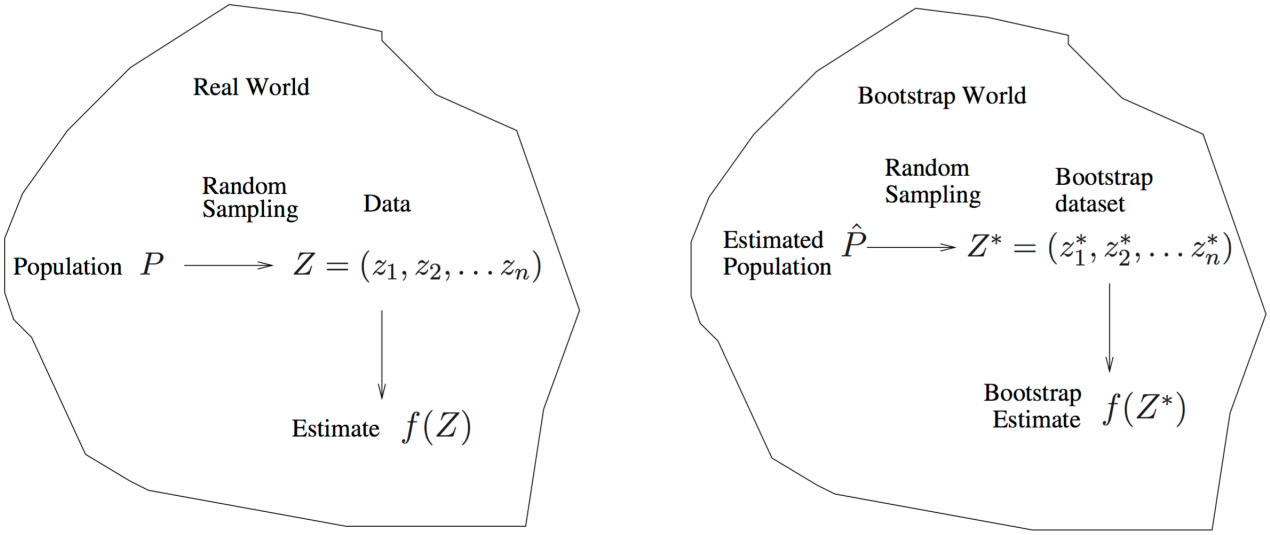
$G_n(t, \hat{F}_n)$ je konzistentný ak $\forall \epsilon > 0, \forall F_0 \in \mathcal{I}$

$$\lim_{n \to \infty} P_n \left[\sup_\tau |G_n(t, F_n) - G_\infty(\tau, F_0)| > \epsilon \right] = 0$$

$$G_n(\tau, \hat{F}_n) \sim G_\infty(\tau, F_n) \sim G_\infty(\tau, F_0) \sim G_n(\tau, F_0)$$

V článku (Beran and Ducharme 1991, [BD91]) sú uvedené nasledujúce postačujúce podmienky na garantovanie konzistencie bootstrapu.

- $\hat{F}_n \rightarrow F_0$ ($\hat{F}_n$ je dobrým odhadom F_0)
- $G_\infty(\tau, F)$ je spojitou funkciou τ pre všetky $F \in \mathcal{I}$ (spojitosť v τ)
- pre každé τ a pre každú postupnosť H_n, takú že $H_n \rightarrow F_0$: $G_n(\tau, H_n) \rightarrow G_\infty(\tau, F_0)$ ("spojitosť" v F_0)



Obr. 5: Bootstrap

7.4 Bias correction

Teraz nás zaujíma bias: $E[\hat{\theta}_n - \theta]$

Krok 1 Vypočítame $\hat{\theta}_n$

Krok 2 Vygenerujeme náhodnú vzorku veľkosti n z distribúcie $\hat{F}_n$: $\{X_i^* : i = 1, ..., n\}$ a vypočítame $\hat{\theta}_n^* = g(\bar{X}^*)$

Krok 3 Zopakujeme (2) veľakrát na výpočet $E^*\hat{\theta}_n^*$. Odhad biasu je $E^*\hat{\theta}_n^* - \hat{\theta}_n$. A náš bias-corrected estimátor je $\hat{\theta}_n - B_n^*$

7.5 Testy hypotéz

Na testy hypotéz potrebujeme dve veci: nulovú hypotézu a vhodnú testovaciu štatistiku (najlepšie pivotálnu), napríklad $\hat{T}_n = n^{1/2}\frac{\hat{\theta}_n - \theta_0}{\hat{se}_{\hat{\theta}_n}}$. Ale fantázii sa pri formulovaní hypotéz medze nekladú.

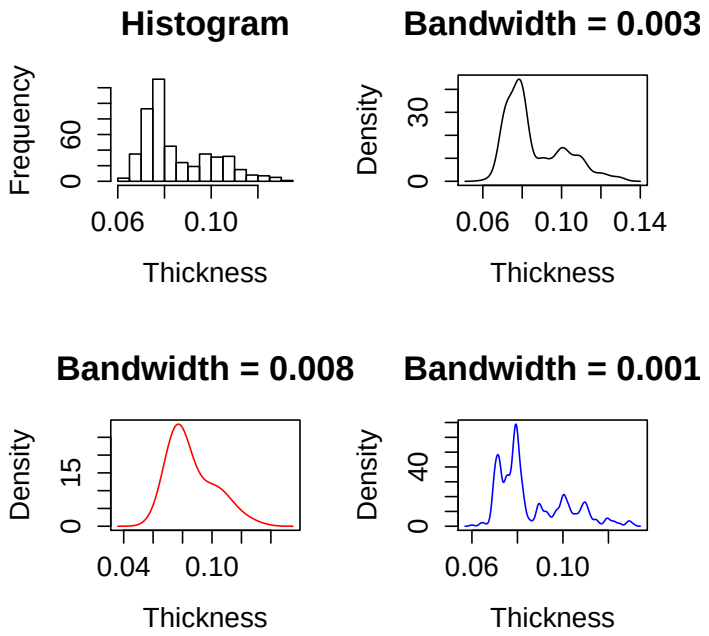
Zvoliť si nulovú hypotézu a testovaciu štatistiku však vôbec nemusí byť jednoznačné alebo jednoduché.

7.5.1 Hrúbka známok

Nasledujúci príklad je o známkach, je založený na štúdii [IS88] a opísaný v [ET94]. Historici považujú za zaujímavú otázku, či sa v minulosti tlačili známky na jeden druh papiera alebo nie. Pozorujeme hrúbky 485 známok, z histogramu nie je jasné, koľko hrbov má skutočná hustota. V závislosti od voľby bandwithu dostaneme rôzne počty hrbov.

Pre naše dáta je najmenšia hodnota vyhladzovacie parametra pri ktorom je neparametrický jadrový odhad hustoty s Gaussovským kernelom unimodálny $\hat{h}_1 = 0.0068$. Budeme testovať nulovú hypotézu H_0 : počet módov rozdelenia = 1. Prirodzenou distribúciou pre naše dáta za predpokladu nulovej hypotézy je $\hat{f}(t; \hat{h}_1)$. Distribúciu $\hat{f}(t; \hat{h}_1)$ ale mierne vylepšíme, kvôli tomu aby mala rovnakú varianciu ako naše dáta, teda $\hat{\sigma}^2$ (dá sa ukázať že náhodná premenná s hustotou $\hat{f}(t; \hat{h}_1)$ má varianciu $\hat{\sigma}^2 + \hat{h}_1^2$). Naša vylepšená distribúcia (aplikovali sme *variance stabilizing transformation*) nech je $\hat{g}(\cdot; \hat{h}_1)$. Budeme sa preto pozerať na

$$ASL_{boot} = P_{\hat{g}(\cdot;\hat{h}_1)}\left(\hat{h}_1^* > \hat{h}_1\right),$$



Obr. 6: Hrúbka známok, počet hrbov závisí od vyhladzovacieho parametra.

kde ASL značí dosiahnutý level signifikantnosti (*achieved significance level*) a $\hat{h}_1^*$ je najmenší možný vyhladovací parameter pri ktorom je distribúcia jadrový odhad hustoty unimodálny.

Nakoľko nesamplujeme bootstrapové datasety priamo z našej vzorky 485 hrúbiek známok, ale z hladkej distribúcie $\hat{g}(\cdot; \hat{h}_1)$, nazývame túto metódu hladký bootstrap (*smooth bootstrap*). Ešte potrebujeme vedieť samplovať z $\hat{g}(\cdot; \hat{h}_1)$, ktoré má tú skvelú vlastnosť, že má varianciu $\hat{\sigma}^2$ a strednú hodnotu rovnakú ako náhodná premenná pochádzajúca z rozdelenia s hustotou $\hat{f}(t; \hat{h}_1)$. To dosiahneme nasledovne: vyberieme bootstrapovú vzorku $y_1^*, ..., y_n^*$ z $\hat{F}_n$ a položíme

$$x_i^* = \bar{y}^* + (1 + \hat{h}_1^2/\hat{\sigma}^2)^{1/2}(y_i^* - \bar{y}^* + \hat{h}_1 \epsilon_i). \tag{7.1}$$

Algoritmus sa dá popísať nasledovne

Krok 1 Vyberieme B bootstrapových datasetov z $\hat{g}(\cdot; \hat{h}_1)$ pomocou (7.1)

Krok 2 Pre každý bootstrapový dataset vypočítame najmenšiu možnú hodnotu vyhladzovacieho parametra, pre ktorý je neparametrický odhad hustoty unimodálny. Označíme si B hodnôt ako $\hat{h}_1(1), ..., \hat{h}_1(B)$.

Krok 3 Aproximujeme ASL_{boot} pomocou

$$\hat{ASL}_{boot} = \#\{\hat{h}_1^*(b) \geq \hat{h}_1\}/B.$$

Pre $B = 5000$ nám vyšlo $\hat{ASL}_{boot} = 0.0002$, čo je menej ako 5%, preto zamietame nulovú hypotézu, že známky boli tlačené na jeden typ papiera na hladine významnosti 5%.

Kreatívna voľby testovacej štatistiky a nulovej hypotézu nám vylepšuje vlastnosti testu, napríklad zvyšuje šancu správneho zamietnutia nulovej hypotézy ak je nesprávna (zvyšuje silu testu). Preto sme napríklad volil ako nulovú hypotézu hodnotu parametra na hranici medzi jednohrbým a dvojhrbým rozdelením.

7.6 Konfidenčné intervaly

Existuje viacero typov konfidenčných intervalov v závislosti od toho ako sa použije bootstrapovaná distribúcia testovanej štatistiky.

- Canonical Bootstrap - Bootstrapová distribúcia použitá len na odhadnutie štandardnej chyby estimátora.

Krok 1 Vypočítame $\hat{\theta}_n$

Krok 2 Vygenerujeme náhodnú vzorku veľkosti n z $\hat{F}_n$: $\{X_i^* : i = 1, ..., n\}$ a vypočítame $\hat{\theta}_n^*$

Krok 3 Zopakujeme (2) veľakrát na výpočet empirickej distribúcie $\hat{\theta}_n^*$ a pomocou nej odhadneme štandardnú chybu $\hat{se}$. Označíme $t_{n,\alpha/2}$ ako $(\alpha/2)$ kvantil studentovho rozdelenia.
A $[\hat{\theta}_n - t_{n,1-\alpha/2} \cdot \hat{se}^*, \hat{\theta}_n - t_{n,\alpha/2} \cdot \hat{se}^*]$ označuje konfidenčný interval.

- Percentile Bootstrap - Kvantily bootstrapovej distribúcie použité priamo na konštrukciu konfidenčného intervalu.

Krok 1 Vypočítame $\hat{\theta}_n$

Krok 2 Vygenerujeme náhodnú vzorku veľkosti n z $\hat{F}_n$: $\{X_i^* : i = 1, ..., n\}$ a vypočítame $\hat{\theta}_n^*$

Krok 3 Zopakujeme (2) veľakrát na výpočet empirickej distribúcie $\hat{\theta}_n^*$. Označíme $\hat{\theta}_{n,\alpha/2}^*$ ako $(\alpha/2)$ kvantil tejto distribúcie. A $[\hat{\theta}_{n,\alpha/2}^*, \hat{\theta}_{n,1-\alpha/2}^*]$ označuje konfidenčný interval.

Tento konfidenčný interval je invariantný na transformáciu θ.

- Studentized Bootstrap Bootstrapová distribúcia použitá na odhadnutie kvantilu pivotovej štatistiky (dvojúrovňový bootstrap môže byť použitý ak nepoznáme vzťah pre štandardnú chybu estimátora).

Krok 1 Vypočítame $\hat{\theta}_n$

Krok 2 Vygenerujeme náhodnú vzorku veľkosti n z $\hat{F}_n$: $\{X_i^* : i = 1, ..., n\}$ a vypočítame $\hat{T}_n^* = n^{1/2}(\hat{\theta}_n^* - \hat{\theta}_n)/\hat{s}_n^*$

Krok 3 Zopakujeme (2) veľakrát na výpočet empirickej distribúcie T_n^*. Označíme $t_{n,\alpha/2}^*$ ako $(\alpha/2)$ kvantil tejto distribúcie. A $[\hat{\theta}_n - t_{n,1-\alpha/2}^*, \hat{\theta}_n - t_{n,\alpha/2}^*]$ označuje konfidenčný interval.

Tento konfidenčný interval nie je invariantný na transformáciu θ ale je presnejší ako percentile bootstrap.

- Bias Corrected and Accelerated Bootstrap - vhodný ak $\hat{\theta} \sim N(\theta + \text{bias}, \hat{se}^2)$ a zároveň umožňuje nekonštantný $\hat{se}$
- ABC - Approximate Bootstrap Confidence Interval - ľahšie počítateľná verzia BCa konfidenčných intervalov.

7.7 Keď bootstrap zlyhá

Kedy bootstrap nefunguje?

- Pre ťažkochvosté distribúcie s neexistujúcou varianciou, X_i je náhodná vzorka z Cauchy rozdelenia, $\hat{T}_n = \bar{X}$
- X_i je náhodná vzorka z $N(\mu, \sigma^2)$, $\hat{T}_n = n^{1/2}(\bar{X}^2 - \mu^2)$ ak $\mu \neq 0$, inak $T_n = n\bar{X}^2$.
- Maximum vzorky: F_0 má suport $[0, \theta_0]$. $\hat{\theta}_n = \max\{X_1, ..., X_n\}$. $\hat{T}_n = n(\hat{\theta}_n - \theta)$, $T_n^* = n(\hat{\theta}_n^* - \hat{\theta}_n)$. $P_n^*(T_n^* = 0) = 1 - (1 - 1/n)^n \to 1 - e^{-1}$ zatiaľčo $P(\hat{T}_n = 0) \to 0$.
- Parameter je na hranici parametrického priestoru: X_i je náhodná vzorka z $N(\mu, 1)$ kde $\mu \in [0, \infty)$ [Andrews (2000)]

Aký je teda problém bootstrapu? Chceme ho použiť v situáciách, ktoré sú komplexné, avšak môžu byť natoľko komplexné, že ani nevieme zaručiť, že bootstrap bude fungovať. V jednoduchých situáciách ho zasa nie je treba.

Nič nie je stratené. Alternatíva k Bootstrapu = Subsampling [PRW99].

- simulujeme **menšie** dátové vzorky **bez** opakovania
- dôležitý rozdiel: naše vzorky pochádzajú zo skutočnej distribúcie (F_0) a nie z nášho odhadnutého modely $(\hat{F}_n)$
- všeobecnejší ako bootstrap
- horší v situáciách kedy bootstrap funguje
- praktický problém → ako zvoliť veľkosť vzorky?

Najznámejšia kniha od objaviteľov bootstrapu je [ET94]. Jednoduchá a prístupná expozícia bootstrapu s príkladmi v jazyku R je napríklad tu: [Lar14], [boo] a [DK02]. Dve dôležité knižnice v Rku sú `bootstrap` (kódy ku všetkým funkciám spomenutých v [ET94]) a `boot`.

Animácie `https://www.stat.auckland.ac.nz/~wild/BootAnim/`

8 Kvantilová regresia

Ak chceme odhadnúť z náhodného výberu $\{Y_i\}_{i=1}^n$ strednú hodnotu $E(Y_i) = \mu$, môžeme to formulovať ako nasledovný minimalizačný problém

$$\hat{\mu} = \arg\min_{\mu \in \mathbb{R}} \sum_{i=1}^{n} (Y_i - \mu)^2.$$

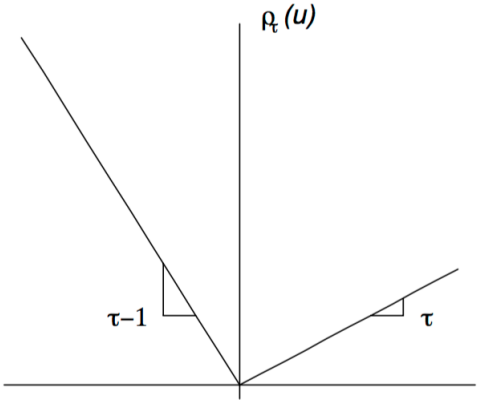
Ak predpokladáme, že podmienená stredná hodnota je lineárnou funkciou prediktorov $E(Y_i|x_i) = x_i^T \beta$, potom môžeme rozšíriť odhadovací problém a riešiť

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (Y_i - x_i^T \beta)^2.$$

Aj odhadnutie kvantilu, teda $Q(\tau) = \inf\{y : F(y) \geq \tau\}$, pre $0 < \tau < 1$, vieme formulovať ako optimalizačný problém. A to konkrétne takýto:

$$\hat{Q}(\tau) = \arg\min_{\xi \in \mathbb{R}} \sum_{i=1}^{n} \rho_\tau(Y_i - \xi),$$

kde $\rho_\tau(u) = u(\tau - I(u < 0))$.



Obr. 7: Funkcia $\rho_\tau(u) = u(\tau - I(u < 0))$.

Podobne ako cdf $F(y) = P(Y \leq y)$ aj kvantilová funkcia $Q(\tau)$ kompletne charakterizuje pravdepodobnostnú distribúciu náhodnej premennej Y.

V prípade ak predpokladáme, že podmienený kvantil Y_i je lineárnou funkciou prediktorov $Q(\tau|x_i) = x_i^T \beta$, pre prirodzené rozšíriť odhadovací problém analogicky ako v prípade strednej hodnoty a riešiť

$$\hat{\beta}(\tau) = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(Y_i - x_i^T \beta),$$

dôležitý špeciálny príklad je pre $\tau = 0.5$, kedy modelujeme podmienený medián a $\rho_{0.5}(u) = |u|$.

Prečo je výsledkom τ-kvantil? Minimalizujme

$$E(\rho_\tau(Y - \xi)) = (\tau - 1) \int_{-\infty}^{\xi} (y - \xi) dF(y) - \tau \int_{\xi}^{\infty} (y - \xi) dF(y),$$

cez ξ, diferencovaním cez ξ dostávame $0 = (1 - \tau) \int_{-\infty}^{\xi} dF(y) - \tau \int_{\xi}^{\infty} dF(y) = F(\xi) - \tau$.

Pri lineárnej regresii modelujeme len podmienenú strednú hodnotu Y_i, kvantilová regresia nám však ponúka oveľa komplexnejší pohľad na distribúciu Y_i. Pre rôzne kvantily Y_i môžu byť citlivosti na prediktory X_i výrazne rozdielne.

Kvantilová regresia teda odpovedá na otázku: 'Ako podmienená distribúcia Y_i závisí od x_i?' Kým lineárna regresia uvažuje efekt prediktorov len na posun distribúcie vysvetľovanej premennej, kvantilová regresia uvažuje aj potenciálne efekty aj na tvar distribúcie. Napríklad rekvalifikačný kurz môže mierne predĺžiť krátke doby nezamestanosti ale na druhej strane výrazne zmenšiť pravdepodobnosť dlhého zotrvania v nezamestanosti. V tomto prípade môže byť priemerný efekt malý ale efekt na tvar distribúcie výrazný.

Dôležitá vlastnosť kvantilovej regresie je, že pre monotónnu transformáciu $h(.)$ platí $Q_{h(Y)}(\tau|x) = h(Q_Y(\tau|x))$. Naopak v lineárnej regresii vo všeobecnosti $E(h(Y)|x) \neq h(E(Y|x))$. Vďaka tejto vlastnosti ak vysvetľujeme podmienený medián transformovanej $h(Y_i)$, potom môžeme $h^{-1}(x_i^T\hat{\beta})$ interpretovať ako podmienený medián pôvodného Y_i pre dané x_i.

8.1 Optimalizačný problém

Pre danú hodnotu τ je úloha nájdenia $\hat{\beta}(\tau)$ ekvivalentná riešeniu úlohy lineárneho programovania (optimalizácia lineárnej funkcie za platnosti lineárnych reštrikcií).

$$\hat{\beta}(\tau) = \arg\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(Y_i - x_i^T\beta),$$

$$\iff$$

$$\min_{(\beta,u,v)} \tau 1_n^T u + (1-\tau)1_n^T v$$

$$s.t.$$

$$y - X\beta = u - v$$

$$\beta \in \mathbb{R}^p, \quad u \geq 0, \quad v \geq 0$$

Základné metódy riešenia

- Simplexová metóda (veľmi pomalá pre veľké problémy)
- Metódy vnútorného bodu
- Metódy vnútorného bodu s preprocessingom
- Metódy vnútorného bodu pre riedke problémy (matica X má veľa núl)

V programe R je najpoužívanejšou knižnica `quantreg`, na ktorej pracoval sám Roger Koenker.

8.2 Efekt na kvantil (quantile treatment effect)

Majme náhodnú premennú $Y \sim F$ a $Y + \Delta(Y) \sim G$. Uvažujme $\Delta(y)$ také, že $F(y) = G(y + \Delta(y))$ a teda $\Delta(y)$ je jednoznačné, a to $\Delta(y) = G^{-1}(F(y)) - y$. Ak nastavíme y ako τ-kvantil rozdelenia F (teda $\tau = F(y)$), potom efekt na kvantil (quantile treatment effect) je

$$\delta(\tau) = \Delta(F^{-1}(\tau)) = G^{-1}(\tau) - F^{-1}(\tau),$$

priemerný efekt dostaneme integrovaním QTE podľa τ:

$$\bar{\delta} = \int_0^1 \delta(\tau)d\tau = \int G^{-1}(\tau)d\tau - \int F^{-1}(\tau)d\tau = \mu(G) - \mu(F)$$

V prípade diskrétnej premennej D_i

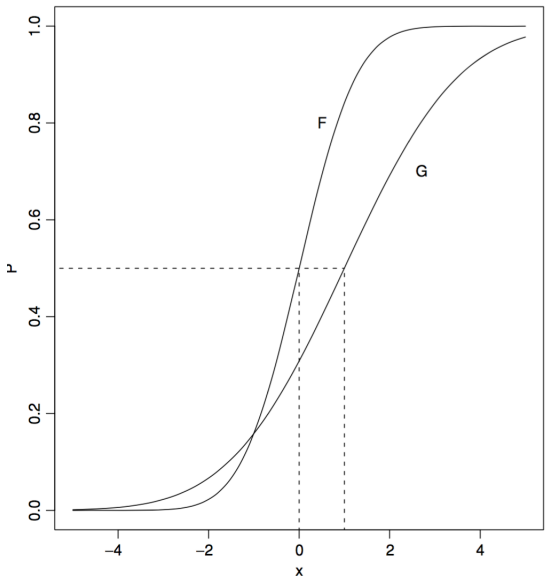
$$Q_{Y_i}(\tau|D_i) = \alpha(\tau) + \delta(\tau)D_i,$$

a QTE je zmena na τ kvantilu, ktorá je spôsobená zmenou z $D_i = 0$ na $D_i = 1$.

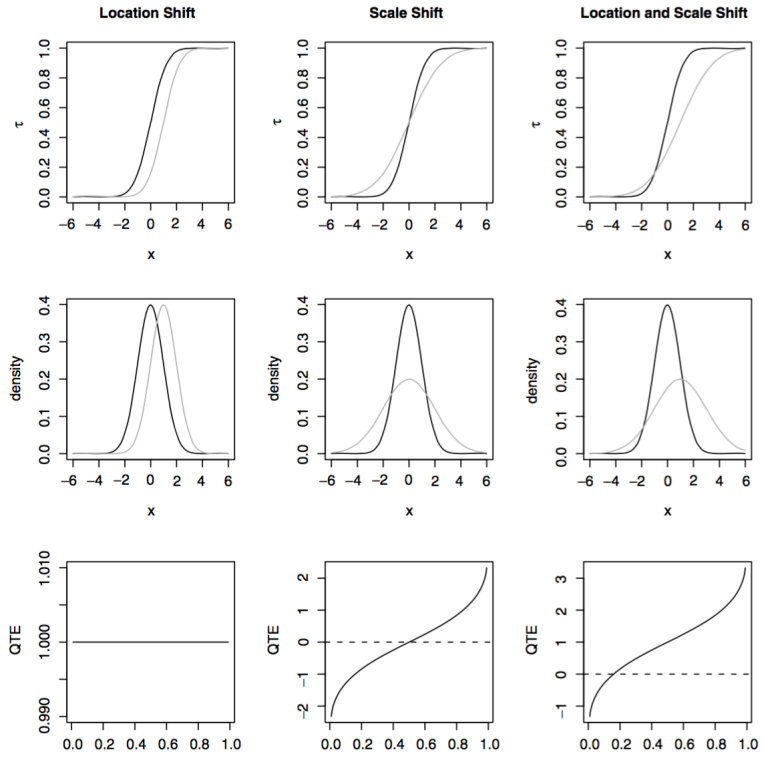
V prípade spojitej premennej X_i

$$Q_{Y_i}(\tau|X_i) = \alpha(\tau) + x_i^T\beta(\tau),$$

a QTE je zmena na τ kvantilu, ktorá je spôsobená zmenou z X_i na $X_i + 1$.

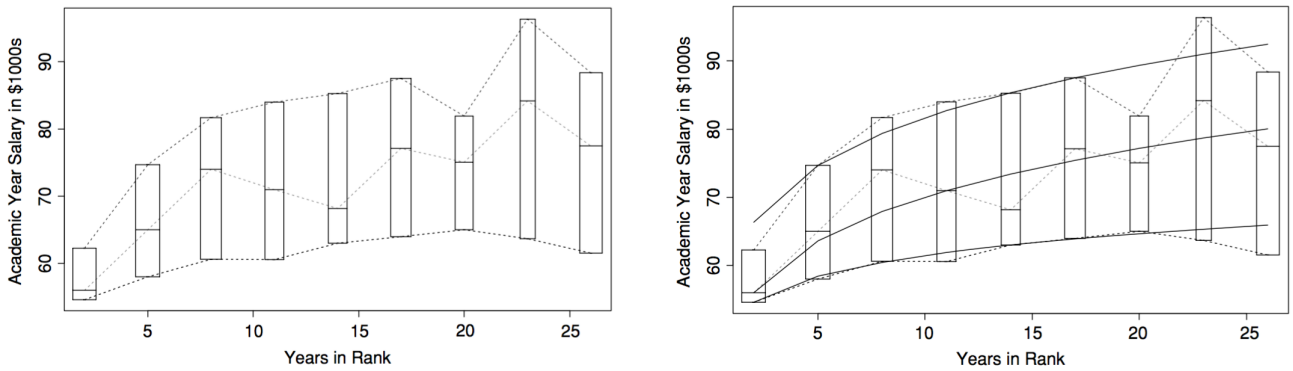


Obr. 8: Horizontálna vzdialenosť medzi G a F je QTE, vidíme, že $\Delta(y)$ má rôzny efekt v rôznych častiach distribúcie. Pre medián a vyššie kvantily je QTE pozitívne, pre nižšie kvantily zase negatívne.

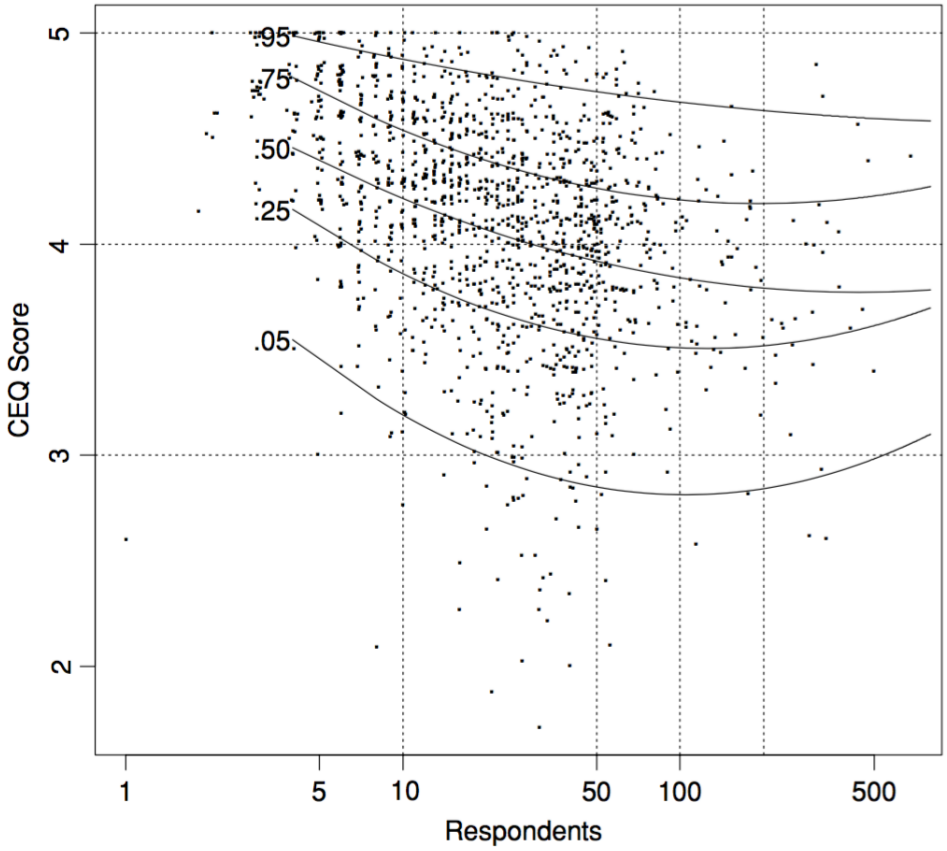


Obr. 9: $\Delta(X)$ môže mať rôzny efekt na distribúciu Y

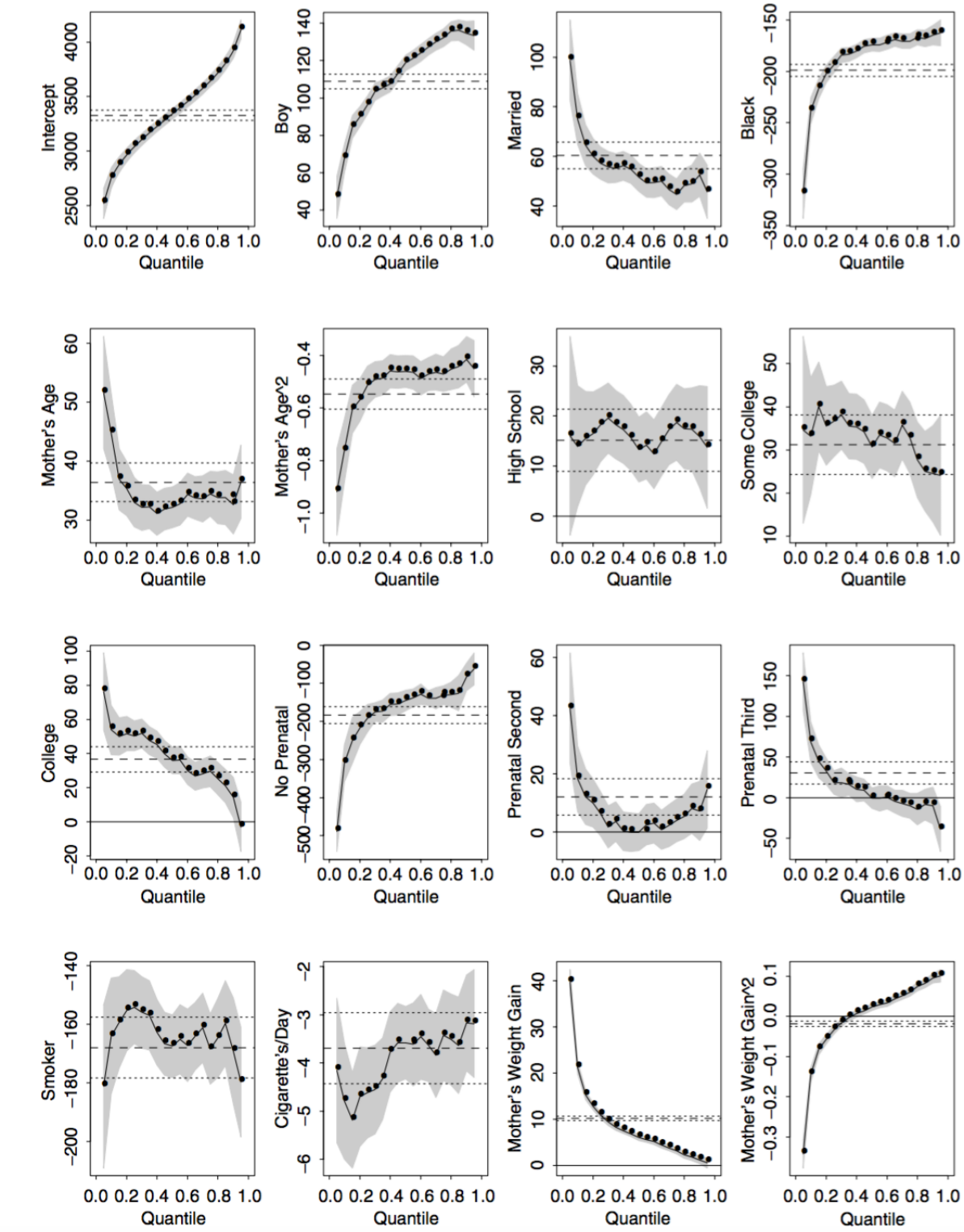
8.3 Príklady



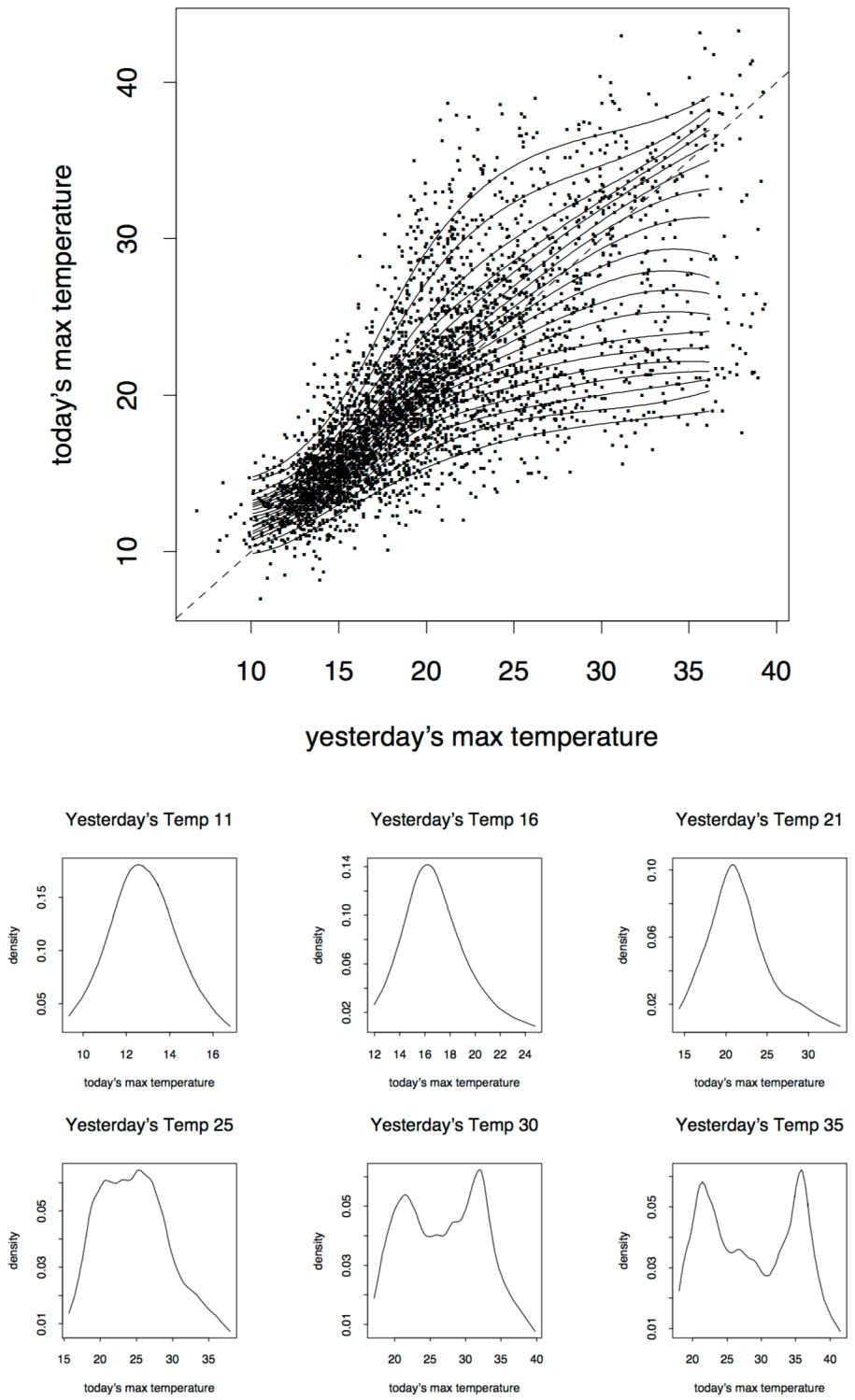
Obr. 10: **Mzdy a pracovné skúsenosti:** Ročná mzda v tisícoch dolárov v závislosti od počtu rokov odpracovaných ako profesor štatistiky. V pravo fit modelu $Q_{\log(Y_i)}(\tau|X_i) = \alpha + \beta \log(X_i)$, pre kvartily. ([Koe05])



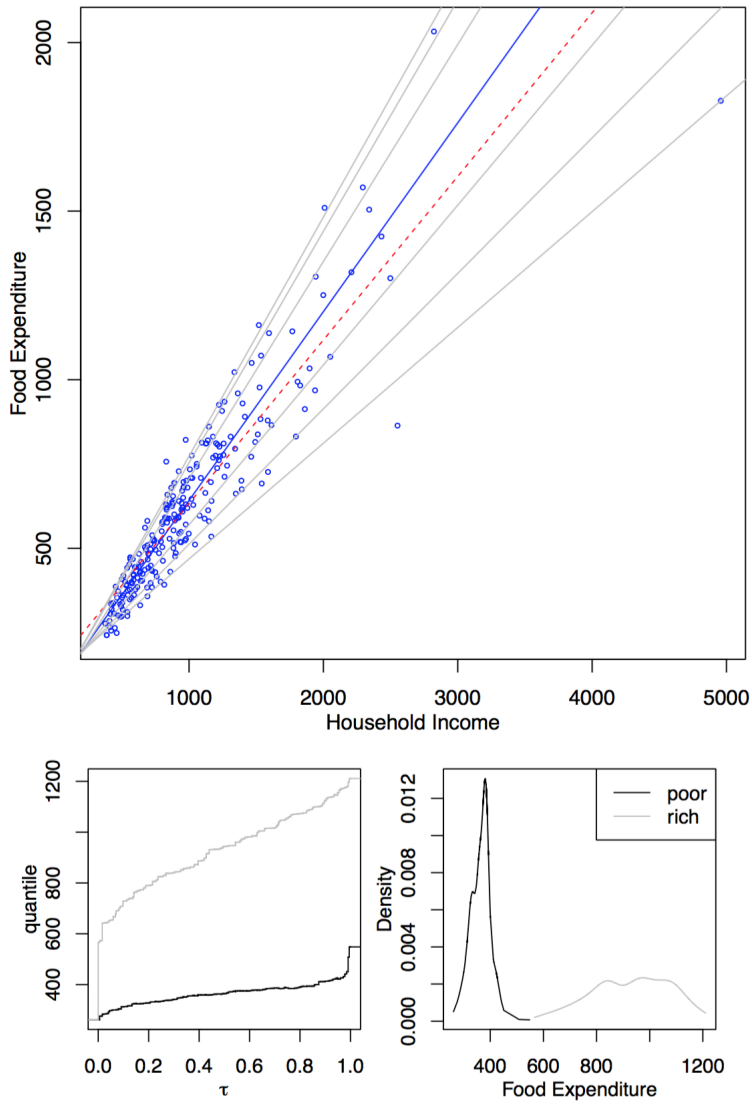
Obr. 11: **Hodnotenie kurzov a veľkosť tried:** Hodnotenie kurzu v závislosti od počtu študentov zapísaných na predmet. $Q_{Y_i}(\tau|X_i) = \beta_0(\tau) + \beta_1(\tau)\text{Size} + \beta_2(\tau)\text{Size}^2$. ([Koe05])



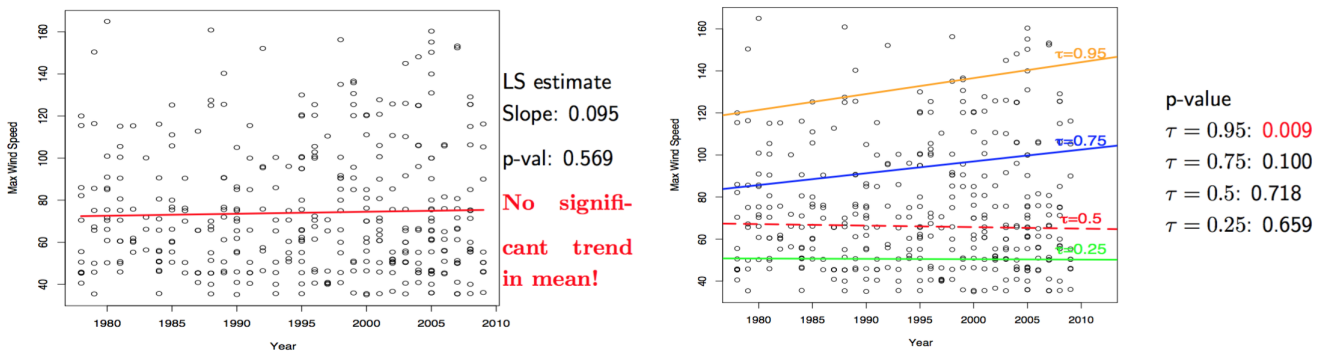
Obr. 12: **Nízka pôrodná váha:** Pôrodná váha dieťaťa v závislosti od rôznych parametrov. $Q_{Y_i}(\tau|X_i) = \beta(\tau)X_i$. ([AD08])



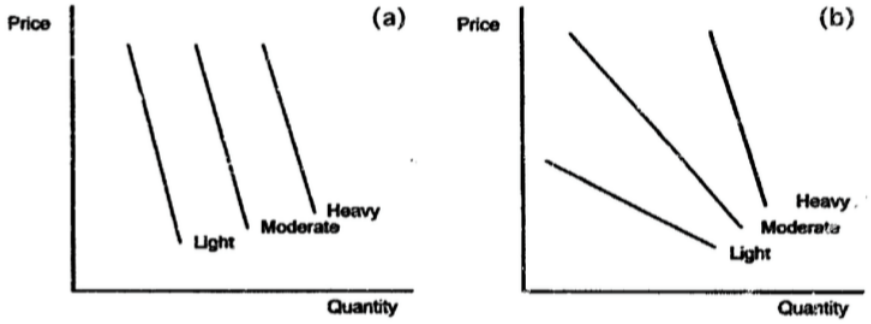
Obr. 13: **Včerajšia teplota predikuje súčasnú:** Kvantilový AR(1) proces. ([Koe05])



Obr. 14: **Výdavky na jedlo a príjem:** Dolu porovnanie rozdelení výdavkov na jedlo pre domácnosti na 10% a 90% kvantile príjmu. Pôvodné dáta z [Eng57].



Obr. 15: **Maximálna rýchlosť vetra v čase:** Mení sa priemerná maximálna rýchlosť v čase? vs. Menia sa kvantily rýchlosti maximálnej rýchlosti vetra v čase? Dáta z [HBG96].



Obr. 16: **Citlivosť dopytu po alkohole:** Je na rozdielnych kvantiloch výrazne rozdielna [MBM95].

9 Zovšeobecnené aditívne modely

9.1 Aditívne model

Majme **lineárny model** s ktorým sa stretávame doteraz.

$$Y = X\beta + \epsilon$$

tento model vie byť veľmi flexibilný, prediktory môžeme transformovať, pridávať interakčné členy. Inšpekcia obrázkov môže tiež pomôcť navrhnúť vhodnú transformáciu. Poľahky si však môžeme čosi nevšimnúť, nakoľko sa pozeráme vždy len na jeden obrázok.

Alternatívou je **neparametrický model**, ktorý je flexibilnejší a vhodnú transformáciu nájde automaticky. Vieme však, že pri vyššom počte prediktorov, na uspokojivé odhady potrebujeme obrovskú dátovú vzorku.

$$Y = f(X) + \epsilon$$

Niečo medzi lineárnym a neparametrickým modelom je **aditívny model**.

$$Y = \beta_0 + \sum_{j=1}^{p} f_j(X_j) + \epsilon,$$

kde f_j sú (dostatočne) hladké funkcie.

- Flexibilnejšie ako lineárny model
- Jednotlivé funkcie f_j sú zakresliteľné a interpretovateľné ako marginálna asociácia medzi prediktorom a vysvetľovanou premennou
- Transformácie f_j sú nachádzané systematickým spôsobom
- Nepotrebujú takú veľkú dátovú vzorku ako neparametrický model

Ale čo s kategorickými premennými? Tam transformácie nie sú zaujímavé.

$$Y = \beta_0 + \sum_{j=1}^{p} f_j(X_j) + Z\gamma + \epsilon,$$

Ako odhadnúť funkcie $\{f_j\}$? Tu je **backfitting algoritmus** od Hastie a Tibshirani (1990) z knižnice `gam`. Tento algoritmus konverguje za veľmi všeobecných podmienok (idea je podobná ako Gauss-Seidelova metóda na riešenie systému lineárnych rovníc).

Krok 1 Inicializujeme $\beta_0 = \bar{Y}$ a $f_j(x) = \hat{\beta}_j x$, kde $\hat{\beta}$ je napríklad odhad MNŠ.

Krok 2 Postupne prechádzame cez $j = 1, ..., p, 1, ..., p, 1, ...$

$$f_j = S(x_j, Y - \beta_0 - \sum_{i \neq j} f_i(X_i)),$$

kde funkcia $S(x, y)$ je vyhladzovač dát (x, y). (neparametrický ako napr. jadrový odhad, splajn, loess alebo parametrický ako napr. lineárny alebo polynomiálny). Na rozdielne prediktory môžeme použiť rozdielne vyhadzovače s rozdielnymi vyhladzovacími oknami. Zastavíme keď sa funkcie $\{f_j\}$ príliš nemenia.

V knižnici `mgcv` sa používajú splajny s automatickou voľbou vyhladzovacieho okna. Každá funkcia $f_j(x) = \sum_i \beta_i \phi_i(x)$ je lineárnou kombináciou splajnových bázických funkcií. Požadujeme aby splajny boli dostatočne hladké a preto penalizujeme $\int [f_j''(x)]^2 dx = \beta_j^T S_j \beta_j$ (kde matica S_j závisí od splajnových bázických funkcií), potom maximalizujeme

$$\log L(\beta) - \sum_j \lambda_j \beta_j^T S_j \beta_j,$$

kde penalizačné parametre zabezpečujúce hladkosť λ_j sú nastavené pomocou krížovej validácie.

Aditívne modely nám môžu pomôcť s nájdením vhodnej transformácie pre lineárny model.

Keď porovnávame modely, potrebujeme vedieť počet stupňov voľnosti. Na to potrebujeme vedieť počet parametrov, to je však pri splajnoch problematické. Keď si zoberieme množinu lineárnych vyhladzovačov $\hat{Y} = PY$, tak lineárnej regresií zodpovedá $P = X(X^TX)^{-1}X$ a počet parametrov je hodnosť matice X a v tomto prípade je to aj stopa matice P.[5] Preto ekvivaltený počet parametrov pre aditívne modely je $tr(P)$.

Vďaka aditívnym modelom môžeme vybádať, že akási závislosť mení sklon. Toto môže byť modelované "hokejkami": Ľavá hokejka nech je $lhs(x) = (c - x)1_{x<c}$ a pravá hokejka $rhs(x) = (x - c)1_{x>c}$. Model s jedným prediktorom, ktorý umožnuje zmenu sklonu je nasledovný

$$Y = \beta_0 + \beta_1 lhs(X_1) + rhs(X_2) + \epsilon,$$

aditívny model nám pomôže s nájdením bodu zlomu c a preto súčasťou exploratívnej dátovej analýzy.

Diagnostické nástroje sú rovnaké ako v prípade lineárneho modelu, podobne heteroskedasticitu môžeme odhadovať iteratívnou zmenou váh ako to môžeme robiť v lineárnom modeli.

9.2 Zovšeobecnený aditívny model

Oproti GLM

$$\eta = X\beta, \quad E(Y) = \mu, \quad g(\mu) = \eta, \quad Var(Y) \propto V(\mu),$$

je jediný rozdiel v tom, že lineárny prediktor bude mať tvar $\beta_0 + \sum_{j=1}^{p} f_j(X_j)$.

Odhadovací algoritmus založený na IRWS sa zmení jedine v tom, že v každom kroku budeme odhadovať funkcie $\{f_j\}$,

9.3 Model s meniacou sa podmienenou strednou hodnotou (Alternating Conditional Expectations)

ACE je špeciálnym typom TBS (transform-both-sides) modelu

$$\theta(Y) = \beta_0 + \sum_{j=1}^{p} f_j(X_j) + \epsilon,$$

kde minimalizujeme $\sum_i (\theta(Y_i) - \sum f_j(X_{ij}))^2$. Aby sme sa vyhli triviálnemu riešeniu $\theta = f_j = 0$, pridáme podmienku $Var(\theta(Y)) = 0$. Na odhadovanie môžeme použiť nasledujúci algoritmus.

Krok 1 Inicializujeme $\theta(Y) = \frac{Y - \bar{Y}}{SD(Y)}$, $f_j(x) = \hat{\beta}_j x$, kde $\hat{\beta}$ je napríklad odhad MNŠ.

Krok 2 Postupne prechádzame cez $j = 1, ..., p, 1, ..., p, 1, ...$

$$f_j = S(x_j, \theta(Y) - \sum_{i \neq j} f_i(X_i)),$$

$$\theta = S(y, \sum_{i \neq j} f_i(X_i)),$$

$$\theta(y) \leftarrow \frac{\theta(y) - \theta(\bar{y})}{SD(\theta(y))}$$

Zastavíme keď sa funkcie $\{f_j\}$ a θ príliš nemenia.

[5] Pre projekčnú maticu platí, že hodnosť sa rovná jej stope. Prečo? $tr(P) = tr(X(X^TX)^{-1}X) = X^TX(X^TX)^{-1} = tr(I_p) = p$. Využili sme $tr(AB) = tr(BA)$.

Vlastnosti ACE modelu

- Má tendenciu overfitovať
- Riešenie je citlivé na marginálne distribúcie prediktorov
- Pre model s jedným parametrom je ACE symetrické v X a Y. Ale ak $Y = X + \epsilon,\ X \sim U(0,1),\ \epsilon \sim N(0,1)$, potom $E(Y|X) = X$ ale $E(X|Y) \neq Y$.
- Je vhodný na exploratórnu analýzu, vďaka ACE modelu môžeme objaviť vhodnú transformáciu.

Kanonické korelácie (Canonical Correlations)

ACE model vyrástol z myšlienky kanonických korelácií. Majme 2 množiny náhodných premenných Y a X, a nájdime vektory dĺžky 1 a a b, také, aby sme maximalizovali koreláciu medzi $a^T X$ a $b^T Y$. Hlavný odkaz je asi nasledovný: ak z kanonickej korelácie alebo ACE modelu vypadne veľmi zlý fit, akoukoľvek dátovou analýzou sa asi ďaleko nedostaneme.

9.4 Aditivita a stabilizácia variancie (Additivity and Variance Stabilization)

Myšlienka je nasledovná: Zvolíme $\{f_j\}$ na to, aby sme dostali dobrý fit a zvolíme θ tak aby sme dostali konštantnú varianciu $Var(\theta(Y)|\sum_{j=1}^{p} f_j(X_j)) = const.$

Ak máme náhodnú premennú Y s varianciou $Var(Y) = V(Y)$, potom nasledovná transformáciu zabezpečí konštantnú varianciu.

$$\theta(t) = \int_0^t \frac{1}{\sqrt{V(u)}} du.$$

Konštantná variancia nám zabezpečí presnejší odhad smerodajných odchýlok estimátorov, nie nutne najlepší fit. Odpovedať na otázku: "chcem presnejšiu predikciu alebo presnejšiu informáciu o neistote?"si už musí odpovedať každý sám.

9.5 Viacrozmerné adaptívne regresné splajny (Multivariate Adaptive Regression Splines)

MARS model je založený na:

$$\hat{f}(x) = \sum_{j=1}^{k} c_j B_j(x),$$

kde bázické funkcie majú tvar $[\pm(x_i - t)]_+^q$. Pre $q = 1$ dostávame po častiach lineárnu funkciu.

Iteratívne pridávame bázické funkcie, ktoré produkujú najlepší fit. Oproti čisto aditívnemu modelu môžeme pridávať aj interakčné členy.

9.6 Literatúra

Kapitola 12 v [Far05], pozor, v poslednom príklade o MARS je chyba a v druhom vydaní tejto knihy je nový príklad. Knižka [HT90] a GAM zasadené v širšom kontexte ostatných metód je [HTFF05].

10 Panelové dáta

Mnohokrát sme v situácii, že rovnaké subjekty (individuáli, rodiny, firmy, mestá, štáty, ...) pozorujeme vo viacerých časoch. V tom prípade máme jeden dataset pre každý čas a otvárajú sa nám nové možnosti, ktoré sme predtým nemali. Predtým, než sa budeme bližšie venovať takýmto **panelovým dátam** si vysvetlíme dva kľúčové pojmy - *fixed effects* a *random effects*.

Problémom je, že nemáme náhodnú vzorku, pozorovania nie sú nezávislé, nepozorované faktory, ktoré ovplyvňujú niekoho mzdu v roku 2010 ju ovplyvňujú aj v roku 2011; nepozorované faktory ovplyvňujúce kriminalitu v nejakom meste v roku 2015 ju budú ovplyvňovať aj v 2020.

Príklady

- Vplyv nezamestanosti na kriminalitu, dáta na úrovni miest
- Súvis odpracovaných hodín a dĺžky spánku, dáta na úrovni individuálov
- Vplyv objasnenia zločinu na kriminalitu. dáta na úrovni miest
- Vplyv dopravného zákona na úmrtnosť na cestách, dáta na úrovni miest
- Vplyv členstva v odboroch na mzdu, dáta na úrovni individuálov.

Majme subjekty $i = 1, ..., n$ (človek, firma, krajina,...) a čas $t = 1, ..., T$ a u_{it} nech je náhodná chyba so strednou hodnotou 0. Majme takýto všeobecný model

$$Y_{it} = \alpha_{it} + X_{it}\beta_{it} + u_{it}$$

tu však máme priveľa parametrov na to aby sme ich odhadli len s $N = n \times T$ pozorovaní. Ak urobíme predpoklad homogenity parametrov, teda $\alpha_{it} = \alpha, \beta_{it} = \beta$ pre všetky i, t, potom takýto zjednodušený model môžeme poľahky odhadnúť metódou najmenších štvorcov ako keby sme mali N pozorovaní

$$Y_{it} = \alpha + X_{it}\beta + u_{it}. \tag{10.1}$$

Takýto model však nie je zaujímavý, pre každého individuála i máme skrátka viacej (konkrétne T pozorovaní) a to je všetko. Môžeme však uvažovať model v ktorom bude v chybách u_{it} akási štruktúra. Môžeme predpokladať, že $u_{it} = \mu_i + \epsilon_{it}$ a teda, že existuje akýsi faktor, ktorý je špecifický pre individuála i.

$$Y_{it} = \alpha + X_{it}\beta + \mu_i + \epsilon_{it}.$$

Komponent chyby μ_i je náhodná premenná. Ak $cov(\mu_i, X_{it}) \neq 0$, potom β odhadnuté pomocou metódy najmenších štvorcov bude nekonzistentná. V tomto prípade náš parameter α nahradíme α_i a odhadneme pomocou MNŠ. To je to isté ako kebyže z naších dát najprv pre každého individuála urobíme priemer v čase a urobíme regresiu len na priemerné pozorovania. Takémuto modelu hovoríme **fixed effects** model. Fixed neznamená nenáhodný ale len to, že sa nemení v čase. Iný spôsob ako sa zbaviť μ_i je odhadnúť parameter β z regresie rozdielov $(Y_{it} - Y_{it-1}) = (X_{it} - X_{it-1})\beta + (u_{it} - u_{it-1})$ (*First Differencing*). Toto je jednoduchý spôsob ako sa zbaviť nepozorovovaných vplyvov, ktoré sú nemeniace sa v čase. Oba tieto spôsoby majú ten problém, že všetkej variácie v prediktoroch, ktorá sa nemení v čase. Nie sú to preto vhodné metódy ak sa prediktory nemenia príliš v čase. Ktorý si vybrať? Ak $T = 2$, tak nám dajú identické výsledky. Ak sú chyby seriálne korelované (napríklad ak sú náhodnou prechádzkou), je lepšie použiť FD estimátor.

V prípade ak $cov(\mu_i, X_{it}) = 0$, tak potom aj u_{ij} je nekorelované s regresormi a preto je odhad MNŠ konzistentný. Takémuto modelu hovoríme **random effects** model. Na druhej strane, takýto odhad zvyčajne nie je efektívny, nakoľko korelačná matica chýb nemusí byť s nulami mimo diagonály. Toto je špeciálny typ heteroskedasticity a tu môžeme (podobne ako v iných prípadoch s nehomoskedastickými chybami) použiť *feasible generalized least squares* estimátor (minimalizujeme najmenšie štvorce vážené inverznou odhadnutou kovariančnou maticou).

Je tu trade-off, random effects model je založený na silnejšom predpoklade a dáva efektívnejšie (teda s menšou varianciou, teda presnejšie) odhady ako fixed effects model v prípade, že je predpoklad $cov(\mu_i, X_{it}) = 0$ splnený. Na druhej strane, ak tento predpoklad nie je splnený, tak RE model nám dáva nekonzistentné výsledky. Na porovnanie RE a FE modelu sa používa Hausmanov test, nulovou hypotézou je, že predpoklad $cov(\mu_i, X_{it}) = 0$ je splnený a preto, FE nie je výrazne rozdielny ako RE.

10.1 Fixed Effects

Prepíšme si $Y_{it} = X_{it}\beta + \alpha_i + u_{it}$ do maticovej podoby

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \beta + \begin{bmatrix} \boldsymbol{i} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{i} & \dots & \mathbf{0} \\ & & \vdots & \\ \mathbf{0} & \mathbf{0} & \vdots & \boldsymbol{i} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \dots \\ \alpha_n \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix},$$

kde Y_i a ϵ_i sú $T \times 1$ vektory, X_i je $T \times p$ matica a $\boldsymbol{i}$ a a $\mathbf{0}$ označujú $T \times 1$ vektor jednotiek.

Tuto môžeme odhadnúť parametre pomocou MNŠ, avšak to môže byť numericky náročné. Ak označíme $D = \begin{bmatrix} \boldsymbol{i} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{i} & \dots & \mathbf{0} \\ & & \vdots & \\ \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{i} \end{bmatrix}$, potom vieme odhadnúť koeficient β nasledovne

$$\hat{\beta} = (X^T M_D X)^{-1}(X^T M_D Y), \tag{10.2}$$

kde

$$M_D = I - D(D^T D)^{-1} D^T,$$

teda $\hat{\beta}$ môžeme dostať tak, že najprv pretransformujeme dáta $X_* = M_D X$ a $Y_* = M_D Y$ a potom urobíme lineárnu regresiu kde Y_* vysvetľujeme pomocou X_*. M_D je matica reziduí z projekcie na konštanty špecifické pre individuála. Maticu M_D vieme prepísať ako $M_D = \begin{bmatrix} M_0 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & M_0 & \dots & \mathbf{0} \\ & & \vdots & \\ \mathbf{0} & \mathbf{0} & \dots & M_0 \end{bmatrix}$, kde $M_0 = I_T - \frac{1}{T}\boldsymbol{i}\boldsymbol{i}^T$. Teda, koniec koncov, ide o regresiu priemerov cez čas!

Koeficienty pri dummy premenných dostaneme ako $\hat{\alpha}_i = \bar{Y}_i - \bar{X}_i\beta$.

Ako otestujeme signifikantnosť efektov špecifických pre individuálov ($\alpha_i = 0$)?

$$F_{n-1,nT-n-p} = \frac{R^2_{LSDV} - R^2_{Pooled}/(n-1)}{(1 - R^2_{LSDV})/(nT - n - p)},$$

kde *LSDV* značí least square dummy variable teda R^2 z regresie kde odhadom je (10.2) a *Pooled* je z regresie (10.1), kde $\alpha_i = 0$.

- Pôvodná formulácia

$$Y_{it} = X_{it}\beta + \alpha + \epsilon_{it}$$

$$S^{total}_{XX} = \sum_{i=1}^{n}\sum_{t=1}^{T}(X_{it} - \bar{X})(X_{it} - \bar{X})^T, \qquad S^{total}_{XY} = \sum_{i=1}^{n}\sum_{t=1}^{T}(X_{it} - \bar{X})(Y_{it} - \bar{Y})$$

- Regresia deviácie zo stredných hodnôt

$$Y_{it} - \bar{Y}_{i.} = (X_{it} - \bar{X}_{i.})\beta + \epsilon_{it} - \bar{\epsilon}_{i.}$$

$$S^{within}_{XX} = \sum_{i=1}^{n}\sum_{t=1}^{T}(X_{it} - \bar{X}_{i.})(X_{it} - \bar{X}_{i.})^T, \qquad S^{within}_{XY} = \sum_{i=1}^{n}\sum_{t=1}^{T}(X_{it} - \bar{X}_{i.})(Y_{it} - \bar{Y}_{i.})$$

- Regresia stredných hodnôt

$$\bar{Y}_{i.} = \bar{X}_{i.}\beta + \alpha + \bar{\epsilon}_{i.}$$

$$S^{between}_{XX} = \sum_{i=1}^{n}\sum_{t=1}^{T}(\bar{X}_{i.} - \bar{X})(\bar{X}_i - \bar{X})^T, \qquad S^{between}_{XY} = \sum_{i=1}^{n}\sum_{t=1}^{T}(\bar{X}_{i.} - \bar{X})(\bar{Y}_{i.} - \bar{Y})$$

Poznamenajme, že $S_{XX}^{total} = S_{XX}^{within} + S_{XX}^{between}$ a $S_{XY}^{total} = S_{XY}^{within} + S_{XY}^{between}$.

Algebraickými manipuláciami sa dá ukázať, že $\hat{\beta}^{total} = [S_{XX}^{total}]^{-1} S_{XY}^{total} = F^{within}\hat{\beta}^{within} + F^{between}\hat{\beta}^{between}$, kde

$$F^{within} = (S_{XX}^{within} + S_{XX}^{between})^{-1} S_{XX}^{within} = I - F^{between}.$$

Teda $\hat{\beta}^{total}$ sa dá vyjadriť ako maticovo váhovaný priemer $\hat{\beta}^{within}$ a $\hat{\beta}^{between}$.

10.2 Random Effects

Preformulujme model nasledovne

$$Y_{it} = \alpha + X_{it}\beta + u_i + \epsilon_{it}$$

Na u_i sa môžeme pozerať ako na súbor faktorov, ktoré nie sú v regresii a sú špecifické pre individuála i. Uvažujme nasledujúcu sadu predpokladov

- $E(\epsilon_{it}|X) = E(u_i|X) = 0$
- $E(\epsilon_{it}^2|X) = \sigma_\epsilon^2$
- $E(u_i^2|X) = \sigma_u^2$
- $E(\epsilon_{it}u_j|X) = 0$, pre všetky i, t, j
- $E(\epsilon_{it}\epsilon_{js}|X) = 0$, pre všetky $t \neq s, i \neq j$
- $E(u_i u_j|X) = 0$, pre všetky $i \neq j$

Korelačná matica pre $\eta_i = [\eta_{i1}, \ldots, \eta_{iT}^T]$, kde $\eta_{ij} = \epsilon_{it} + u_i$ je

$$\Sigma = \begin{bmatrix} \sigma_\epsilon^2 + \sigma_u^2 & \sigma_u^2 & \sigma_u^2 & \ldots & \sigma_u^2 \\ \sigma_u^2 & \sigma_\epsilon^2 + \sigma_u^2 & \sigma_u^2 & \ldots & \sigma_u^2 \\ & \ldots & & & \\ \sigma_u^2 & \sigma_u^2 & \sigma_u^2 & \ldots & \sigma_\epsilon^2 + \sigma_u^2 \end{bmatrix} = \sigma_\epsilon^2 I_T + \sigma_u^2 \boldsymbol{i}_T \boldsymbol{i}_T^T$$

a celková kovariančná matica pre všetkých nT pozorovaní je (lebo pozorovania individuálov i a j sú nezávislé)

$$\Omega = \begin{bmatrix} \Sigma & 0 & 0 & \ldots & 0 \\ 0 & \Sigma & 0 & \ldots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \ldots & \Sigma \end{bmatrix} = I_n \otimes \Sigma$$

Odhad GLS je

$$\hat{\beta} = (X^T \Omega^{-1} X)^{-1} X^T \Omega^{-1} y$$

kde $\Omega^{-1/2} = [I_n \otimes \Sigma]^{-1/2}$. Dá sa ukázať, že

$$\Sigma^{-1/2} = \frac{1}{\sigma_\epsilon}\left[I - \frac{\theta}{T}\boldsymbol{i}_T \boldsymbol{i}_T^T\right],$$

kde

$$\theta = 1 - \frac{\sigma_\epsilon}{\sqrt{\sigma_\epsilon^2 + T\sigma_u^2}}.$$

Transformácia zabezpečujúca homoskedastické chyby je preto

$$\Sigma^{-1/2} y_i = \frac{1}{\sigma_\epsilon} \begin{bmatrix} Y_{i1} - \theta\bar{Y}_{i.} \\ Y_{i2} - \theta\bar{Y}_{i.} \\ \vdots \\ Y_{iT} - \theta\bar{Y}_{i.} \end{bmatrix}.$$

Špeciálny prípad je pre $\theta = 1$ a to je LSDV model.

Podobne aj tento GLS estimátor je maticovo-vážený priemer within a between estimátorov.

$$\hat{\beta} = \hat{F}^{within}\hat{\beta}^{within} + (I - \hat{F}^{within})\hat{\beta}^{between},$$

kde

$$\hat{F}^{within} = [S_{XX}^{within} + \lambda S_{XX}^{between}]^{-1} S_{XX}^{within}, \qquad \lambda = \frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + T\sigma_u^2} = (1-\theta)^2$$

Špeciálne prípady

- $\lambda = 1$, GLS nám dá to isté ako obyčajná MNŠ. $\sigma_u^2 = 0$
- $\lambda = 0$, GLS nám dá to isté ako LSDV model pre fixed effects model, $\sigma_\epsilon^2 = 0$

10.3 Test: Fixed Effects vs. Random Effects?

- **Hausman test**

$$(\hat{\beta}^{FE} - \hat{\beta}^{RE})^T \left[Var(\hat{\beta}^{FE}) - Var(\hat{\beta}^{RE})\right]^{-1} (\hat{\beta}^{FE} - \hat{\beta}^{RE}) \sim \chi^2_{p-1}$$

- **Breusch-Pagan test (LM test)**

$$H_0 : \sigma_u^2 = 0, \quad H_1 : \sigma_u^2 \neq 0$$

10.4 Literatúra

Kapitoly 9 a 13 v [Gre03], knižnica `plm` [CM+08] a `lme4` [BMB+14]

Technickejšia literatúra [Are03], menej technickejšia tu [Woo15]. Veľmi prístupný úvod k zmiešaným modelom [Win13] a panelovým dátam [TR]. Videá na youtube od Ben Lamberta.

11 Náhodné lesy

"Pozdravujem vás, lesy, hory, z tej duše pozdravujem vás!"

Pavol Országh-Hviezdoslav

P.O.H. v lese.

11.1 Regresné stromy

Predstavte si, že prídete k lekárovi a ten sa vás začne vypytovať všelijaké otázky:

- Máte diastolický krvný tlak vyšší ako 100?
- Máte v rodine niekoho kto mal problémy so srdcom?
- Športujete viac ako 60 minút týždenne?
- Máte viac ako 30 rokov?

pomocou týchto ako aj iných **áno/nie otázok** vás zaradí do akejsi rizikovej skupiny a môže navrhnúť ďalší test alebo adekvátnu liečbu.

Na začiatku Vám dá tie **najdôležitejšie** otázky, a **podľa toho ako odpovedáte** Vám dáva stále podrobnejšie a podrobnejšie otázky tak, aby čo najlepšie rozlíšil tých rizikových pacientov od menej rizikových.

Teda lekár má akýsi receptár (v ang. guidelines), ktorým sa riadi. Jediné čo sa od Vás dozvie je zopár áno/nie odpovedí. Je to skvelé lebo je to ohromne jednoduché.[6]

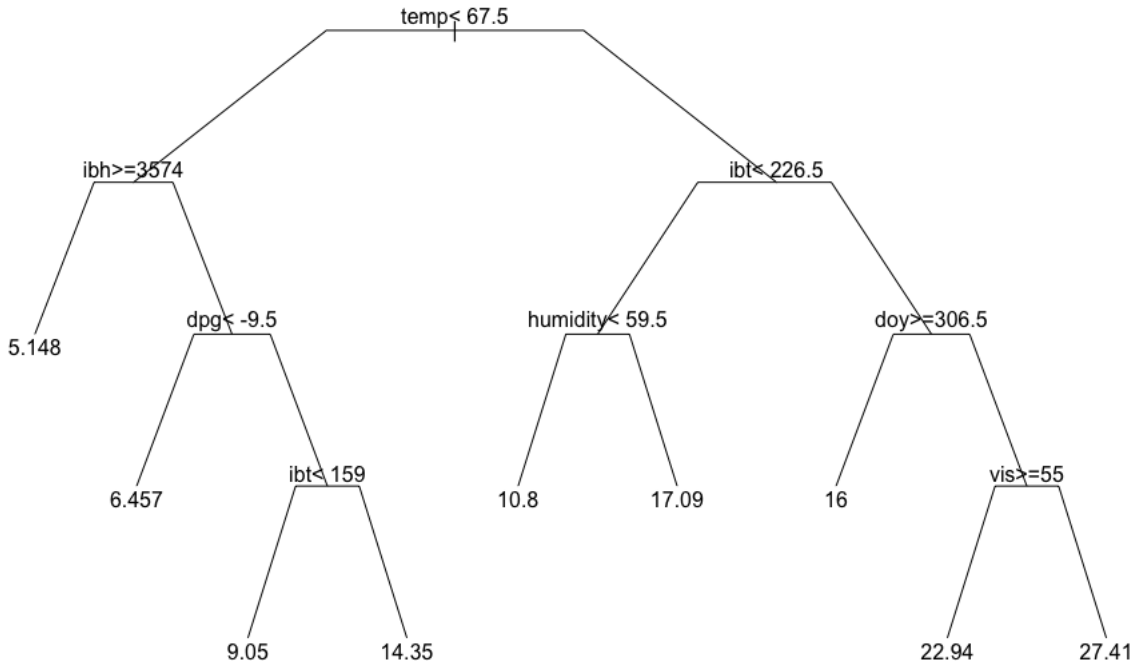
Zaujímavou otázkou je, či vieme nájsť nejaký systematický spôsob, ako majú byť tieto otázky kladené. Teda ako si vybrať premennú na základe ktorej rozlišujeme pacientov (krvný tlak, problémy so srdcom, šport, vek) a v prípade, že ide o numerickú premennú, tak ako nájsť hodnotu (100, 60, 30), ktorá nám pomôže čo najlepšie rozlišovať medzi pacientami. Ako v každej situácii v štatistike, keď používame slovo "najlepšie"potrebujeme mať nejaký spôsob ako túto kvalitu rozlišovania kvantifikovať.

Predstavme si teraz, že chceme predikovať nejakú spojitú premennú Y, a to len na základe takýchto áno/nie otázok týkajúcich sa regresorov $X = (X_1, X_2, ..., X_p)^T$. Našou úlohou je nájsť nejaké hodnoty $k \in \{1, ..., p\}$ a skalárnu veličinu $c \in \mathbb{R}$ tak, aby sme naše dáta rozdelili na 2 časti a aby nám miera neistoty čo najviac klesla. Teda aby Y pre $X_k \geq c$ (časť 1) bolo výrazne rôzne ako Y pre $X_k < c$ (časť 2). Prirodzenou mierou je porovnať pôvodné RSS a potom $RSS(part_1) + RSS(part_2)$. Teda chceme vybrať k a c tak, aby $RSS(part_1) + RSS(part_2)$ bolo čo najmenšie. V oboch týchto častiach budeme predikovať priemer Y. Sú aj iné múdre spôsoby posúdenie kvality rozdelenia do dvoch častí ale začať s týmto znie fajn.

Teraz sme si rozdelili naše dáta na dve časti pomocou jednej áno/nie otázky. Ak sú všetky naše premenné usporiadané to sme museli vyskúšať nie viac ako $p(n-1)$ kombinácií (zamyslite sa prečo) a my sme vybrali tú najlepšiu. Dobrý začiatok. Teraz môžeme pokračovať takýmto spôsobom a dve podčasti deliť na ďalšie časti. Toto sa dá elegantne znázorniť ako postupné vetvenie, preto ten názov - strom.

Takto si môžeme dáta rozdeliť na n rôznych častí. Toto však nie je príliš užitočné, takýto predikčný model by dokonale predikoval v rámci našej dátovej vzorky ale veľmi zle mimo nej (malý bias ale vysoká variancia). Takže sa treba zamyslieť, kedy vetvenie zastavíme.

[6] Toto bulvárne zjednodušenie je len didaktickou pomôckou.



Obr. 17: Predikcia koncentrácie atmosférického ozónu v Los Angeles v roku 1979 na základe rôznych prediktorov. Prebraté z [Far14].

Ako vetvenie zastavíme? No môžeme si zadefinovať, že požadujeme zlepšenie RSS aspoň o ϵ, tu ale nie je jasné, akú zmysluplnú hodnotu by sme si mali zvoliť. Strom môžeme hľadať taký, ktorý dobre predikuje, takže ktorý nám dá najmenšiu kros-validačnú chybu. Môžeme použiť tradičné leave-one-out alebo k-fold kros-validáciu. K-fold kros-validácia je úspornejšia ale rozdelenie na k častí je náhodné, preto náš strom závisí od náhodného seedu.

Iným spôsobom je nechať nariasť poriadne veľký strom a potom mu usekávať postupne vetvičky - toto sa nazýva *pruning*. Budeme minimalizovať

$$CC(strom) = \sum_{listy\ i} RSS_i + \lambda(\#\ listov),$$

kde parameter λ je penalta za komplexitu stromu, podobne ako pri model selection technicke AIC. Zo stromu veľkosti n vieme urobiť strom veľkosti $n-1$ tak, že spojíme susedné hrče (*nodes* - pozor listy sú špeciálne *nodes*, a to *terminal nodes*), tak aby sme stratili čo najmenej informácie, čo je prirodzené. Teraz máme receptár zmenšovania stromu tak môžeme použiť kros-validáciu, ale už len na λ, teda vyberieme tento parameter tak, aby náš strom mal čo najmenšiu kros-validačnú chybu.

Ohromne dôležitá výhoda stromových modelov je ich jednoduchosť a názornosť. Nevýhodou je, že nemáme v dispozícii confidenčné intervaly.

11.2 Náhodný les - mnoho náhodných regresných stromov

Regresné stromy však majú aj nejaké nevýhody. Binárna (áno/nie) nátura stromov spôsobuje, že nebudú predikovať príliš dobre, a to najmä ak sa pozorovania nachádzajú tesne pri hranici, ktorá rozdeľuje strom na vetvy. Takže trošku zmenené dáta nám môžu veľmi zmeniť náš regresný strom, čo je rozhodne nežiadúca vlastnosť. Alebo ak by sme náhodne rozdelili dáta na 2 polovice, tak z každej polovice dát nám môže vzniknúť výrazne iný strom.

Kebyže však máme veľmi veľa stromov, každý je trochu nerobustný ale potom tieto stromy spriemerujeme, dostaneme akýsi priemerný strom, ktorého variancia bude už výrazne nižšia.

Otázkou je, ako z jedného datasetu získať mnoho stromov. Prirodzeným nápadom je použiť bootstrap.

Takže

- Vytiahneme si z nášho datasetu vzorku rovnakej veľkosti s opakovaním
- Na tejto vzorke urobíme strom
- Väčšinou zhruba tretinu dáta nemáme v tejto vzorke, takže túto môžeme použiť na overenie toho, ako dobre náš strom predikuje.

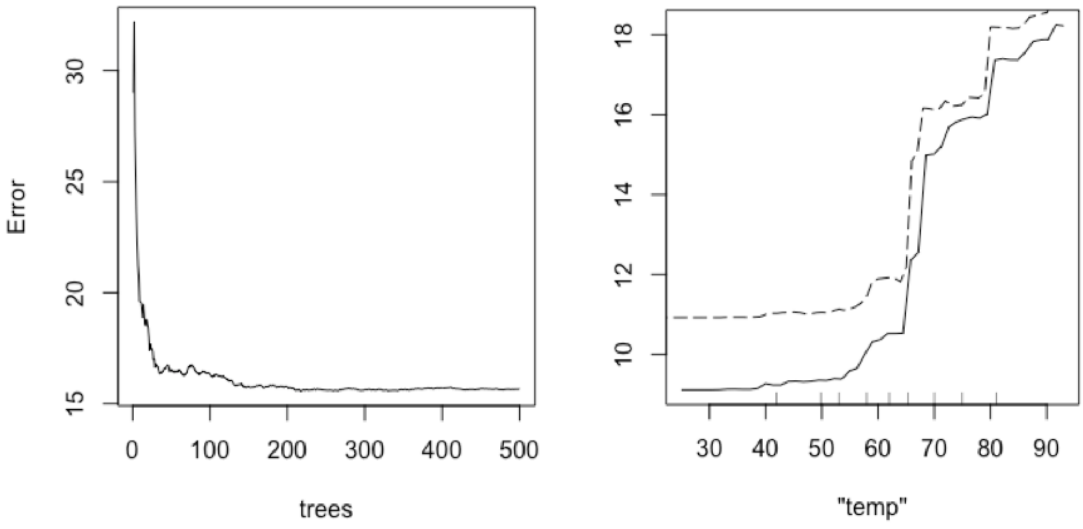
Teraz máme mnoho stromov.

Tieto stromy sú však dosť podobné a presne to nechceme. My potrebujeme, aby naše stromy boli rôzne, lebo potrebujeme variabilitu, ktorú nám rovnaké stromy neposkytnú. Znáhodníme to tak, že dovolíme vyberať len niektoré prediktory a to ktoré. tak to bude náhodné. Týmto spôsobom v nejakých stromoch už nie sú tie najdôležitejšie ale tým pádom aj tie menej prediktívne regresory môžu pomôcť. Pravidlom palce je náhodne vybrať zhruba $\sqrt{p}$ regresorov. Trochu menej, trochu viacej, vyberieme to tak, aby predikcia na dátach nepoužitých na konštrukciu stromu bola čo najlepšia.

Teraz keď potrebujeme predikovať pre novú hodnotu regresorov, tak každý strom nám dá nejakú predikciu a urobíme priemer.

Aký veľký les je dosť veľký? Taký pre ktorý predikčná chyba už neklesá.

Nevieme však popísať efekt jednej premennej tak pekne ako v prípade lineárnej regresie. Môžeme urobiť partial dependence metódu od Davida Freidmana, takže fixneme hodnotu prediktora na nejakú hodnotu a nanútime ju každému nášmu pozorovaniu. Pre každé takéto pozorovanie vypočítame predikovanú hodnou odozvy a spriemerujeme. Alebo môžeme zafixovať ostatné hodnoty trebárs na priemernej hodnote.



Vľavo: predikčná chyba v závislosti od veľkosti lesa. Vpravo partial dependence metóda (plná čiara) a metóda, kde sú ostatné premenné zafixované na priemerných hodnotách (prerušovaná čiara). Prebraté z [Far14].

Ako zistiť ktoré premenné sú dôležité. Na to máme mieru, ktorá sa nazýva *importance*. Zobereme jednu premennú a náhodne ju spermutujeme. Touto mierou bude rozdiel v MSE spermutovaného da-

tasetu mínus pôvodného. Toto použijeme len na '*out-of-bag* dáta, takže na tie, ktoré neboli použité pri konštrukcii stromov.

11.3 Klasifikačné stromy a Klasifikácia pomocou náhodného lesa

Fungujú veľmi podobne ako regresia. Len teraz namiesto RSS použijeme iné kritérium. Napríklad devianciu, entropiu alebo najčastejšie sa používa GIni index.

Stromy sú výhodné v tom, že sa v rámci nich prirodzene pracuje s chýbajúcimi pozorovaniami. Ak sa v rámci stromu nevieme rozhodnúť lebo informácia nám chýba, vybereme si väčšiu skupinu a hotovo. Chýbajúce pozorovania môžeme tiež dopĺňať a pozerať ako dobre sú klasifikované v lese ale iné triky.

Stromy boli veľmi úspešné napríklad vo financiách, spamových filtroch, medicíne. Oproti neurónovým sieťam sú napríklad oveľa menej náročné na výpočtovú silu.

11.4 Literatúra

Okrem tradičného [Far14] odporúčam [HTFF05], [JWHT13] a vynikajúce videá of Josha Starmera na Youtube (StatQuest).

A Čriepky z elementárnej pravdepodobnosti a štatistiky

A.1 Spojite rozdelený náhodný vektor

Náhodný vektor je kolekcia náhodných premenných, teda súbor $\mathcal{F}$-merateľných funkcií na pravdepodobnostnom priestore $(\Omega, \mathcal{F}, P)$

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \tag{A.1}$$

Pre spojite rozdelené X vieme pravdepodobnostné správanie úplne popísať **distribučnou funkciou** F_X alebo funkciou hustoty $f_X(x_1, \ldots, x_n)$. Platí medzi nimi nasledujúci vzťah.

$$F_X(t_1, \ldots, t_n) = P(X_1 \leq t_1, \ldots, X_n \leq t_n) = \int_{-\infty}^{t_1} \cdots \int_{-\infty}^{t_n} f_X(z_1, \ldots, z_n)\, dz_1 \ldots dz_n$$

Špeciálnym prípadom náhodného vektora je keď sú jeho jednotlivé časti **nezávislé**. V tomto prípade

$$f_X(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n).$$

Stredná hodnota náhodného vektora je definovaná nasledovne

$$E(X) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_n) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} = \mu \tag{A.2}$$

ide teda o vektor rovnakej dimenzie ako samotná náhodná premenná X, poznamenajme, že $E(X)$ je vektor čísiel nie náhodných premenných. Jeho jednotlivé komponenty dostaneme pomocou funkcie husoty f_X nasledovným spôsobom

$$\mu_i = E(X_i) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} z_i f_X(z_1, \ldots, z_n)\, dz_1 \ldots dz_n.$$

Pokiaľ chceme úplne popísať pravdepodobnostné správanie len jedného komponentu X_i, nasledovným spôsobom získame f_{X_i} z f_X, tejto funkcii hustoty hovoríme **marginálna**, zatiaľčo pôvodná f_X je **združená** hustota.

$$f_{X_i}(x_i) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} z_i f_X(z_1, \ldots, z_{i-1}, x_i, z_{i+1} \ldots, z_n)\, dz_1 \ldots dz_{i-1} dz_{i+1} \ldots dz_n.$$

Poznamenajme, že $\mu_i = \int_{-\infty}^{\infty} z_i f_{X_i}(z_i) dz_i$
Variancia jednotlivých komponentov X_i je

$$\sigma_i^2 = \int_{-\infty}^{t_1} \cdots \int_{-\infty}^{t_n} (z_i - \mu_i)^2 f_X(z_1, \ldots, z_n)\, dz_1 \ldots dz_n$$

a kovariancia

$$\sigma_{ij}^2 = Cov(X_i, X_j) = \int_{-\infty}^{t_1} \cdots \int_{-\infty}^{t_n} (z_i - \mu_i)(z_j - \mu_j) f_X(z_1, \ldots, z_n)\, dz_1 \ldots dz_n$$

Všetky variancie aj kovariancie sa dajú popísať úspornejšie pomocou kovariančnej matice

$$\Sigma = E\left[(X - \mu)(X - \mu)^T\right] = E\left[XX^T\right] - \mu\mu^T = \begin{bmatrix} \cdots & \cdots & \cdots \\ \vdots & \sigma_{ij}^2 & \vdots \\ \cdots & \cdots & \cdots, \end{bmatrix}$$

teda jedná sa o maticu **čísiel** a nie náhodných premenných.
Pre lineárnu transformáciu náhodného vektora $Y = AX$ platí

$$\begin{aligned} E(Y) &= E(AX) = AE(X) = A\mu, \\ Var(Y) &= E\left[(Y - E(Y))(Y - E(Y))^T\right] = E\left[(AX - A\mu)(AX - A\mu)^T\right] \\ &= E\left[A(X - \mu)(X - \mu)^T A^T\right] = A\, E\left[(X - \mu)(X - \mu)^T\right] A^T = AVar(X)A^T = A\Sigma A^T, \end{aligned} \tag{A.3}$$

kde v jednom z krokov sme využili maticovú identitu $(AX)^T = X^T A^T$.

Pripomeňme, že pre normálne náhodne rozdenený náhodný vektor X je jeho lineárna transformácia tiež z normálneho rozdelenia, preto platí $X \sim N(\mu, \Sigma) \implies Y = AX \sim N(A\mu, A\Sigma A^T)$.

Viacrozmerné normálne rozdelenie s parametrami (μ, Σ) má nasledovnú funkciu hustoty

$$f_X(x_1, \ldots, x_n) = \frac{1}{\sqrt{(2\pi)^k \det \Sigma}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right),$$

kde x označuje stĺpcový vektor.

A.2 Niektoré dôležité pravdepodobnostné distribúcie

Okrem normálho rozdelenia sa pri štatistickom testovaní hypotéz častokrát objavujú rozdelenia, ktoré sú odvodené od normálneho. Tieto rozdelenia majú komplikované funkcie hustoty.

Nech $\{X_i\}_{i=1}^k$ sú i.i.d a nech $X_i \sim N(0,1)$. Potom $Y \equiv X_1^2 + \ldots X_k^2$ má rozdelenie chí-kvadrát s k stupňami voľnosti, označujeme aj $Y \sim \chi_k^2$. $E(Y) = k$ a $Var(Y) = \sqrt{2k}$. S týmto rozdelením sa stretáme často najmä kvôli tomu, že normované štvorce reziduí pri lineárnom regresnom modeli s normálnymi chybami majú takéto rozdelenie. Stretneme sa s ním tiež pri teste pomerom vierohodností.

Nech $X \sim N(0,1)$ a nech $Y \sim \chi_k^2$, potom $Z \equiv \frac{X}{\sqrt{(Y/k)}}$ má t-rozdelenie (Studentovo) s k stupňami voľnosti, označujeme ako $Z \sim t_k$. Špeciálny prípad je t_1 rozdelenie, ktoré sa nazýva *Cauchyho* rozdelenie. Je známe tým, že má ťažké chvosty a stredná hodnota nie je dobre definovaná. Studentovo rozdelenie má podobne ako normálne rozdelenie zvonovitý tvar ale s ťažšími chvostami. Pre veľký počet stupňov voľnosti sa čoraz viac podobá na normálne, ktoré je jeho limitou. S t-rozdelením sa stretneme pri testovaní signifikantnosti jedného parametra β_i pri lineárnom regresnom modeli s normálnymi chybami. Podobne keď chceme zostrojiť konfidenčný interval pre neznámy parameter strednej hodnoty a nepoznáme smerodajnú odchýlku (ak poznáme, môžeme použiť aproximáciu normálnym rozdelením).

Nech $X \sim \chi_s^2$ a $Y \sim \chi_t^2$, potom $Z \equiv \frac{X/s}{Y/t}$ má F-rozdelenie s s a t stupňami voľnosti, označujeme $Z \sim F_{s,t}$. Dva chí-kvadráty dávame do pomeru sumy štvorcov pri testovaní platnosti menšieho modelu oproti väčšiemu (kde menší vznikol pomocou lineárnych reštrikcií) pri lineárnom regresnom modeli s normálnymi chybami.

A.3 Základné pojmy štatistického testovania hypotéz

Princíp frekventistického testovania hypotéz je nasledovný: určí sa **nulová** hypotéza, napríklad $\beta_i = 0$, to je čosi čo nás zaujíma. Ideme zistiť či sme schopný zamietnuť túto hypotézu v prospech ineh **alternatívnej hypotézy**, napríklad $\beta_i \neq 0$. Za predpokladu správnosti modelu **a** nulovej hypotézy vieme, že akási **testovacia štatistika** má nám známe rozdelenie. Z našich dát však máme len jednu jedinú hodnotu, jedinú realizáciu tejto štatistiky. Pozrieme sa na to, ako veľmi extrémna je to hodnota. Ak je veľmi, tak **zamietneme** nulovú hypotézu v prospech alternatívnej hypotézy. Ak nie je veľmi extrémna, tak **nezamietneme** nulovú hypotézu v prospech alternatívnej hypotézy. Pozor, to však neznamená, že nulová hypotéza je pravdivá. Pokojne môžeme mať len príliš malú dátovú vzorku. Za platnosti modelu a nulovej hypotézy sa pravdepodobnosť padnutia ešte extrémnejšej hodnoty ako našej realizovanej štatistiky nazýva **p-hodnota**. Ak je malá, tak to znamená, že naša štatistika je veľmi nepravdepodobná a zamietneme nulovú hypotézu v prospech alternatívnej. V praxi to funguje tak, že sa zvolí akési malé čislo α, ktoré nám hovorí ako veľmi budeme nesprávne zamietať nulovú hypotézu aj keď bude platná, toto sa volá **chyba prvého druhu** alebo **hladina významnosti**. Štandardne sa volí ako 5% ale toto je len konvencia a niet žiadneho iného dôvodu prečo nezobrať inú hodnotu. Teda ak je p-hodnota menšia ako α zamietnem nulovú hypotézu, inak nezamietnem. Rozhodovacie pravidlo, teda funkcia, ktorá dostane dáta a odpovie zamietni/nezamietni, sa nazýva **test**.

Existuje aj iná kvalita testu ako veľkosť chyby prvého druhu, a to je napríklad ako dobre vie môj test, teda rozhodovacie pravidlo, zamietnuť nulovú hypotézu, keď nie je pravdivá. Pravdepodobnosť správneho zamietnutia sa nazýva *sila* testu a jeden mínus sila testu sa nazýva **chyba druhého druhu**. Samozrejme by sme chceli aby sila testu bola 1. Niektoré testy sú optimálne v zmysle, že pre fixnú hladinu významnosti α minimalizujú chybu druhého druhu.

Podobne ako $\beta_j = 0$ vieme testovať aj $\beta_j = b$. Množina všetkých možných čisiel b, ktoré by náš test nezamietol sa nazýva **interval spoľahlivosti** (confidence interval) pre neznámy parameter. Častokrát sa preto pozeráme, či obsahuje alebo neobsahuje daný interval spoľahlivosti nulu. Pri fixnej hladine významnosti významnosti sa nazýva $100(1-\alpha)\%$-ný interval spoľahlivosti. Interval spoľahlivosti je náhodný interval, pretože je skonštruoovaný z dát, ktoré sú náhodné. Ak by som vygeneroval veľké množstvo dátových vzoriek a pre každú dátovú vzorku vypočítal interval spoľahlivosti, potom by, za predpokladu správnosti modelu a nulovej hypotézy, $100(1-\alpha)\%$ z nich pokrývalo skutočný parameter. Toto je jediná správna pravdepodobnostná interpretácia. Nie je príliš uspokojujúca, pretože my máme v dispozícii len jednu dátovú vzorku a len jednu testovaciu štatistiku. Interpretácia: "S pravdepodobnosťou $100(1-\alpha)\%$ sa neznámy parameter nachádza v nami vypočítanom intervale spoľahlivosti" je zavádzajúca a nesprávna. Neznámy parameter je fixné číslo a interval spoľahlivosti je realizácia náhodných dát a nie naopak.

Príklad - test pre strednú hodnotu so známou varianciou:

- Model: $\{X\}_{i=1}^n$ sú iid a $\forall i : E(X_i) = \mu, Var(X_i) = 1$.
- Objekt nášho záujmu: $\mu \in \mathbf{R}$
- Testovacia štatistika: $\sqrt{n}\left(\frac{X_1+\cdots+X_n}{n} - \mu\right)$ má za predpokladu platnosti modelu podľa centrálnej limitnej vety asymptoticky normálne rozdelenie $N(0,1)$.
- Nulová hypotéza: $H_0 : \mu = 0$
- Alternatívna hypotéza: $H_1 : \mu \neq 0$
- Hladina významnosti: $\alpha = 0.05$
- Kritická hodnota testovacej štatistiky: $z_{1-\alpha/2} = 1.96$
- Test: Ak $|\sqrt{n}\frac{X_1+\cdots+X_n}{n} - 0| > 1.96 = z_{1-\alpha/2}$ zamietni nulovú hypotézu, ak $|\sqrt{n}\frac{X_1+\cdots+X_n}{n} - 0| \leq 1.96$ nezamietni nulovú hypotézu.
- Dátová vzorka: $\{1, -3, -2, 1, 5, -6, 4, 2\}$
- Realizácia testovacej štatistiky: $\sqrt{n}\frac{X_1+\cdots+X_n}{n} = 0.5657 \leq 1.96$ preto nezamietame H_0.
- P-hodnota: $\Phi(0.5657) = 0.714 \geq 0.05$ preto nezamietame H_0.
- Interval spoľahlivosti pre neznámy parameter μ: $CI = (\frac{X_1+\cdots+X_n}{n} - \frac{\sigma}{\sqrt{n}}, \frac{X_1+\cdots+X_n}{n} + \frac{\sigma}{\sqrt{n}}) = (0.1536, 0.554)$. $0 \in CI$, preto nezamietame nulovú hypotézu.

A.4 Metóda maximálnej vierohodnosti (Maximum Likelihood)

Funkcia vierohodnosti, alebo likelihood funkcia, je pravdepodobnosť dátovej vzorky. V prípade i.i.d. pozorvaní náhodnej premennej s hustotou $f(y_i|\beta)$ je to

$$L(\beta) = \prod_{i=1}^{n} f(y_i|\beta).$$

Na likelihood L sa pozeráme ako na **funkciu parametra pri fixnej dátovej vzorke**.

Odhad metódou maximálnej vierohodnosti, je $\hat{\beta}_{ML} = \arg\max_\beta L(\beta)$. Častokrát je numericky výhodnejšie pracovať s logaritmom, pretože pri väčšej dátovej vzorke násobíme veľmi malé čísla avšak kedže logaritmus je monotónna transformácia $\hat{\beta}_{ML} = \arg\max_\beta \log L(\beta)$. V niektorých situáciách máme analytický predpis pre $\hat{\beta}_{ML}$ ako napríklad pre klasický lineárny regresný model s normálnymi chybami, avšak väčšinou nie a preto si musíme pomôcť optimalizačným softvérom. Pokiaľ je parametrov veľa, toto môže byť veľmi náročný problém sám o sebe.

Predpokladajme, že existuje jediný skutočný parameter β_0, ktorý vygeneroval dáta. Odhad ML je **konzistentný**, takže

$$\hat{\beta}_{ML} \to_P \beta_0,$$

teda konverguje podľa pravdepodobnosti ku skutočnej hodnote β_0, $\forall \epsilon > 0 : P(|\hat{\beta}^n_{ML} - \beta_0| < \epsilon) \to 0$ pre $n \to \infty$ (n je veľkosť dátovej vzorky). Konzistencia je prirodzenou požiadavkou na odhadcu, bez konzistencie sa ďaleko nedostaneme. Teda pre veľkú dátovú vzorku n, odhadca nám bude dávať hodnoty čoraz bližšie a bližsie ku skutočnému θ_0.

Variancia $\hat{\beta}_{ML}$ klesá priamo úmerne n a $\sqrt{n}(\hat{\beta}_{ML} - \beta_0) \to_D N(0, V(\beta_0))$, teda ML odhadca sa pre veľké n podobná normálnemu rozdeleniu, kde $V(\beta)$ je funkciou prvej a druhej derivácie likelihood funkcie.

Prvá derivácia log-likelihoodu podľa parametra sa nazýva **score** funkcia a tu bude označená ako $u(\beta)$

$$u(\beta) = \frac{\partial \log L(\beta)}{\partial \beta},$$

teda ide o riadkový vektor dĺžky rovnakej ako β. Nutnou podmienkou pre optimum $\hat{\beta}_{ML}$ je $u(\beta) = 0$. O tom aký presný je odhad máme informáciu z druhej derivácie log-likelihoodu, ak je $\log L(\beta)$ veľmi ohnutá v optime, znamená to, že okolité body majú oveľa menšiu vierohodnosť. Súhrnná informácia o zahnutosti log-likelihoodu okolo optimálnej hodnoty sa volá **Fisherova informácia** alebo **Fisherova informačná matica**, je definovaná nasledovne

$$I(\beta) = var(u(\beta)) = E\left(\frac{\partial u(\beta)}{\partial \beta}\frac{\partial u(\beta)}{\beta}^T\right)$$

a v prípade korektnej špecifikácie sa matica druhých derivácií log-likelihoodu $H(\beta)$ rovná

$$H(\beta_0) = E\left(\frac{\partial^2 \log L(\beta_0)}{\partial \beta \partial \beta^T}\right) = -E\left(\frac{\partial u(\beta_0)}{\partial \beta}\frac{\partial u(\beta_0)}{\beta}^T\right)^{-1} = -I(\beta_0)^{-1}$$

Dá sa ukázať, že ak je **model korektne špecifikovaný**, tak variancia β_{ML} sa dá rozumne odhadnúť nasledovne

$$var(\hat{\beta}_{ML}) = I^{-1}(\hat{\beta}_{ML}),$$

niekedy však namiesto očakávanej hodnoty druhých derivácií dosadíme priamo $\frac{\partial^2 \log L(\hat{\beta})}{\partial \beta \partial \beta^T}$. Teda $V(\hat{\beta}) = I^{-1}(\hat{\beta}_{ML})$ a $\sqrt{n}(\hat{\beta}_{ML} - \beta_0) \to_D N(0, I^{-1}(\beta_0))$.

Ak však **model nie je korektne špecifikovaný**, potom varianciu vieme odhadnúť nasledovne

$$var(\hat{\beta}_{ML}) = H(\hat{\beta}_{ML})^{-1} I(\hat{\beta}_{ML}) H(\hat{\beta}_{ML})^{-1},$$

a

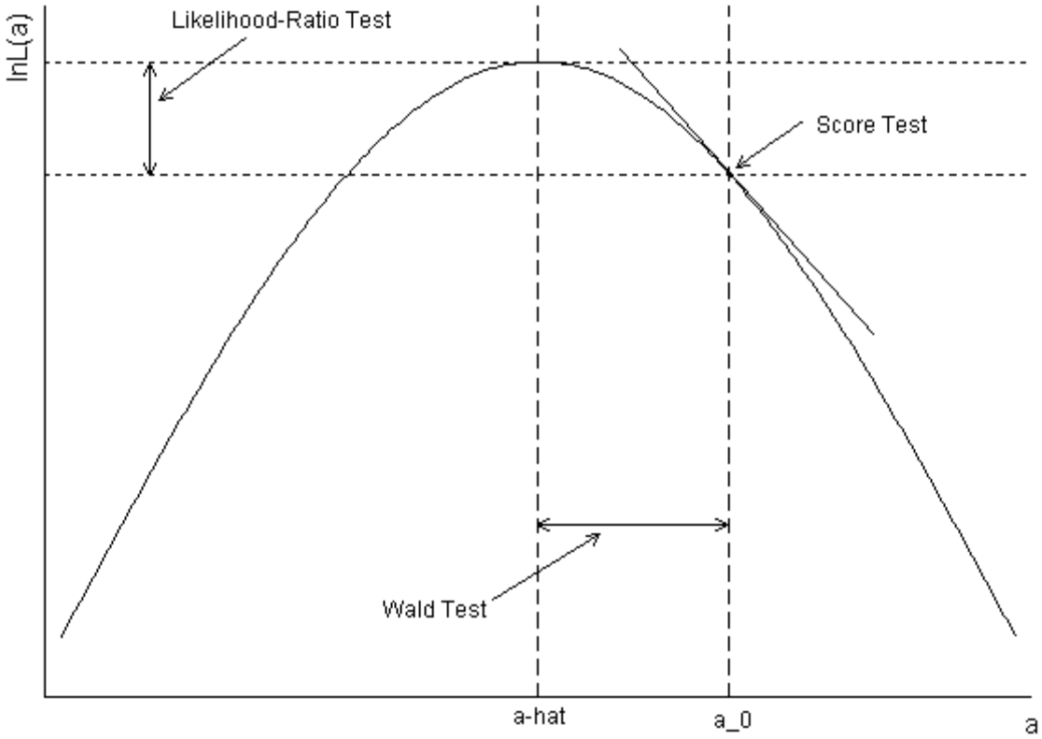
$$\sqrt{n}(\hat{\beta}_{ML} - \beta_*) \to_D N\left(0, H(\beta_*)^{-1} I(\beta_0) H(\beta_*)^{-1}\right),$$

kde β_* je minimizátor KL divergencie (viď časť (A.4.1)).

Na základe likelihood funkcie máme **tri typy testov** na porovnávanie dvoch vnorených modelov, každý je asymptoticky rozdelený ako χ^2. Majme dva modely: veľký model s l parametrami a likelihoodom L_{large} a malý model s s parametrami, ktorý je špeciálna verzia veľkého modelu za predpokladu lineárnych reštrikcií na parametre. Pozor χ^2 aproximácia nefunguje ak sa parameter nachádza na hranici priestoru parametrov (napr. $\sigma^2 = 0$, pretože $\sigma^2 \in [0, \infty)$).

- **Likelihood ratio test** - testovacia štatistika vyzerá nasledovne $2 \log \frac{L_{large}}{L_{small}} \sim \chi^2_{l-s}$
- **Waldov test** - testuje $H_0 : \beta = \beta_0$ a testovacia štatistika vyzerá nasledovne $(\hat{\beta}_{ML} - \beta_0)^T I(\hat{\beta}_{ML})(\hat{\beta}_{ML} - \beta_0) \sim \chi^2_{l-s}$
- **Score test** - testuje $H_0 : \beta = \beta_0$ a testovacia štatistika vyzerá nasledovne $u(\beta_0)^T I^{-1}(\beta_0) u(\beta_0) \sim \chi^2_{l-s}$

LR test potrebuje dve optimalizácie, Waldov test jednu a score test žiadnu, takže LR môže byť numericky náročný. Pokiaľ sa nám dá, odporúča sa používať LR test. Tieto testy sú graficky zobrazené na Obr. 18.



Obr. 18: Porovnanie testov založených na likelihoode, zdroj: [Fox97].

A.4.1 Maximum Likelihood ako minimizátor KL divergencie

Vieme, že

$$\hat{\beta}_{ML} = \arg\max_\beta \prod_{i=1}^n f(y_i|\beta) = \arg\max_\beta \sum_{i=1}^n \log f(y_i|\beta) = \arg\min_\beta \frac{1}{n} \sum_{i=1}^n -\log f(y_i|\beta)$$

naviac vieme, že podľa silného zákona o veľkých číslach platí

$$\frac{1}{n} \sum_{i=1}^n -\log f(Y_i|\beta) \to_{a.s.} E\left(-\log f(Y|\beta)\right). \tag{A.4}$$

Nech β_0 je skutočný parameter (ktorý nepoznáme a snažíme sa odhadnúť), teda nech $Y_i \sim f(\cdot|\beta_0)$, potom

$$E\left(-\log f(Y|\beta)\right) = \int -\log\left(f(y|\beta)\right) f(y|\beta_0)\, ds$$

a zároveň

$$E\left(\log f(Y|\beta_0) - \log f(Y|\beta)\right) = \int \log\left(\frac{f(y|\beta_0)}{f(y|\beta)}\right) f(y|\beta_0)\, ds \equiv KL\left(f(\cdot|\beta), f(\cdot|\beta_0)\right) \geq 0.$$

Funkcia $KL(\cdot,\cdot)$ sa nazýva **Kuhlback-Leiblerova divergencia** a platí $KL(f,g) = 0 \iff f = g$.

Z rovnice (A.4) vidíme, že

$$\hat{\beta}_{ML} \to_{a.s.} \arg\min_\beta KL\left(f(\cdot|\beta_0), f(\cdot|\beta)\right),$$

Obr. 19: Oblak ako cválajúci kôň.

nakoľko $E\left(\log f(Y|\beta_0)\right)$ nezávisí od β.

Funkcia $KL(f,g)$ má zaujímavú **interpretáciu**. Tu prezentujeme tú pochádzajúcu od informatikov. Nech P je diskrétna pravdepodobnostná distribúcia z ktorej nám padajú nejaké symboly x z množiny $\mathcal{X}$. Predstavme si, že každý takýto symbol vieme zakódovať do binárneho stringu (napríklad 1001001011) a nech L je jeho dĺžka. Existuje bijekcia medzi L a $\log_2 Q$, kde Q je pravdepodobnostná distribúcia symbolov. Ak sú symboly samplované z P, potom je očakávaná dĺžka kódu $EL = \sum_{x\in\mathcal{X}} P(x)L(x)$. Ďalej ak x pochádza z P, potom akýkoľvek spôsob zakódovania symbolov x musí platiť $EL \geq -\sum_{x\in\mathcal{X}} P(x)\log_2 P(x) \equiv H(P)$, kde funkciu $H(P)$ nazývame **entropia** alebo aj množstvo informácie. Toto znamená, že $L(x) = -\log_2(x)$ je najlepší možný spôsob zakódovania $\mathcal{X}$ v zmysle priemernej dĺžky kódu.

Predstavme si teraz hypotetickú situáciu, že si nesprávne myslíme, že dáta (teda x-ká) pochádzajú (sú samplované) z Q. Teraz nakoľko si myslím, že Q je správna distribúcia zakódujem to tak, že dĺžka binárneho stringu kódujúceho x bude $L(x) = \log_2 Q(x)$ nuž a skutočná očakávaná dĺžka stringu bude potom $E_P\left(\log Q(x)\right)$. Nuž ale vieme, že ak by sme vedeli, že dáta sú z P, vedeli by sme to zakódovať lepšie a to tak, že $E_P\left(\log P(x)\right) = H(P)$ by bola očakávaná dĺžka kódu. Preto Kuhlback-Leiblerovu divergenciu

$$KL(P,Q) = E_P\left(\log\frac{Q(X)}{P(X)}\right),$$

interpretujeme ako **o koľko dlhší bude môj priemerný kód ak nesprávne považujem Q za data-generujúci proces, keď je to v skutočnosti P.** Viac nájdete napr. tu [Yu08] alebo v jednom z najslávnejších vedeckých článkov všetkých čias [Sha48].

A.5 Výber modelu - Informačné kritériá

Ak sme v situácii, že chceme predikovať alebo vysvetľovať nejakú premennú p regresormi, máme v dispozícii 2^p modelov, každý z p regresorov môžeme totiž pridať alebo nepridať do modelu. A to ešte vôbec neberieme do úvahy možné transformácie premenných, členy vyšších rádov alebo interakcie! Kedže 2^p je zvyčajne veľa, máme problém. Stratégií čo robiť je viacero.

Kroková eliminácia regresorov. Môžeme napríklad fitnúť najväčší model s p parametrami a potom postupne odoberať "najhoršie"regresory. Najhoršie napríklad v zmysle najväčšej p-hodnoty. Zastaviť môžeme napríklad vtedy, keď už sú všetky regresory signifikantné na našej dopredu zvolenej tolerovanej chyby prvého druhu (napríklad 5%). Tento pristup trpí jedným veľkým problémom. David Freedman [Fre83] toto ilustroval na príklade, kde vysvetľoval šum ďalšími 50 šumami pri dátovej vzorke veľkosti 100. Postupnou (iba dvojkrokovou) elimináciou nesignifikantných parametrov dostal viacero regresorov, ktoré boli signifikantné. Ak niekto len slepo robí krokovú elimináciu môže nájsť vzťah medzi Y a X aj tam, kde žiaden nie je! Pri malej dátovej vzorke a veľa parametroch bude nejaký z regresorov signifikatný náhodou, podobne ako oblak môže vyzerať ako cválajúci kôň (Obr. 19). Rovnaký problém nastane aj pri väčšej dátovej vzorke a menšom počte parametrov, nie je však až taký alarmujúci. Podobne ako môžeme eliminovať regresory môžeme aj postupne pridávať regresory k chudobnému modelu.

Informačné kritériá. Prečo preferujeme malé modely oproti veľkým modelom? Veľké modely fitujú dobre dáta ale zle predikujú (nízky bias a veľká variancia) a malé modely sú príliš chudobné na to aby dostatočne vysvetlili premennú, ktorá nás zaujíma (vysoký bias a malá variancia). Chceli by sme akýsi kompromis medzi týmito dvoma extrémami. Presne toto robia informačné kritériá. Majú formu

IC = -[vhodnosť fitu] + [penalta za veľkosť modelu] a preferujeme model s najnižšou IC.

Najznámejšie sú nasledovné

- Akkaikeho informačné kritérium: $AIC = -2\log L(\hat{\theta}_{ML}) + 2K$
- Bayesovské informačné kritérium: $BIC = -2\log L(\hat{\theta}_{ML}) + 2K\log n$

kde $\log L(\hat{\theta}_{ML})$ je maximum log-likelihood modelu, K je počet parametrov v modeli a n je veľkosť dátovej vzorky. AIC je vhodné keď nás zaujíma predikcia, tak AIC je **efektívny** v zmysle minimalizovania priemernej kvadratickej predikčnej chyby. Na druhej strane BIC, ktorý má väčšiu penaltu, je **konzistentný**, teda ak náš model obsahuje skutočnú (pre nás neznámu) dáta generujúcu distribúciu, tak potom ju BIC odhalí pri veľkej dátovej vzorke.

Náčrt teoretického dôvodenia v prospech AIC a BIC: AIC aj BIC sa minimalizujú priemernú Kuhlback-Leiblerovu vzdialenosť skutočnej dáta-generujúcej distribúcie g a modelu odhadnutého cez maximum likelihood $f(\cdot|\hat{\theta}_{ML})$, teda

$$E_g\left[KL(g, f(\cdot|\hat{\theta}_{ML}))\right] = \int g \log g \; dy - E_g\left[g \log f(y|\hat{\theta}_{ML})\right],$$

kde prvý člen nezávisí od θ a druhý sa dá odhadnúť z dát. Naivný estimátor druhého člena je $\frac{l(\hat{\theta}_{ML})}{n}$, teda maximum log-likelihood. Tento odhad je však vychýleným odhadom $E_g\left[g \log f(y|\hat{\theta}_{ML})\right]$. AIC a BIC sa líšia v tom akým spôsobom odhadnú tento bias.

Existuje ešte mnoho iných IC, každé má špecifické vlastnosti, ktoré ho robia atraktívnym pre danú situáciu.

Na rozdiel od krokovej eliminácie informačné kritériá vedia porovnávať aj modely, ktoré nie sú do seba vnorené (nested). Informačné kritériá sú detailne spracované v učebnici [CH$^+$08].

Fundamentálny problém.. Oba prístupy však trpia neriešiteľným problémom a ten je, že ak použijeme tie isté dáta na výber modelu a potom na odhadnutie parametrov dostane výsledky, ktoré budú prehnane optimistické (príliš dobrý fit). Sériu depresívnych článkov o tomto fenoméne napísali Hannes Leeb a Benedikt Pötscher [LP05], [LP06], [LP08]. Jednoduchým riešením je náhodne rozdeliť dáta na dve časti, na jednej vyberieme najlepší model v zmysle nejakého kritéria. Na druhej ho potom odhadneme. Toto sa však, zdá sa, v praxi nerobí toľko, koľko by sa malo, najmä však vtedy ak cieľom analýzy nie je predikcia.

Alternatívne skoro-riešenie: Ďalšou alternatívou je použiť metódu, ktorá odhaduje lineárny model tak, že rovno nastaví niektoré parametre na nulu ako napríklad LASSO. Pri lineárnom modeli neminimalizujeme len štvorce odchýlok $\sum(y_i - \hat{y}_i)^2$ ale máme k tomu penalizačný člen, ktorý má tvar $\gamma \sum |\beta_i|$. Je to, akoby sme mali šnúru a z nej môžeme nastrihať nejaké β_i-y. Najprv odstrihneme na tie najdôležitejšie a na tie menej dôležité nám už neostane špagát. Pre malé γ máme dlhú šnúru a pre veľké zasa malú. Táto myšlienka sa nazýva regularizácia sa dá použiť nielen pri klasickom lineárnom modeli. Problémom zostáva ako zvoliť penaltu γ, to môžeme napríklad cez krížovú validáciu.

Literatúra

[A+07] Alan Agresti et al. An introduction to categorical data analysis. hoboken, 2007.

[AD08] Jason Abrevaya and Christian M Dahl. The effects of birth inputs on birthweight: evidence from quantile estimation on panel data. *Journal of Business & Economic Statistics*, 26(4):379–397, 2008.

[Agr15] Alan Agresti. *Foundations of linear and generalized linear models*. John Wiley & Sons, 2015.

[AK11] Alan Agresti and Maria Kateri. *Categorical data analysis*. Springer, 2011.

[Are03] Manuel Arellano. Panel data econometrics. *Oxford University Press, OxfordArrelano M, Bond S (1991) Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. Rev Econ Stud*, 58(2):277297Balestra, 2003.

[Aus16] David Austin. We recommend a singular value decomposition, 2016. http://www.ams.org/samplings/feature-column/fcarc-svd.

[BD91] Rudolf J Beran and Gilles R Ducharme. *Asymptotic theory for bootstrap methods in statistics*. Centre de Recherches Mathematiques, 1991.

[BMB+14] Douglas Bates, Martin Maechler, Ben Bolker, Steven Walker, et al. lme4: Linear mixed-effects models using eigen and s4. *R package version*, 1(7), 2014.

[boo] Bootstrap example. http://www.r-bloggers.com/bootstrap-example/.

[CH+08] Gerda Claeskens, Nils Lid Hjort, et al. *Model selection and model averaging*, volume 330. Cambridge University Press Cambridge, 2008.

[CM+08] Yves Croissant, Giovanni Millo, et al. Panel data econometrics in r: The plm package. *Journal of Statistical Software*, 27(2):1–43, 2008.

[Cra12] Michael J Crawley. *The R book*. John Wiley & Sons, 2012.

[CT05] A Colin Cameron and Pravin K Trivedi. *Microeconometrics: methods and applications*. Cambridge university press, 2005.

[CT10] Adrian Colin Cameron and Pravin K Trivedi. *Microeconometrics using stata*, volume 2. Stata Press College Station, TX, 2010.

[DK02] AC Davison and Diego Kuonen. An introduction to the bootstrap with applications in r. *Statisticial Computing and Statistical Graphics Newsletter*, 13(1):6–11, 2002.

[DT01] John DiNardo and Justin L Tobias. Nonparametric density and regression estimation. *The Journal of Economic Perspectives*, 15(4):11–28, 2001.

[Eng57] Ernst Engel. Die produktions-und konsumptionsverhältnisse des königreichs sachsen. *Zeitschrift des Statistischen Bureaus des Königlich Sächsischen Ministeriums des Innern*, 8:1–54, 1857.

[ET94] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[Far05] Julian J Faraway. *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. CRC press, 2005.

[Far14] Julian J Faraway. *Linear models with R*. CRC Press, 2014.

[Fox97] John Fox. *Applied regression analysis, linear models, and related methods*. Sage Publications, Inc, 1997.

[Fre83] David A Freedman. A note on screening regression equations. *The American Statistician*, 37(2):152–155, 1983.

[glma] Assumptions of generalised linear model. http://stats.stackexchange.com/questions/32285/assumptions-of-generalised-linear-model.

[glmb] Checking (g)lm model assumptions in r. http://www.r-bloggers.com/checking-glm-model-assumptions-in-r/.

[Gre03] William H Greene. *Econometric analysis*. Pearson Education India, 2003.

[Han09] Bruce E Hansen. Lecture notes on nonparametrics. *Lecture notes*, 2009.

[HBG96] Rob J Hyndman, David M Bashtannyk, and Gary K Grunwald. Estimating and visualizing conditional densities. *Journal of Computational and Graphical Statistics*, 5(4):315–336, 1996.

[HT90] Trevor J Hastie and Robert J Tibshirani. *Generalized additive models*, volume 43. CRC Press, 1990.

[HTFF05] Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005.

[IS88] Alan J Izenman and Charles J Sommer. Philatelic mixtures and multimodal densities. *Journal of the American Statistical association*, 83(404):941–953, 1988.

[JWHT13] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[Ken01] Peter E Kennedy. Bootstrapping student understanding of what is going on in econometrics. *The Journal of Economic Education*, 32(2):110–123, 2001.

[Koe05] Roger Koenker. *Quantile regression*. Number 38. Cambridge university press, 2005.

[Lar14] Bret Larget. Chapter 3 r bootstrap examples, 2014. http://www.stat.wisc.edu/∼ larget/stat302/chap3.pdf.

[LF06] J Scott Long and Jeremy Freese. *Regression models for categorical dependent variables using Stata*. Stata press, 2006.

[LP] Lewis Lehe and Victor Powell. Simpson's paradox. http://vudlab.com/simpsons/.

[LP05] Hannes Leeb and Benedikt M Pötscher. Model selection and inference: Facts and fiction. *Econometric Theory*, 21(01):21–59, 2005.

[LP06] Hannes Leeb and Benedikt M Pötscher. Can one estimate the conditional distribution of post-model-selection estimators? *The Annals of Statistics*, pages 2554–2591, 2006.

[LP08] Hannes Leeb and Benedikt M Pötscher. Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory*, 24(02):338–376, 2008.

[MBM95] Willard G. Manning, Linda Blumberg, and Lawrence H. Moulton. The demand for alcohol: The differential response to price. *Journal of Health Economics*, 14(2):123 – 148, 1995.

[not] Video lectures on nonparametric regression - lectures 7,8,9,10. https://www.youtube.com/watch?v=YHnC1ddWUx0.

[PRW99] Dimitris N Politis, Joseph P Romano, and Michael Wolf. Subsampling springer series in statistics, 1999.

[Rod15] German Rodriguez. Lecture notes on generalized linear statistical models, 2015. http://data.princeton.edu/wws509/.

[Sha48] C.E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948. cited By 2654.

[SL87] Steven G Self and Kung-Yee Liang. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398):605–610, 1987.

[TR] Oscar Torres-Reyna. Getting started in fixed/random effects models using r. http://www.princeton.edu/∼ otorres/Panel101R.pdf.

[Uni16] The Pennsylvania State University. Lecture notes on analysis of discrete data, 2016. https://onlinecourses.science.psu.edu/stat504/.

[Win13] Bodo Winter. Linear models and linear mixed effects models in r with linguistic applications. *arXiv preprint arXiv:1308.5499*, 2013.

[Woo15] Jeffrey Wooldridge. *Introductory econometrics: A modern approach*. Nelson Education, 2015.

[Yu08] Bin Yu. Lecture notes on tutorial: Information theory and statistics, 2008. http://www.icml-conference.org/icml08/slides1.pdf.

[Zei06] Achim Zeileis. Object-oriented computation of sandwich estimators. 2006.

[ZKJ07] Achim Zeileis, Christian Kleiber, and Simon Jackman. Regression models for count data in r. 2007.