Draft poznámok k predmetu

# Moderná Aplikovaná regresia 1

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.

Lukáš Lafférs*

13. decembra 2021

## 1 Úvod

Na začiatku akejkoľvek dátovej analýzy by sme sa mali zamerať na dáta. Je užitočné konzultovať daný problém s odborníkom v danej oblasti. Predtým, ako sa pustíme do štatistického spracovania dát, sa pýtame (nielen) nasledovné otázky.

- Akým spôsobom boli dáta zbierané? Napr. telefonický prieskum, výstup meracieho prístroja, internetový prieskum. To, či ide o náhodnú vzorku alebo nie, má obrovský dopad na analýzu dát. Zlatým štandardom je experiment, avšak v mnohých prípadoch (najmä - ale nielen - v ekonomických aplikáciách) nie sú dáta náhodnou vzorkou. Existujú aj chýbajúce pozorovania? Aká je povaha dôvodu chýbania? Ľudia, ktorí odmietnu odpovedať na volebný prieskum, môžu byť výrazne iní ako tí ktorí odpovedajú a toto môže viesť ku skresleným výsledkom.
- Čo ktoré premenné znamenajú? V akých jednotkách sú premenné? Odpovede na tieto otázky potrebujeme poznať čo najpodrobnejšie, nemá zmysel púšťať sa do hlbších úvah, ak nerozumieme, s čím pracujeme.
- Aká je otázka, na ktorú chceme odpovedať? Môže ma zaujímať pochopenie akéhosi fenoménu alebo ma môže zaujímať predikcia (počasie, finančné trhy). Problémom je tiež, že ak sa budeme dívať na dáta dostatočne dlho, môžeme tam niečo nájsť, napriek tomu, že tam nič nie je a to čo vidíme je len výsledkom náhodnej variácie.
- Existujú v súbore dát chybne zaznamenané dáta? V kolónke vek môžete nájsť "Banská Bystrica".

Pri spolupráci klienta s aplikovaným štatistikom je veľmi dôležitá komunikácia. Musíme poznať potreby klienta, pochopiť kostru skúmaného problému a zvoliť vhodný štatistický model (častokrát jednoduchší model je presvedčivejší). Ďalej musíme vedieť interpretovať výsledky a obhájiť vhodnosť voľby modelu.

### 1.1 Dáta

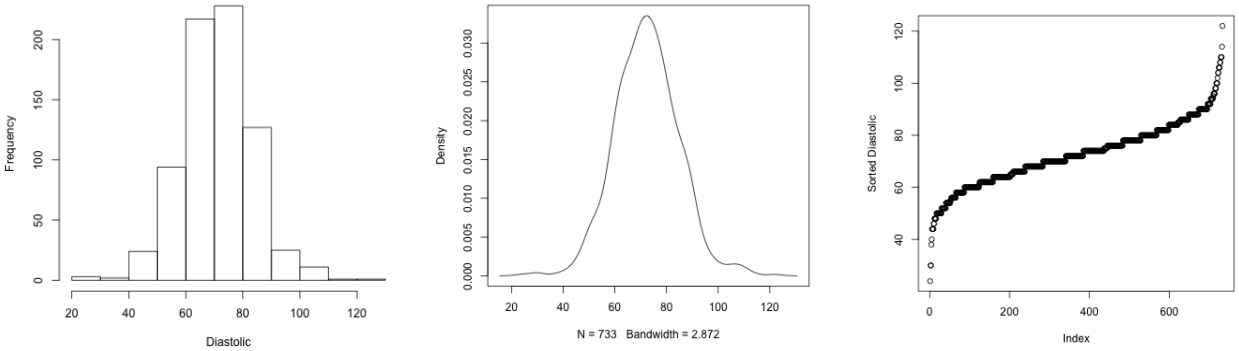
Nesmierne dôležitým nástrojom na komunikáciu je vizualizácia. Dobre zvolený obrázok vie častokrát odkomunikovať viac ako mnoho tabuliek.

Ako prvé sa pozrieme na hrubé dáta. Hľadáme nejaké chybne zaznamenané dáta alebo podozrivo veľké a malé hodnoty. Sumárne štatistiky nám tiež môžu pomôcť odhaliť chyby. Napr. v dátach môžeme vidieť mnoho pacientov, ktorí majú zaznamenaný diastolický krvný tlak nula. Tu je nula nezmyselná a

*E-mail: lukas.laffers@umb.sk. Vrelo ďakujem Samuelovi Hudecovi a Michaele Mihokovej za pripomienky.

zrejme len kóduje chýbajúce pozorovanie. Bez toho, aby sme sa pozreli na sumárne štatistiky, by sme na toto nemuseli prísť a dospeli by sme ku chybným záverom.

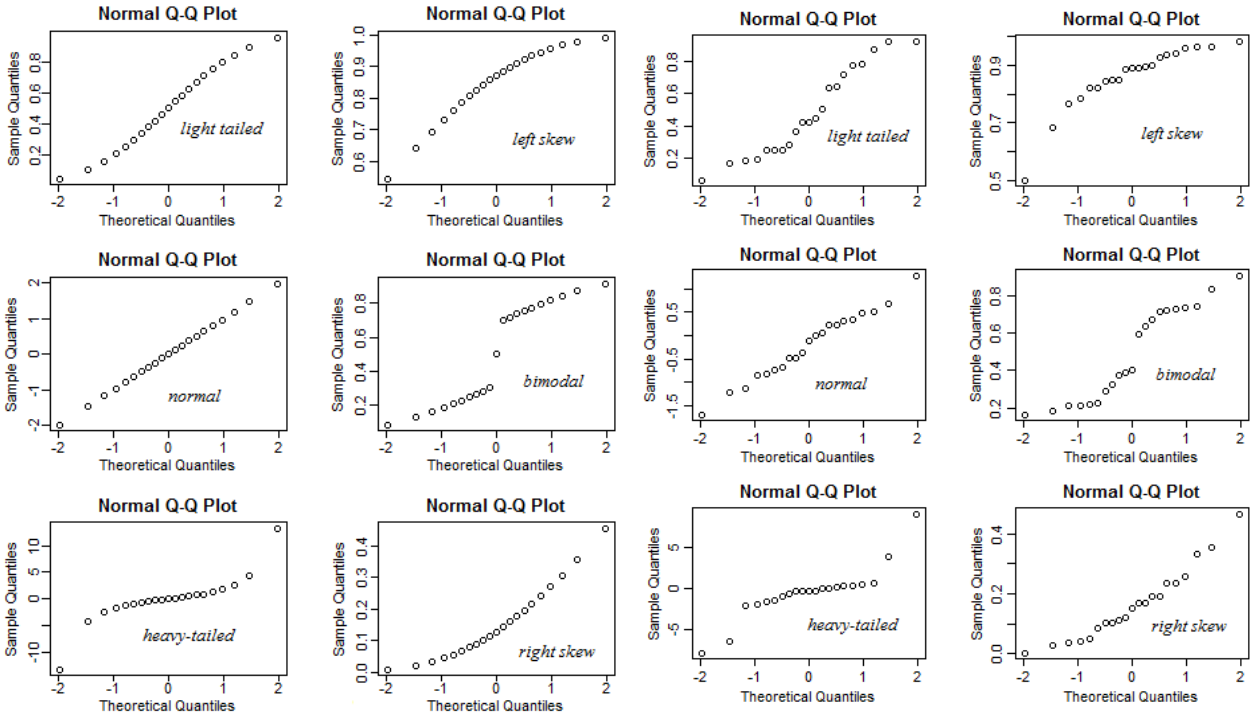
Dôležité je rozlišovať numerické a kategorické premenné. Ak si nedáme pozor, budeme počítať s nezmyselnými štatistikami ako napríklad priemerné PSČ.




Obr. 1: Histogram, vyhladnený jadrový odhad hustoty a graf usporiadaných dát. (Zdroj: [Far14])

Pri histograme musíme zvoliť šírku a počet "chlievikov", pri odhade hustoty zasa mieru vyhladenia. Z posledného obrázka môžeme vyčítať diskrétnu povahu dát.

Na skúmanie tvaru pravdepodobnostnej distribúcie nám môže pomôcť QQ-plot. Porovnáva nejakú dopredu danú teoretickú distribúciu (napr. normálneho rozdelenia) s empirickou distribúciou získanou z dátovej vzorky. Môžeme napríklad porovnať, či naša dátovej vzorka je viacej naklonená doľava alebo či má ťažšie alebo ľahšie chvosty ako teoretická distribúcia.




Obr. 2: QQ-plot pre rôzne tvary distribúcií, vľavo priemerný QQ-plot, vpravo QQ-plot pre konkrétnu dátovú vzorku (Zdroj: http://stats.stackexchange.com/questions/101274/how-to-interpret-a-qq-plot)

.

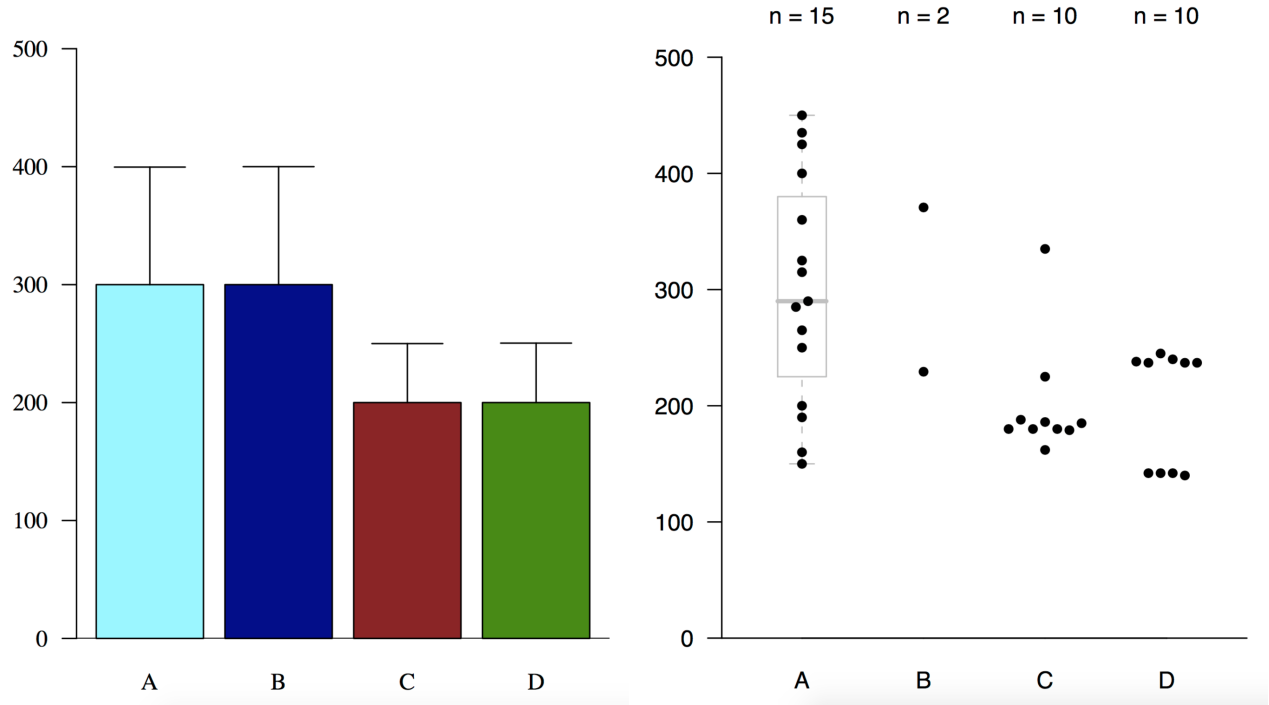
## 1.2 Exploratórna analýza dát

Exploratórna analýza dát je o tom, ako ďaleko sa dostaneme v skúmaní dát bez akéhokoľvek štatistického modelu. Kniha Johna Tukey-ho [Tuk77] je štandardnou referenciou.

## 1.3 Zakladné princípy vizualizacie dát

- Sumárne štatistiky neodpovedajú na všetky otázky. Častokrát je lepšie ukázať celú distribúciu. Na druhej strane, pre normálne rozdelené náhodné premenné sú stredná hodnota a variancia *postačujúce štatistiky*, teda zahrňujú kompletnú informáciu o pravdepodobnostom rozdelení.
- Je užitočné zobraziť surové dáta. Surové dáta sú postačujúcou štatistikou akejkoľvek dátovej vzorky.
- Každému dátovému bodu zodpovedá jedna machuľa atramentu a ďalšie pridávame len ak nám pomôžu pochopiť niečo nové.
- Dobrá vizualizácia ukáže na problém v dátach, moment prekvapenia. Človek sa niečo dozvie len z obrázku.
- Graf by mal byť samovysvetľujúci, ak niekto nečíta celý text ale padne mu zrak len na obrázok, mal by pochopiť, čo ten obrázok hovorí. Preto nešetríme dôsledným označením, či vysvetlením, najmä ak je použitý nejaký komplexnejší model.
- Vyhýbame sa zbytočnej sumarizácii, najmä cez zdroje variácie.
- Farby lepšie vyniknú na tmavšom podklade ako na čisto bielom.
- Farby sa zväčša používajú na vysvetľovanú premennú, nie na prediktory.
- Individuálnym výstupom rozumieme na základe porovnávania podobných objektov.

Pokiaľ to je možné, je lepšie ukázať radšej surové dáta. Nasledujúci príklad ukazuje, že výrazne rozdielne distribúcie môžu zdieľať tie isté sumárne štatistiky. Špeciálny zlý je napr. "dynamitový graf", ktorý zobrazuje strednú hodnotu a smerodajnú odchýlku.




Obr. 3: Dynamitový graf je rovnaký pre výrazne rôzne distribúcie. Prečo zahadzovať informáciu? Zdroj: http://biostat.mc.vanderbilt.edu/ wiki/pub/Main/TatsukiRcode/Poster3.pdf

.

Mnoho zaujímavých princípov vizualizácie dát nájdete v [Don11].

Naviac, v programovacom jazyku R existuje vynikajúca knižnica na produkovanie obrázkov `ggplot2` [Wic09].

## 1.4 Príklad z histórie

V minulosti potrebovali námorníci spoľahlivé spôsoby, aby zistili svoju polohu. V 18. storočí nemalo veľa námorníkov GPS navigácie a taktiež satelitov nebolo toľko čo dnes. Zemepisnú šírku vedeli určiť

pomocou Polárky, so zemepisnou dĺžkou to bolo ťažšie. Tobias Mayer na to chcel použiť informáciu o polohe krátera Manilius na mesiaci. Po použití rôznych geometrických identít a zanedbaní faktorov, ktorých vplyv je malý, dospel k lineárnemu vzťahu (detaily aj celý príbeh v [Sti86])

$$\texttt{arc} = \beta + \alpha\texttt{sinang} + \gamma\texttt{cosang},$$

kde $\beta, \alpha, \gamma$ sú neznáme parametre, ktoré potrebujeme na to, aby sme z polohy krátera zistili zemepisnú dĺžku, a `arc`, `sinang` a `cosang` sú pozorované veličiny. Mayer mal k dispozícii 27 pozorovaní a potreboval odhadnúť 3 neznáme parametre. Veličiny sú pozorované s chybou a pri odvodení sme zanedbali "malé" členy, preto neexistuje trojica parametrov tak, aby 27 rovníc platilo presne. Čo urobil Mayer? Rozdelil si 27 pozorovaní do 3 rovnako veľkých skupín podľa toho, ako veľmi rôzne boli hodnoty `sinang`. V každej skupine sčítal všeky pozorovania a potom riešil sústavu 3 lineárnych rovníc o 3 neznámych.

Legendre publikoval v roku 1805 metódu najmenších štvorcov. Namiesto arbitrárneho zoskupovania pozorovaní do rôznych skupín explicitne zaviedol chybový člen $\epsilon_i$,

$$\texttt{arc}_i = \beta + \alpha\texttt{sinang}_i + \gamma\texttt{cosang}_i + \epsilon_i, \quad \text{pre} \;\; i = 1, ..., 27$$

a navrhol zvoliť parametre $\beta, \alpha, \gamma$ tak, aby boli minimalizované štvorce odchýliek $\sum_{i=1}^{27} \epsilon_i^2$. To, prečo je toto dobrý nápad, sa dozvieme neskôr.

## 1.5 Prečo sa regresia volá regresia

Vysokí rodičia majú vysoké deti, nízki rodičia majú nízke deti. V priemere. Otázkou teraz je, či najvyšší rodičia budú mať najvyššie deti a najnižší rodičia budú mať najnižšie deti. Odpoveďou je, že nie tak celkom.

Galton sa v roku 1875 pozrel na to, ako výška rodičov predikuje výšku detí.

$$\texttt{childHeight} = \alpha + \beta\texttt{midparent} + \epsilon,$$

Model $y = \alpha + \beta x + \epsilon$, kde parametre sú odhadnuté metódou najmenších štvorcov vieme upraviť na $\frac{y-\bar{y}}{SD_y} = r\frac{x-\bar{x}}{SD_x}$, kde $r$ je korelačný koeficient medzi $x$ a $y$.

Možno by sme si boli mysleli, že rodičia, ktorí sú o jednu štandardnú odchýlku vyšší ako priemer rodičov, majú aj dieťa, ktoré je o jednu štandardnú odchýlku vyššie ako priemerné dieťa, To by však znamenalo, že $r = 1$. Parametre $\alpha, \beta$, ktoré zodpovedajú $r = 1$, sú označené prerušovanou čiarou.

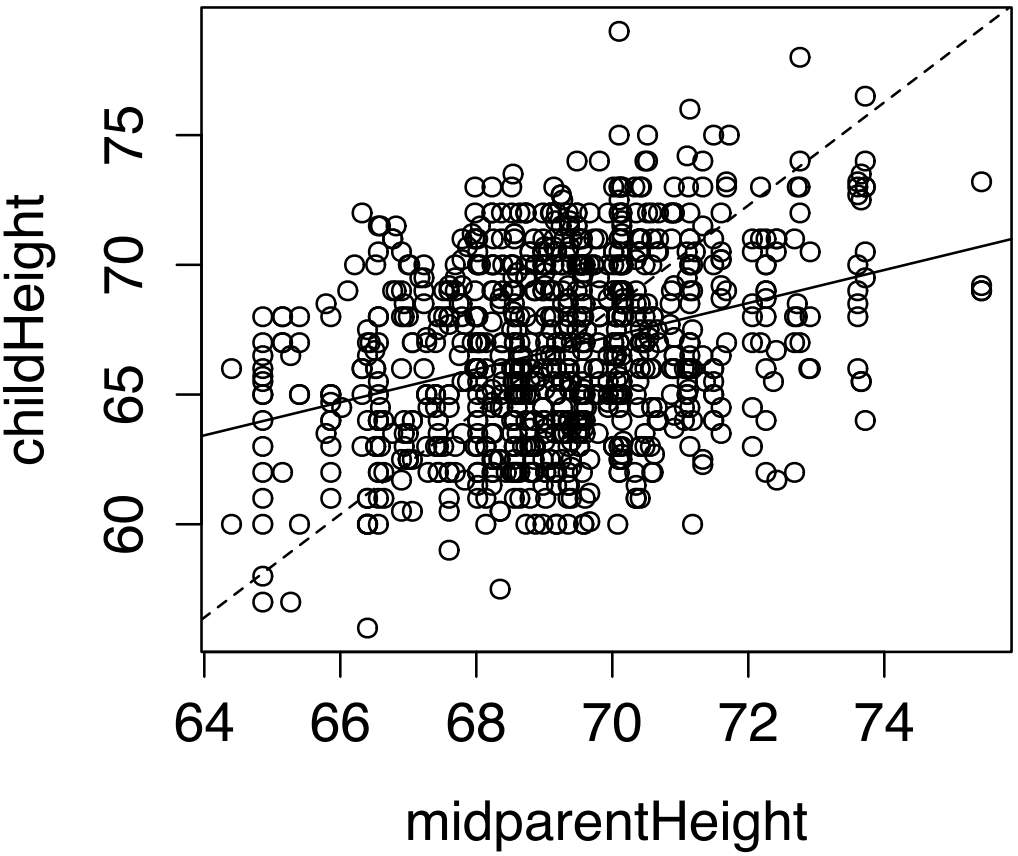
Galton toto pozorovanie nazval *regression to mediocrity* (návrat k priemernosti) a odtiaľ máme názov regresia, ktorý používame dodnes.

Tento efekt je spôsobený tým, že výška rodičov je síce faktor, ktorý ovplyvňuje výšku detí, avšak nie je jediný! Je to čiastočne spôsobené aj náhodnou kombináciou génov, spôsobom života, prostredím. Dieťa, ktoré je najvyššie v skupine, skrátka mohlo mať "šťastie".

Tento efekt sa vyskytuje v mnoho rôznych oblastiach a častokrát zostáva nepovšimnutý. Čudujeme sa, že športovcom, ktorým sa veľmi darí jeden rok a ďalší rok sa im darí len dobre. Alebo, že študenti, ktorí mali najhoršie výsledky, sa najviac zlepšili (preto hodnotenie škôl len podľa toho, ako veľmi sa zlepšili, nie je dobrý nápad).

Výnosnosť firiem v konkurenčnom prostredí je v priemere konštantná, teda ak sa firme darí veľmi dobre, môžeme očakávať, že v budúcnosti sa jej môže dariť o kus bližšie k priemeru. Keď toto ukázal na obrovskom množstve dát v roku 1933 Secrist vo svojej knihe [HS33], recenzent Hotelling to skomentoval ako "proving the multiplication table by arranging elephants in rows and columns, and then doing the same for numerous other kinds of animals" [SHR+34].

Ak zavedieme rýchlostné kamery na miesta, kde nastáva veľa nehôd, ale neberieme do úvahy regression to mediocrity efekt, veľmi pravdepodobne preceníme efekt zavedenia kamier.




Obr. 4: Vysokí rodičia majú v priemere vysoké deti ale nie až tak veľmi, ako sú oni sami vysokí.

# 2 Odhadovanie

Predtým, ako sa začneme zamýšľať nad tým, čo je rozumný spôsob odhadovania neznámych parametrov, musíme si uzrejmiť, čo je to model. Modelom nazývame množinu predpokladov týkajúcich sa náhodných premenných, ktorých dátovú vzorku pozorujeme. Tieto predpoklady môžu byť rozumné alebo nerozumné a môžu byť testovateľné alebo netestovateľné.

Dobrý model je ako dobrá mapa, popisuje dôležité aspekty premenných a vzťahov medzi nimi, ale zároveň nie je zbytočne (a častokrát nesprávne) podrobný. Vhodnosť modelu je posudzovaná na základe účelu použitia. Model na predikciu musí dobre predikovať a model na vysvetlenie nejakého fenoménu ho musi zrozumiteľne a dôveryhodne vysvetliť.

Jeden z predpokladov modelu môže byť akási funkčná závislosť medzi premennými. Niektoré z týchto premenných môžu byť aj nepozorované. O tých musíme zväčša povedať aj čosi viacej, aby bol náš model užitočný. Napr. majme nasledujúcu funčnú závislosť medzi porozovanými náhodnými premennými $Y$ a $X_1, X_2, X_3$ a nepozorovanou náhodnou chybou $\epsilon$.

$$Y = f(X_1, X_2, X_3) + \epsilon,$$

o chybe $\epsilon$ väčšinou predpokladáme, že je náhodná, teda, že nie je systematicky ovplyvnená premennými $X_1, X_2, X_3$, čo môže byť, že $\epsilon$ je nezávislá od $(X_1, X_2, X_3)$ alebo slabšiu verziu nulovej podmienenej strednej hodnoty a teda, že $E(\epsilon|X_1, X_2, X_3) = 0$.

Čo to znamená, že je model lineárny? Myslíme tým, že funkčná závislosť medzi $Y$ a $X$ je lineárna v *parametroch*. Teda, zdanlivo paradoxne, môžeme pomocou lineárneho modelu modelovať aj prudko nelineárne vzťahy

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 \log X + \epsilon.$$

Niektoré funkčné závislosti však nemôžu byť zapísané ako lineárne v parametroch, napr.

$$Y = \beta_0 + \beta_1 X^{\beta_2} + \epsilon;$$

niekedy však transformácia pomôže k linearizácii, napr.

$$Y = \beta_0 X^{\beta_1} \epsilon \quad \rightarrow \quad \log(Y) = \log(\beta_0) + \beta_1 \log(X) + \log(\epsilon) = \beta_0^* + \beta_1^* \log(X) + \log(\epsilon).$$

Lineárny neznamená nutne jednoduchý, ako by sa mohlo zdať. Lineárny je len v parametroch a lineárny model môže dobre vysvetľovať rôzne nelineárne závislosti medzi premennými.

Akým spôsobom môžeme dospieť k modelu?

- fyzikálnou teóriou - teda, napríklad vieme, že dĺžka pružiny je proporciálna hmotnosti závažia,
- skúsenosťou z predošlých modelov,
- pozeraním na dáta a skúšaním.

Dáta je užitočné zobrazovať v maticovej forme.

Funkčnú závislosť

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$$

pre $i = 1, \ldots, n$ vieme napísať v maticovej forme nasledovne

$$y = X\beta + \epsilon,$$

kde

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{pmatrix}, \; X = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ \ldots & & \ldots & \\ 1 & x_{n1} & x_{n2} & x_{n3} \end{pmatrix}, \; \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}, \; \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \ldots \\ \epsilon_n \end{pmatrix}.$$

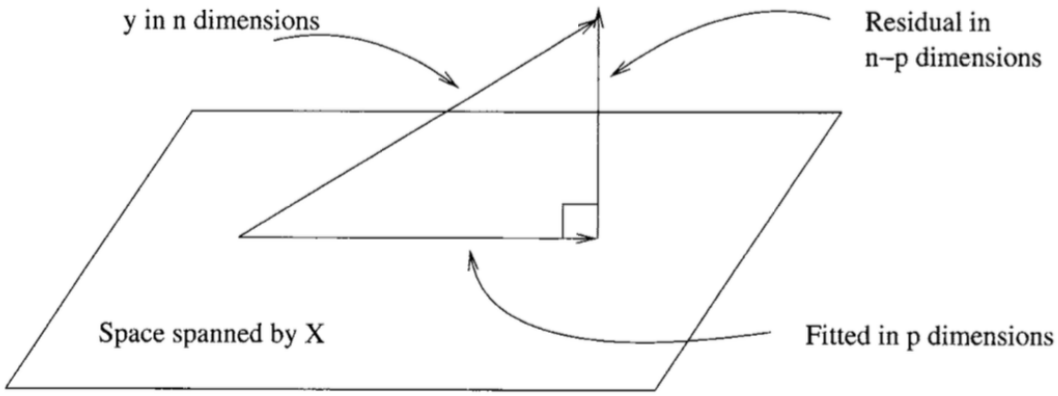
Pozorovaní máme $n$, no k dispozícii len 4 parametre. Ako rozumne zvoliť vektor $\beta$? Jeden spôsob sme už uviedli, a to navoliť ich tak, aby $\sum_{i=1}^n \epsilon_i^2 = \epsilon^T \epsilon$ bolo čo najmenšie.

Chceli by sme, aby systematická časť $X\beta$ bola čo najbližšie $y$ a náhodná zložka $\epsilon$ bola čo najmenšia. Vektor $y$ je bod v $n$-dimenzionálnom priestore, avšak ten sa nenachádza (alebo nemusí nutne nachádzať) v $p$ dimenzionálnom priestore, ktorý je generovaný stĺpcami matice $X$. Minimalizovaním štvorcov odchýliek vlastne minimalizujeme euklidovskú vzdialenosť vektora $y$ od podpriestoru generovaného stĺpcami matice $X$.

Minimalizujeme $\epsilon^T\epsilon = (y - X\beta)^T(y - X\beta)$. Zdiferencovaním cez $\beta$ dostávame podmienky prvého rádu pre extrém:

$$X^T X\hat{\beta} = X^T y.$$

Ak je $X^T X$ invertovateľná, tak máme $\hat{\beta} = (X^T X)^{-1}X^T y$. Potom $\hat{y} = X\hat{\beta} = X(X^T X)^{-1}X^T y = Hy$. Maticu $H$ nazývame aj *hat matrix* a je maticou ortogonálnej projekcie do priestoru generovaného stĺpcami matice $X$. Projekčná matica je symetrická ($H = H^T$) a idempotentná ($H = HH$). Nakoľko minimalizujeme štvorce odchýliek modelu od skutočných hodnôt $y$, táto odhadovacia metóda sa nazýva **metóda najmenších štvorcov** (MNŠ alebo v angličtine *ordinary least squares (OLS) method*).




Obr. 5: Geometria metódy najmenších štvorcov (Zdroj: [Far14])

Teraz máme analytickú formulu na výpočet MNŠ odhadcu $\hat{\beta}$, avšak ak chceme $\hat{\beta}$ vypočítať, nerobíme to priamo, to by bolo totižto numericky veľmi neefektívne a nestabilné, invertovanie matíc je totiž numericky náročné.[1] Necháme, nech to urobí za nás $R$ , ktoré na to použije QR-rozklad matice $X$. QR rozklad je založený na tom, že matica $X$ sa rozloží na $X = Q\begin{pmatrix} R \\ \mathbf{0} \end{pmatrix}$, kde $Q$ je ortogonálna $n \times n$ matica, $R$ je horná trojuholníková matica a $\mathbf{0}$ je $(n-p) \times p$ matica.

$$RSS = (y-X\beta)^T(y-X\beta) = ||y-X\beta||^2 = ||Q^T y - Q^T X\beta||^2 = \left|\left|\begin{pmatrix} f \\ r \end{pmatrix} - \begin{pmatrix} R \\ \mathbf{0} \end{pmatrix}\beta\right|\right|^2 = ||f-R\beta||^2 + ||r||^2,$$

takže reziduálna suma štvorcov môže byť minimalizovaná riešením hornej trojuholníkovej sústavy, čo je numericky výhodné.[2] Týmto spôsobom sa vyhneme priamemu invertovaniu matice $X^T X$, čo je nestabilné a numericky neefektívne.

## 2.1 Gauss-Markovova veta

Teraz sa dozvieme, prečo je MNŠ veľmi rozumná voľba vektora parametrov $\beta$.

- Je geometricky zrozumiteľná.
- Ak sú chyby $\{\epsilon_i\}$ nezávislé a normálne rozdelené s rovnakou varianciou, potom je to aj odhad metódou maximálnej vierohodnosti (maximum likelihood estimator).

[1] Na invertovanie matice $p \times p$ pomocou Gauss-Jordanovej eliminácie potrebujeme rádovo $p^3$ operácií.

[2] Označíme $Q^T y = \begin{pmatrix} f \\ r \end{pmatrix}$.

- Je najlepší (v zmysle najmenšej variancie) odhad z triedy nevychýlených lineárnych odhadov. (o tomto je GM veta)

Čo to znamená nevychýlený? $E(\hat{\beta}) = \beta$. To znamená, že náhodný vektor $\hat{\beta}$ je centrovaný okolo skutočného, ale nám neznámeho vektora $\beta$.

Čo to znamená lineárny odhad? To znamená, že sa dá napísať ako $Ay$, teda lineárna kombinácia $y$. V prípade MNŠ je to $A = (X^T X)^{-1} X^T$

Čo to znamená najlepší v zmysle variancie? To znamená, že pre hocijakú kombináciu $c^T\beta$ má $c^T\hat{\beta}$ najmenšiu možnú varianciu. A to je fajn, lebo pre odhadcu je lepšie, aby bol čo najpresnejší (čo najmenej variabilný), aby mal čo najmenšiu varianciu.

**Gauss-Markovova veta**: Nech $E(\epsilon) = 0$, nech $var(\epsilon) = \sigma^2 I$ a nech $E(y) = X\beta$. Ďalej nech $\Psi = c^T\beta$ je *odhadnuteľná funkcia*, teda nech existuje lineárna kombinácia $a^T y$ taká, že $E(a^T y) = c^T\beta$ pre všetky $\beta$. Potom spomedzi všetkých lineárnych nevychýlených odhadov má odhad $\hat{\Psi} = c^T\hat{\beta}$ najmenšiu varianciu a takýto (lineárny a varianciu minimalizujúci) odhad je jednoznačne určený.

**Dôkaz Gauss-Markovovej vety**: Nech $a^T y$ je nevychýleným odhadom $c^T\beta$ (začneme lineárnym a nevychýleným odhadom neznámeho parametra $c^T\beta$, lebo o iných odhadoch GM veta ani nehovorí), takže $\forall\beta: \; E(a^T y) = a^T E(y) = a^T X\beta = c^T\beta$, a preto $a^T X = c^T$. Preto $c$ musí ležať v podpriestore generovanom stĺpcami matice $X^T$, a teda aj v podpriestore generovanom stĺpcami matice $X^T X$, a preto musí existovať (nejaké) $\lambda$ také, že $c = X^T X\lambda$. Takže dostávame, že pre náš MNŠ odhad $c^T\hat{\beta}$ platí:

$$c^T\hat{\beta} = \lambda^T X^T X\hat{\beta} = \lambda^T X^T X(X^T X)^{-1} X^T y = \lambda^T X^T y.$$

Zoberme si teraz hocijaký lineárny odhad $a^T y$ a pozrime sa na jeho varianciu

$$\begin{gathered} var(a^T y) = var(a^T y - c^T\hat{\beta} + c^T\hat{\beta}) = \\ var(a^T y - \lambda^T X^T y + c^T\hat{\beta}) = \\ = var(a^T y - \lambda^T X^T y) + var(c^T\hat{\beta}) + \\ 2cov(a^T y - \lambda^T X^T y, \lambda^T X^T y). \end{gathered}$$

Posledný člen

$$\begin{gathered} cov(a^T y - \lambda^T X^T y, \lambda^T X^T y) = \\ (a^T - \lambda^T X^T)\sigma^2 I X\lambda = \\ (a^T X - \lambda^T X^T X)\sigma^2 I\lambda = (c^T - c^T)\sigma^2 I\lambda = 0, \end{gathered}$$

preto

$$\begin{gathered} var(a^T y) = var(a^T y - \lambda^T X^T y) \\ + var(c^T\hat{\beta}) \geq var(c^T\hat{\beta}). \end{gathered}$$

Čo sa jednoznačnosti týka: rovnosť nastane, ak $var(a^T y - \lambda^T X^T y) = 0$, a teda ak $a^T = \lambda^T X^T$. Toto však znamená, že $a^T y = \lambda^T X^T y = c^T\hat{\beta}$. Teda rovnosť nastáva, keď $a^T y = c^T\hat{\beta}$, preto je lineárny a varianciu minimalizujúci estimátor jednoznačne určený. □

Takže ak platí lineárny model, chyby $\epsilon$ sú nekorelované a majú rovnakú varianciu, MNŠ je veľmi rozumný spôsob odhadovania parametrov.

A čo keď predpoklady GM vety neplatia? Toto bude neskôr:

- Ak majú chyby nerovnakú varianciu alebo ak sú navzájom korelované, môžeme použiť zovšeobecnenú metódu najmenších štvorcov.
- Ak má distribúcia chýb príliš ťažké chvosty, tak je zväčša rozumné použiť robustnú regresiu, ktorá je nelineárna v $y$.
- Ak sú stĺpce matice $X$ veľmi korelované, potom nás môžu zaujímať vychýlené estimátory (majú menšiu varianciu ako nevychýlené).

## 2.2 Miera vhodnosti fitu

Ako dobre fituje náš model dáta? Užitočná metrika, ktorá toto vyjadruje jedným číslom, sa nazýva koeficient determinácie alebo $R^2$. Je definovaný nasledovne:

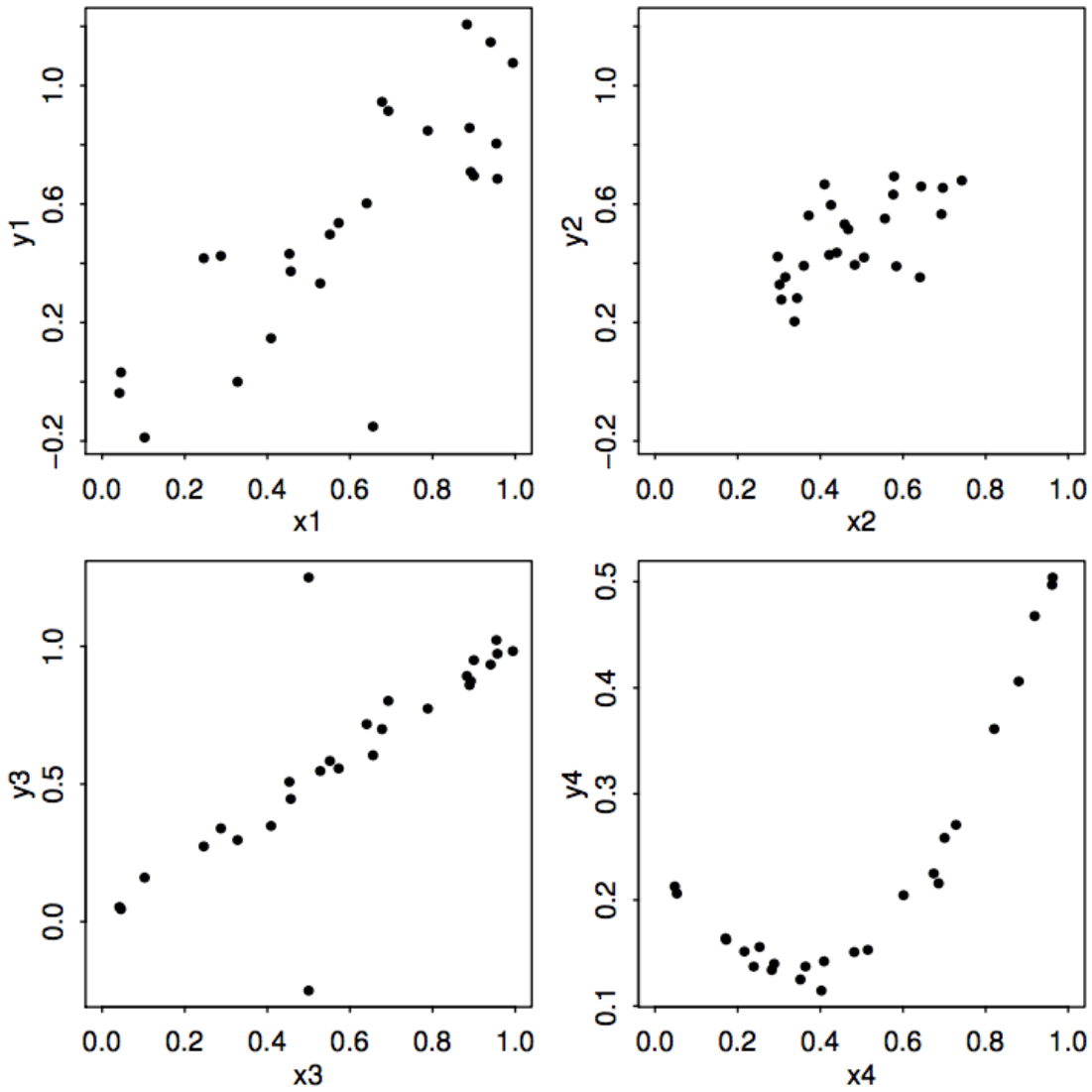
$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2} = 1 - \frac{RSS}{TSS},$$

RSS je skratka pre *residual sum of squares* a TSS je skratka pre *total sum of squares*. Dá sa ukázať, že

$$R^2 = (cor(\hat{y}, y))^2 = \frac{\sum(\bar{y}_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2}.$$

Teda $R^2$ je korelácia (miera lineárnej závislosti) na druhú lineárneho prediktora $y$ a samotného $y$.

Čo je veľké a čo je malé $R^2$ záleží od aplikácie. Kým vo fyzike to bude veľké číslo (napr. 0.95), v ekonómii vzťahy medzi premennými nepoznáme a je tam veľa šumu. Ako vždy keď máme čosi zložité (dáta) reprezentované len jedným číslom, tak je to skresľujúce. Tu je obrázok 4 datasetov s rovnakým $R^2$.




Obr. 6: Rovnaké $R^2$ pri výrazne rôznych datasetoch. (Zdroj: [Far14])

V čom je teda $R^2$ skresľujúce? V tom, že hovorí o kvalite lineárneho fitu. Avšak v určitých situáciách, najmä ak dáta vykazujú výrazné nelinearity, toto nie je zaujímavá informácia. Podobne outliery výrazne skresľujú výpovednú hodnotu tohto čísla. Na $R^2$ sa teda budeme pozerať ako na mieru mechanického fitu a nebudeme nutne podľa neho posudzovať kvalitu modelu.

Hodnote $R^2$ sa vágne hovorí aj percento vysvetlenej variácie v $y$

## 2.3 Identifikovateľnosť

Ak matica $X^T X$ nie je invertovateľná, tak potom neexistuje unikátna hodnota, ktorá minimalizuje sumu štvorcov chýb. Nastane to vtedy, keď stĺpce matice $X$ sú lineárne závislé, takže nejaký prediktor je lineárnou kombináciou iných prediktorov. To sa stane napríklad keď:

- Máme veľa prediktorov a kus neporiadok v nich a stane sa, že máme ten istý prediktor (ale napríklad v iných jednotkách) dvakrát.
- Máme viac premenných ako pozorovaní, teda ak $p > n$.

Koncept identikácie alebo identifikovateľnosti je zložitejší a ide o to, či štatistický model dokopy s distribúciou pozorovaných veličín jednoznačne určuje parameter. Napríklad ak poznám pravdepodobnostnú distribúciu $(y, X)$ a zároveň stĺpce matice $X$ nie sú kolineárne, potom parameter $\beta$ v lineárnom modeli je jednoznačne určený.

## 2.4 Ortogonalita

Ak sú stĺpce matice $X$ ortogonálne, potom hodnoty parametrov nie sú citlivé na to, ktorý prediktor figuruje alebo nefiguruje v regresii. Z regresie

$$Y = X\beta + \epsilon = X_1\beta_1 + X_2\beta_2 + \epsilon$$

dostaneme rovnaké $\hat{\beta}_1$ ako z regresie

$$Y = X_1\beta_1 + \epsilon.$$

Toto je užitočné najmä vtedy, keď chceme parametre nejako interpretovať. Predsa len; je nepríjemné, ak sa po pridaní ďalšieho prediktora výrazne zmení hodnota ostatných parametrov. Na druhej strane je to signálom toho, že predošlý model nebol dobrý.

# 3 Štatistická inferencia

Doteraz sme si hovorili o tom, ako odhadnúť parametre lineárneho modelu. Vysvetlili sme si, prečo je za určitých podmienok odhad metódou najmenších štvorcov rozumnou voľbou. Zatiaľ však nevieme nič o pravdepodobnostnom správaní takýchto odhadcov (estimátorov). Tým pádom ani nevieme, ako veľmi presné alebo nepresné sú tieto odhady, teda ako veľmi môžeme týmto číslam dôverovať. Čím je naša dátová vzorka väčšia, tým presnejšie naše odhady budú.

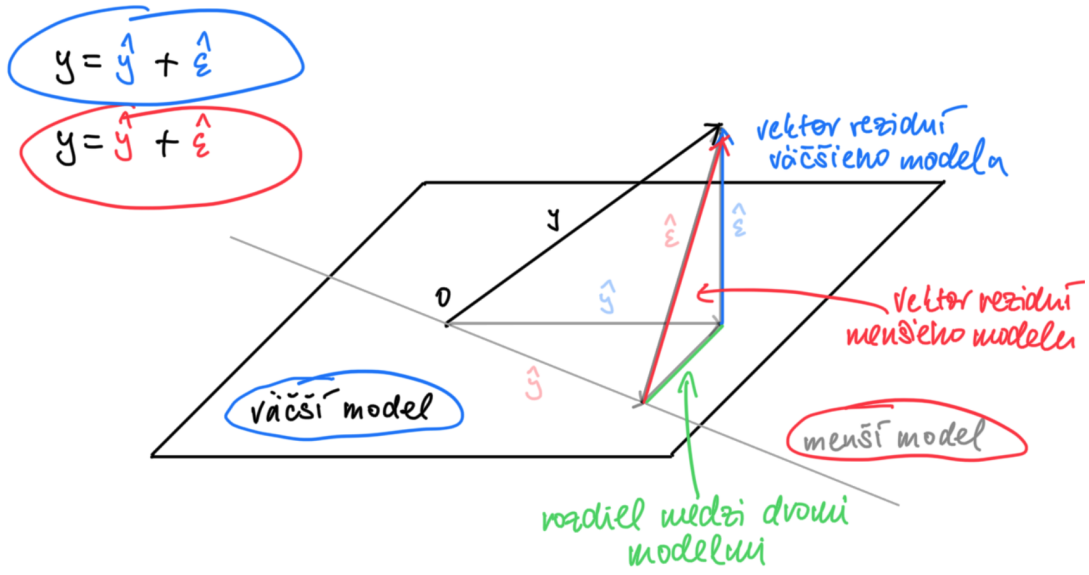
Budeme predpokladať, že chyby $\epsilon$ sú normálne rozdelené, doteraz sme predpokladali, že sú nezávislé a rovnako rozdelené , že majú strednú hodnotu 0 a rovnakú varianciu. Dokopy to dáva $\epsilon \sim N(0, \sigma^2 I)$ a nakoľko $y = X\beta + \epsilon$, dostávame $y \sim N(X\beta, \sigma^2 I)$.

To, že my predpokladáme, že chyby sú normálne rozdelené ešte neznamená, že je to **pravda**. Je to predpoklad, je to prostriedok na to, aby sme si spravili (dúfajme že) užitočný model. Môžeme sa pozrieť na odhadnutý model a vidím pomocou QQ-plotu, že odhady chýb, teda reziduá, vôbec nevyzerajú byť normálne rozdelené. Normalita je výhodná v tom, že odhad MNŠ je potom odhad metódou maximum likelihood, teda je v istom zmysle optimálny. Normalita je prirodzene výsledkom centrálnej limitnej vety, vieme, že pre iid premenné sa priemer správa pre veľkú dátovú vzorku stále viac a viac ako normálne rozdelenie.[3] Poznamenajme, že Gauss-Markova veta nepotrebovala predpoklad normality. Predpoklad normality však robí MNŠ odhad optimálny z oveľa väčšej triedy odhadcov ako len lineárne a nevychýlené.

Za predpokladu normality teda dostávame:

$$\hat{\beta} = (X^T X)^{-1} X^T y \sim N\left(\beta, (X^T X)^{-1}\sigma^2\right).$$

Predstavme si, že chceme testovať hypotézu, či dáta pochádzajú z akéhosi "malého"modelu $\omega$, ktorý je špeciálnym prípadom veľkého modelu $\Omega$, napríklad, že v malom modeli je menej prediktorov, teda niektoré komponenty vektora parametrov $\beta$ sú nulové. Väčší model $\Omega$ bude vždy lepšie fitovať dáta ako malý model $\omega$, ale ak budú rozdiely maličké, preferujeme menší model, lebo je jednoduchší. Toto je zobrazené na obrázku 7.




Obr. 7: Geometria najmenších štvorcov (Inšpirácia: [Far14])

Fit môžeme merať veľkosťou RSS, takže napríklad

$$\frac{RSS_\omega - RSS_\Omega}{RSS_\Omega}$$

[3] Hovoríme, že konverguje podľa distribúcie k normálnemu rozdeleniu.

vyzerá ako rozumná testovacia štatistika. Dá sa ukázať, že je to štatistika testu pomocou pomeru vierohodností (likelihood ratio statistic).

Nasledujúca kvantita má, za predpokladu normality, $F$ rozdelenie s $(p-q)$ a $(n-p)$ stupňami voľnosti.

$$F = \frac{(RSS_\omega - RSS_\Omega)/(p-q)}{RSS_\Omega/(n-p)} =$$

$$\frac{(RSS_\omega - RSS_\Omega)/(df_\Omega - df_\omega)}{RSS_\Omega/(df_\Omega)} \sim F_{p-q,n-p}$$

kde $n$ je veľkosť dátovej vzorky, $p$ je počet parametrov vo veľkom modeli $\Omega$ a $q$ je počet parametrov v malom modeli $\omega$. Hodnoty $df_\omega = n-q$ a $df_\Omega = n-p$ označujú počet stupňov voľnosti v modeloch.

Ako vyzerajú rozličné hypotézy, ktoré môžeme testovať?

**Testovanie významnosti všetkých prediktorov naraz**

$$H_0 : \beta_1 = \beta_2 = ... = \beta_{p-1} = 0$$

Tu porovnávame veľký model $\Omega$ so všetkými prediktormi a "hlúpy"model len s konštantou *(null model).* Testovacia štatistika má nasledovnú formu:

$$F = \frac{(TSS - RSS)/(p-1)}{RSS/(n-p)} \sim F_{p-1,n-p}.$$

**Testovanie významnosti len jedného konkrétneho prediktora**

Testujeme hypotézu, či efekt $i$-teho prediktora je štatisticky významný:

$$H_0 : \beta_i = 0.$$

Testovacia štatistika má tvar

$$t_i = \frac{\hat{\beta}_i}{se(\hat{\beta}_i)} \sim t_{n-p}.$$

Poznamenajme, že $t_i^2 \sim F_{1,n-p}$. Pokiaľ $n$ je dostatočne veľké, tak je $t$-rozdelenie veľmi podobné normálnemu kritická hodnota obojstranného testu na hladine významnosti 95% je približne 2.

**Testovanie významnosti viacerých prediktorov**

Môžme testovať hypotézu, či je efekt $i$-teho a $j$-teho prediktora rovnaký:

$$H_0 : \beta_i = \beta_j = 0$$

Testovacia štatistika má tvar:

$$F = \frac{(RSS_\omega - RSS_\Omega)/(2)}{RSS_\Omega/(n-p)} \sim F_{2,n-p}.$$

**Testovanie významnosti viacerých lineárnych reštrikcií na parametre modelu**

Vieme testovať všelijaké reštrikcie, napr.

$$H_0 : \beta_1 = \beta_2,$$

ide len o to šikovne si zvoliť ten menší model, v tomto prípade to bude:

$$y = \beta_0 + \beta_1^*(x_1 + x_2) + \beta_3 x_3 + ... + \beta_p x_p + \epsilon,$$

a znova skonštruovať $F$ štatistiku rovnako ako vo všeobecnom prípade.

## 3.1 Permutačný test

Čo keď naše chyby nie sú normálne rozdelené, resp. dátová vzorka nie je dostatočne dobrá na to, aby centrálna limitná veta pomohla urobiť estimátor dostatočne podobný normálnemu rozdeleniu.

Predstavme si, že chceme otestovať nulovú hypotézu $H_0 : \beta_1 = ... = \beta_{p-1} = 0$. To znamená, že prediktory vôbec nepomôžu s predikciou $y$. To znamená, že keď náhodne zamiešam (spermutujem) vektor $y$, tak potom $F$ štatistika bude mať stále také isté rozdelenie. Týmto spôsobom môžem nasamplovať veľa realizácií F štatistiky, dostanem odhad distribúcie tejto štatistiky za platnosti nulovej hypotézy. Teraz sa pozriem na moje pôvodné dáta a tú jednu jedinú realizáciu F štatistiky čo mám. Pozriem sa, ako veľmi je extrémna oproti "nasamplovanej"distribúcii.

## 3.2 Akým spôsobom sme získali dátovú vzorku?

Z našej dátovej vzorky sa chceme čomusi priučiť o celej populácii.

- Náhodná vzorka - toto je zlatý štandard, v medicíne štandardom, v sociálnych vedách výnimkou.
- Nenáhodná vzorka - niektorí pacienti odmietnu liečbu, možno na to majú nejaké dôvody, ktoré spôsobujú, že liečba by na nich mala iný efekt ako na priemerného pacienta.
- Sample of convenience - niekedy vzorku nevyberáme náhodne ale podľa nejakého mechanizmu o ktorom veríme, že nemá efekt na kvalitu vzorky.
- Celá populácia - tu nie je ani treba žiadnu štatistickú inferenciu a stačí popísať čo vidíme v dátach. Na druhej strane, štatistický test môže mať interpretáciu ako modelovanie pravdepodobnosti výberu našej populácie z akejsi množiny alternatívnych svetov.

## 3.3 Konfidenčné intervaly

Konfidenčný interval pre neznámy parameter $\beta_i$ je množina hodnôt parametrov $c$, ktoré by neboli zamietnuté v teste $H_0 : \beta_i = c$ pri fixnej hladine významnosti. Pre skalárny parameter $\beta_i$ je to

$$CI^\alpha = [\hat{\beta}_i - t_{n-p}^{(\alpha/2)} se(\hat{\beta}_i), \hat{\beta}_i + t_{n-p}^{(\alpha/2)} se(\hat{\beta}_i)],$$

kde $t_{n-p}^{(\alpha/2)}$ je $(\alpha/2)$-percentný kvantil Studentovho t-rozdelenia s $(n-p)$ stupňami voľnosti.

Konfidenčný interval je spravidla kvalitnejší typ informácie ako samotná p-hodnota nejakého testu. Korektná interpretácia je však dôležitá. Konfidenčný interval je náhodný interval, ktorý má nasledovnú vlastnosť. Ak by sme dáta generovali veľa krát, to znamená generovali veľa rovnako veľkých datasetov, tak v 95% prípadov by konfidenčný interval pokryl skutočnú hodnotu. To je však problém, my máme len 1 dataset a tým pádom len 1 konfidenčný interval. Možno pokrýva skutočnú hodnotu $\beta_i$ a možno nie. Interpretácia: "Konfidenčný interval na 95% pokrýva skutočný parameter"je mätúca, lebo nie je jasné čo je samplovací proces.

Zhrnuté: parameter je **fixná hodnota** a konfidenčný interval je **náhodný interval**. Konfidenčný interval chápeme ako akúsi množinu rozumných hodnôt pre neznámy fixný parameter.

Ak nás zaujíma viac než jeden parameter môžeme zobraziť **konfidenčnú množinu**, t.j. množinu hodnôt vektora parametrov $\beta$, pre ktoré platí

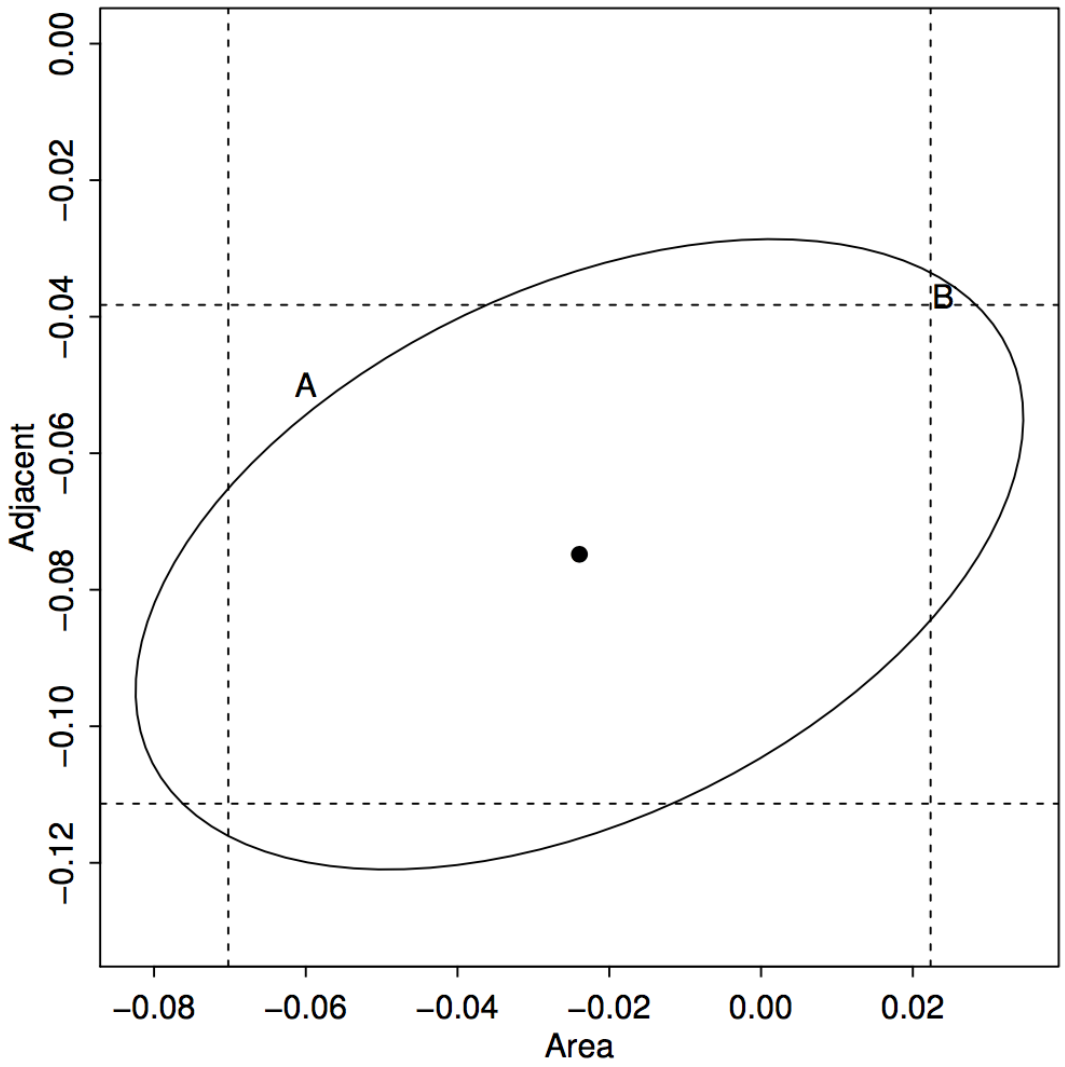
$$(\hat{\beta} - \beta)^T X^T X (\hat{\beta} - \beta) \leq p\hat{\sigma}^2 F_{p,n-p}$$

v dvoch rozmeroch to môže vyzerať napríklad ako na obrázku 9.

## 3.4 Konfidenčné intervaly založené na Bootstrape

Konfidenčné intervaly, ktoré sme si predstavili sú založené na $F$ alebo $t$ štatistikách, ktorých základom sú normálne rozdelené chyby $\epsilon$. Ak je tento predpoklad podozrivý alebo priamo nedôveryhodný, môžeme si pomôcť **bootstrapom**.

Predstavme si situáciu, že by sme poznali distribúciu $\epsilon$ov. Potom by sme nasledovný postup mohli zopakovať mnohokrát a dostali by sme distribúciu $\hat{\beta}$.




Obr. 8: 95% konfidenčná množina (Zdroj: [Far14])

(1) Vygeneruj náhodnú realizáciu $\epsilon$ zo známej distribúcie.
(2) Pre naše fixné $X$ a známe $\beta$ vypočítaj $y = X\beta + \epsilon$
(3) Vypočítaj $\hat{\beta}$.

Pomocou takejto simulácie by sme napríklad mohli skúmať vlastnosti estimátora. Túto myšlienku ideme použiť na zostrojenie konfidenčného intervalu. Rozdiel bude v tom, že namiesto toho aby sme poznali distribúciu $\epsilon$ sa budeme tváriť, že distribúcia $\epsilon$ je tá, ktorú odhadneme z lineárneho modelu. Toto urobíme veľakrát.

(1) Vygeneruj náhodnú realizáciu $\epsilon^*$ vyberaním hodnôt s opakovaním z $\hat{\epsilon}_1, ..., \hat{\epsilon}_n$.
(2) Pre naše fixné $X$ vypočítaj $y^* = X\hat{\beta} + \epsilon^*$
(3) Pomocou $(X, \hat{y})$ vypočítaj $\hat{\beta}^*$

Tým, že máme distribúciu $\hat{\beta}^*$ z nej poľahky urobíme napr. 95% konfidenčný interval tak, že zoberieme 2.5%ný a 97.5%ný kvantil.

# 4 Lineárna regresia ako nástroj predikcie a vysvetľovania

To či nás zaujíma predikcia alebo chceme vysvetliť nejaký fenomén sú dve veľmi rôzne veci. Pri predikcii nám ide o to dobre predikovať, napríklad preto, že správne predikcie nám pomôžu zarábať peniaze. Pokiaľ je model úspešný nevadí nám až tak, že nerozumieme prečo vlastne funguje.

## 4.1 Predikcia

Predstavme si lineárny regresný model, $y = X\beta + \epsilon$ s normálnymi nekorelovanými chybami $\epsilon \sim N(0, \sigma^2 I)$. Chceli by sme predikovať výstup $y$ pre ďalšie pozorovanie s vektorom prediktorov $x_0$. Lineárny model, ktorého parametre $\beta$ sú odhadnuté pomocou MNŠ predikuje

$$\hat{y}_0 = x_0^T \hat{\beta}.$$

Tento objekt, $\hat{y}_0$ je náhodná premenná. Jej stredná hodnota je $x_0^T\beta$ nás však môže zaujímať neistota spojená s týmto odhadom. Náhodnosť prichádza z $\hat{\beta}$ nakoľko prediktory $x_0$ považujeme za fixné. Chceli by sme kvantifikovať neistotu spojenú s tým, že máme len náhodnú vzorku nejakej fixnej dĺžky. Pre varianciu náhodnej premennej $\hat{y}_0$ platí

$$var(\hat{y}_0) = x_0^T (X^T X)^{-1} x_0 \sigma^2.$$

Neistotu budeme kvantifikovať cez konfidenčné intervaly.

Môžeme mať ambíciu predikovať dva rôzne objekty

- priemernú hodnotu $\hat{y}_0$, pre ňu máme konfidenčný interval $\hat{y}_0 \pm t_{n-p}^{(\alpha/2)} \hat{\sigma} \sqrt{x_0^T (X^T X)^{-1} x_0}$
- budúcu hodnotu, teda $\hat{y}_0 + \epsilon_0$, pre ňu máme konfidenčný interval $\hat{y}_0 \pm t_{n-p}^{(\alpha/2)} \hat{\sigma} \sqrt{1 + x_0^T (X^T X)^{-1} x_0}$. Poznamenajme, že nový člen 1 pod odmocninou je tam kvôli, tomu, že budúca hodnota má väčšiu variabilitu práve kvôli náhodnej chybe ktorú pričítavame k $\hat{y}_0$, lebo

$$var(\hat{y}_0 + \epsilon_0) = var(\hat{y}_0) + var(\epsilon_0) =$$

$$x_0^T (X^T X)^{-1} x_0 \sigma^2 + \sigma^2 = (1 + x_0^T (X^T X)^{-1} x_0)\sigma^2,$$

kde prvá rovnosť platí kvôli nekorelovanosti $\hat{y}_0$ a $\epsilon_0$. Teda

$$[\text{šírka intervalu}] \propto [\text{neistota v odhade } \hat{\beta}] + [\text{neistota v chybe } \epsilon_0]$$

Z analytického vyjadrenia pre konfidenčné intervaly vidíme, že konfidenčný interval pre priemernú hodnotu je užší ako pre budúcu hodnotu. Intervaly sú širšie pre také $x_0$, ktoré sú ďaleko od pozorovaní vo vzorke.

O čom sme tu doteraz vôbec nerozprávali je však **neistota v modeli**. Nech to znie depresívne ako chce, stále uvažujeme o lineárnom modeli s normálnymi chybami a neistota ktorú kvantifikujeme je podmienená tým, že náš model je korektne špecifikovaný. Nuž ale on nemusí byť.

Na čo si dávať pozor pri predikcii?

(1) Nesprávny model.
(2) Predikujeme pre hodnoty $x_0$ výrazne iné ako tie v našej dátovej vzorke.
(3) Externá validita : Predikujeme pre porozovania z inej populácie. Urobíme experiment na študentoch a budeme na základe toho tvrdiť niečo o celej populácii.
(4) Overfitting: Model výborne vysvetľuje naše dáta ale veľmi zle predikuje. Je príliš (a zbytočne) zložitý.
(5) Extrémne pozorovania: v našej dátovej vzorke môžu ale nemusia byť. Vo financiách model funguje výborne v časoch stability ale nevie predpovedať krízu.

Našťastie existujú spôsoby ako čiastočne overiť či predikujeme dobre alebo nie.

## 4.2 Vysvetľovanie

Niekedy máme ambíciu sa dozvedieť niečo z dát, porozumieť niečomu, narozdiel 'len' od predikovania. V istých situáciách by sme dokonca chceli popísať kauzálny efekt nejakej premennej na inú premennú, napr. že fajčenie zvyšuje riziko rakoviny alebo že človek svojou činnosťou prispieva ku globálnemu otepľovaniu. Toto je veľmi ambiciózne a z dát môžeme vyčítať všeličo avšak definitívna odpoveď, že $X$ spôsobuje $Y$ to nebude. Aj keď nevieme získať definitívnu odpoveď (v zmysle matematického dôkazu) vo veľa prípadoch situácia nie je až taká zúfalá a lineárna regresia môže kvantitatívne podporiť expertný názor. Povieme si aké vlastnosti majú situácie, kedy je kauzalita parsimónnym a plauzibilným vysvetlením nejakého fenoménu.

Majme lineárny model s jedným prediktorom. Ak by sme natvrdo a hlúpo verili nášmu modelu

$$y = \beta_0 + \beta_1 x_1 + \epsilon,$$

potom by sme povedali, že

~~*zvýšenie $x_1$ o jednu jednotku zvýši $y$ v priemere o $\beta_1$ jednotiek.*~~

To je však dosť nepravdepodobné, že náš model je pravdivý a vieme, že ak pridáme do regresie ďalší prediktor $x_2$, tak odhad parametra $\beta_1$ v modeli

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon,$$

nielenže môže mať inú hodnotu ale dokonca aj znamienko. Okrem toho je nerozumné uvažovať, že zvýšenie nadmorskej výšky ostrova v Galapágoch zvýši počet živočíšnych druhov na tomto ostrove o $\beta_2$. Je to nezaujímavá úvaha. Skrátka vidíme, že vo vyššie položených ostrovoch je viacej/menej živočíšnych druhov, preinterpretúvanie tohoto odhadu neznámeho parametra $\beta_2$ je nepotrebné a nezrozumiteľné. V niektorých fyzikálnych aplikáciách majú však parametre lineárneho modelu nejakú reálnu interpretáciu.

Okrem toho, interpretácia parametra je podmienená modelom! Preto pre model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon,$$

hovoríme, že

*náš model predikuje, že zvýšenie $x_1$ o jednu jednotku zvýši $y$ v priemere o $\beta_1$ jednotiek, za predpokladu, že sa nezmení hodnota $x_2$.*

Častokrát by však zmena $x_1$ zmenila $x_2$. Ak napríklad porovnávam mužov ($x_1 = 1$) a ženy ($x_1 = 0$) a mám fixovanú hodnotu výška v cm ($x_2 = 170$). Tak hodnota $\beta_1$ porovnáva nízkeho muža s vysokou ženou.

Interpretácia záleží aj od transformácie. Ak máme model

$$\log(\texttt{mzda}) = \beta_0 + \beta_1 \texttt{vzdelanie} + \beta_2 \texttt{vek} + \epsilon$$

tak $\beta_1$ interpretujeme ako:

*náš model predikuje, že pre rovnako starých ľudí, nárast počtu rokov vzdelania o 1 rok zvýši v priemere mzdu o $\beta_1 \cdot 100\%$ percent.*

Je rozumné používať jemnejší jazyk.

## 4.3 Interpretácia koeficientov pri logaritmickej transformácii

- $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

  $[x_1 \to x_1 + \delta] \implies [y \to y + \beta_1 \delta]$

  *náš model predikuje, že zvýšenie $x_1$ o jednu jednotku zvýši $y$ v priemere o $\beta_1$ jednotiek, za predpokladu, že sa nezmení hodnota $x_2$.*

- $$\log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$
  $[x_1 \to x_1 + \delta] \implies [y \to \exp(\beta_0 + \beta_1(x_1 + \delta) + \beta_2 x_2 + \epsilon)) = y \cdot exp(\beta_1 \delta) \approx y(1 + \beta_1 \delta)]$
  *náš model predikuje, že zvýšenie $x_1$ o jednu jednotku zvýši $y$ v priemere o približne $\beta_1 \cdot 100\%$, za predpokladu, že sa nezmení hodnota $x_2$.*
- $$y = \beta_0 + \beta_1 \log(x_1) + \beta_2 x_2 + \epsilon$$
  $[x_1 \to x_1 \cdot (1 + \delta)] \implies [y \to y + \beta_1 \log(1 + \delta) \approx y + \beta_1 \delta]$
  *náš model predikuje, že zvýšenie $x_1$ o 1% zvýši $y$ v priemere o približne $\beta_1/100$ jednotiek, za predpokladu, že sa nezmení hodnota $x_2$.*
- $$\log(y) = \beta_0 + \beta_1 \log(x_1) + \beta_2 x_2 + \epsilon$$
  $[x_1 \to x_1 \cdot (1 + \delta)] \implies [y \to y \cdot \exp(\beta_1 \log(1 + \delta)) = y \cdot (1 + \delta)^{\beta_1} \approx y \cdot (1 + \beta_1 \delta)]$
  *náš model predikuje, že zvýšenie $x_1$ o 1% zvýši $y$ v priemere o približne $\beta_1$%, za predpokladu, že sa nezmení hodnota $x_2$.*

## 4.4 Kauzalita

To čo je kauzálny efekt a čo nie je, je veľmi zložitá otázka. Existujú rôzne pohľady na kauzalitu v rôznych vedách. V štatistike je zlatým štandardom randomizovaný experiment. Ak chcem vedieť či nejaká intervencia má efekt na odozvu $y$, subjektom náhodne priradím či dostanú alebo nedostanú intervenciu. Potom sa pozriem či sú tieto dve skupiny dostatočne podobné, teda či je moja vzorka vybalancovaná. Ak áno, tak namerané rozdiely v odozve vysvetľujeme intervenciou.

Nech pacient $i$ dostane liečbu $T = 1$ alebo nedostane liečbu $T = 0$ a nech $y_i^T$ označuje jeho potenciálnu odozvu. Potom $\delta_i = y_i^1 - y_i^0$, je individuálny kauzálny efekt. Problémom je, že pre jedného pacienta pozorujem alebo $y_i^1$ alebo $y_i^0$ avšak nie obe naraz.

Tu je problém aj s definíciou kauzality pokiaľ nevieme intervenciu kontrolovať. Vieme si ľahko predstaviť efekt liečby, pretože si vieme predstaviť, že pacient môže a nemusí dostať liečbu. Ak by sme však chceli zistiť vplyv pohlavia na výšku mzdy, tu to už ide ťažšie. Mnohí výskumníci sú názoru, že *'No causation without manipulation'*, teda, že bez toho aby sme vedeli manipulovať premennú ktorej efekt skúmame nemôžeme identifikovať kauzalitu.

### 4.4.1 Prirodzený experiment

Predstavme si, že chceme skúmať efekt zavedenia minimálnej mzdy na zamestnanosť. Tu nikdy nebudeme mať experiment, že niekomu zvýšime minimálnu mzdu a niekomu zasa nie. Môžeme mať však čosi podobné čomu ekonómovia hovoria *prirodzený experiment*, teda môže sa stať, že budeme mať dva veľmi podobné štáty, kým v jednom sa zmení minimálna mzda a v druhom nie. Príklady prirodzených experimentov:

- V 2012 v ČR bol po tom ako viacero ľudí umrelo na otravu metanolom zakázaný predaj alkoholu na dva týždne. Toto umožnuje skúmať ako vplýva zákaz alkoholu na počet dopravných nehôd.
- V 1854 prepukla v Londýne epidémia cholery. Snow ukázal, že epicentrá cholery boli v miestach, ktorým bola dodávaná voda kontaminovaná splaškami. To či ľudia mali alebo nemali kontaminovanú vodu nemal nik pod kontrolou a tento prípad je preto prirodzeným experimentom.
- Ekonómovia chceli ukázať efekt veľkosti rodiny na mzdu matky. Tu nie je jasný smer kauzality, nakoľko matky môžu odložiť plánovanie rodiny ak sa im kariérne darí. Ukazuje sa však, že rodiny s dvomi chlapcami alebo dvomi dievčatami majú väčšiu pravdepodobnosť, že budú mať ďalšie dieťa. Tento fakt prekvapivo nezávisí od bohatstva, vierovyznania alebo štátnej príslušnosti. Preto môžeme porovnávať matky, ktoré majú dve deti rovnakého pohlavia a ktoré majú chlapca a dievča (pohlavie narodeného dieťaťa je náhodné, toto je preto prirodzený experiment) a na základe toho odhadnúť kauzálny efekt tretieho dieťaťa na mzdu matky.

- Počas vojny vo Vietname boli mladí muži draftovaní do vojny nasledovným spôsobom. Vytiahli si náhodné číslo a ľudia s vyšším čislom mali väčšiu pravdepodobnosť, že budú odrukovaní na vojnu. Teda máme prirodzený experiement a môžeme porovnávať mužov s väčším a menším vyžrebovaným číslom a určiť kauzálny efekt účasti vo vojne na mzdu.

### 4.4.2 Matching vs. Covariate Adjustment

Už sme si hovorili, že parametre interpretujeme vzhľadom na model. Môže sa však stať, že nejaký dôležitý prediktor nepozorujeme a tento dôležitý prediktor je korelovaný s prediktorom, ktorého efekt na odozvu nás zaujíma. Ak je tomu tak, naše odhady budú nedôveryhodné. Predstavme si, že vysvetľujeme rast miezd pomocou vzdelania.

$$\log(\texttt{mzda}) = \beta_0 + \beta_1 \texttt{vzdelanie} + \beta_2 \texttt{vek} + \epsilon$$

Zaujíma nás ako veľmi sa oplatí študovať, teda aký je efekt ďalšieho roka štúdia na nárast mzdy. V tomto modeli nám však čosi chýba. Ak by sme vedeli urobiť experiment a donútiť ľudí študovať toľko koľko im povieme, potom by bolo všetko fajn. Takto tomu však nie je, ľudia sa totiž rozhodujú koľko budú študovať a ľudia, ktorí sa rozhodli študovať sú iní ako tí, ktorí sa rozhodli neštudovať. Je uveriteľné predpokladať, že ľudia, ktorí sa rozhodnú dlhšie študovať sú v priemere šikovnejší ako tí, ktorí sa rozhodli študovať menej. Avšak táto "šikovnosť"ľudí je nepozorovaná a je korelovaná s nárastom mzdy. Ak je teda korektný model

$$\log(\texttt{mzda}) = \beta_0^* + \beta_1^* \texttt{vzdelanie} + \beta_2^* \texttt{vek} + \beta_3^* \texttt{ability} + \epsilon$$

a zároveň

$$\texttt{ability} = \gamma_0 + \gamma_1 \texttt{vzdelanie} + \epsilon'$$

teda náš pôvodný model

$$\begin{aligned}\log(\texttt{mzda}) &= \beta_0^* + \beta_1^* \texttt{vzdelanie} + \beta_2^* \texttt{vek} \\ &+\beta_3^*(\gamma_0 + \gamma_1 \texttt{vzdelanie} + \epsilon') + \epsilon \\ \log(\texttt{mzda}) &= (\beta_0^* + \beta_3^*\gamma_0) + (\beta_1^* + \beta_3^*\gamma_0)\texttt{vzdelanie} \\ &+\beta_2^* \texttt{vek} + (\beta_3^*\epsilon' + \epsilon)\end{aligned}$$

teda to čo nám dá náš jednoduchší model ($\beta_1 = \beta_1^* + \beta_3^*\gamma_0$) je vlastne zmiešaný efekt vzdelania a šikovnosti. Nuž ale keď chceme peniaze na reformu školstva, tak nás zaujíma efekt vzdelania a nie nejaký zmiešaný efekt. Šikovnosť, teda `ability` je teda nepozorovaný *confounder* (mätúca premenná alebo fičúria). Tomuto problému sa hovorí aj *ommited variable bias,* lebo náš estimátor koeficientu pri vzdelaní je vychýlený. Chcelo by to premennú ktorá by zachytávala túto šikovnosť, napríklad iq kvocient, aj keď táto miera môže byť spochybnená. Tákuto premennú však máme aj tak málokrát v dispozícii.

Teda ak si subjekty v našej dátovej vzorke nevyberajú prediktor, ktorého efekt na odozvu nás zaujíma, náhodne ale tak, že je korelovaný s nejakým iným dôležitým prediktorom, najjednoduchšie je ho pridať do regresie ak sa to dá. Teda v regresii by sme mali mať všetky dôležité prediktory, a ak aj nejaké nemáme, tak nech sú radšej nekorelované s tými čo máme v regresii. Pridanie mätúcej premennej sa nazýva **covariate adjustment**.

V prípade, že nás zaujíma efekt nejakej intervencie- binárnej premennej (dostal liek/nedostal liek, zavedieme/nezavedieme daňovú reformu,...), ale domnievame sa, že existuje mnoho mätúcich faktorov, korelovaných s intervenciou, napríklad starší ľudia sú náchylnejší užívať nejaké lieky. Mohli by sme dať vek do regresie ako ďalší prediktor, ale tým pádom robíme akýsi parametrický predpoklad, nie je však jasné či vek vstupuje do regresie lineárne. **Matching:** Čo môžeme urobiť namiesto toho je, že napárujeme subjekty podľa veku, z našeho datesetu vytvoríme dvojice podobne starých ľudí. Ak je mätúcich premenných viacej, tak môžeme urobiť také párovanie aby boli subjekty čo najpodobnejšie čo sa týka veku, rasy, pohlavia a náboženského vyznania. Čo to znamená najpodobnejšie? Vhodné párovanie je závislé od konkrétnej aplikácie.

Matching alebo covariate adjustment? Záleží. Matching je zložitejší na výpočet a nie je priamočiare aký typ matchingu si vybrať (existuje riziko, že výskumník si vyberie taký matching, sby dostal taký výsledok aký sa mu páči). Covariate adjustment je priamočiary ale predpokladá, že poznáme akým spôsobom ovplyvňuje mätúca premenná odozvu $y$.

### 4.4.3 Kvalitatívna podpora kauzality

Čím viacej sú splnené tieto podmienky, tým ľahšie sa nám hovorí o kauzalite.

- Efekt je výrazný. Nie, že je štatisticky významný ale že je veľký z praktického hľadiska.
- Efekt je konzistentný. Podobný efekt namerali pre iné subjekty, v inom štáte za iných podmienok.
- Efekt je špecifický. Skúmaný faktor ovplyvňuje len skúmanú odozvu $y$ a nie 10 ďalších vecí.
- Efekt rešpektuje časovú následnosť. Ak malo $x$ spôsobiť $y$, tak $x$ muselo nastať skôr.
- Efekt je monotónny. Viacej $x$ spôsobuje výraznejšiu zmenu v $y$.
- Efekt je hodnoverný. Experti v danej oblasti poznajú mechanizmus pomocou ktorého mohlo $x$ spôsobiť $y$ a tento mechanizmus je hodnoverný.
- Efekt je potvrdený experimentom.

# 5 Diagnostika

Každý model je založený na **predpokladoch**, tie môžu byť rozumné alebo menej rozumné. Po tom ako odhadneme model je dobré skontrolovať ako veľmi adekvátne tieto predpoklady sú. Predpoklady, ktoré robíme:

- Predpokladáme, že naše chyby sú $\epsilon \sim N(0, \sigma^2 I)$, teda
  - chyby sú nezávislé
  - chyby sú normálne rozdelené
  - chyby majú rovnakú varianciu
- Neexistujú pozorovania, ktoré by náš model nepopisoval. Náš model vie byť do veľkej miery ovplynený zopár nezvyčajnými pozorovaniami.
- Predpokladáme $y = X\beta + \epsilon$

Detekovať problém môžeme graficky alebo pomocou štatistického testu. Oba spôsoby majú svoje výhody aj nevýhody. Grafická detekcia (t.j. pomocou obrázku) je flexibilná ale vyžaduje úsudok. Na druhú stranu detekcia formou štatistického testu je presne stavaná na úzke použitie, čo je aj dobré aj zlé. Pomocou diagnostických nájstrojov môžeme zistiť ako model vylepšiť, čo je užitočné.

## 5.1 Predpoklady týkajúce sa chýb

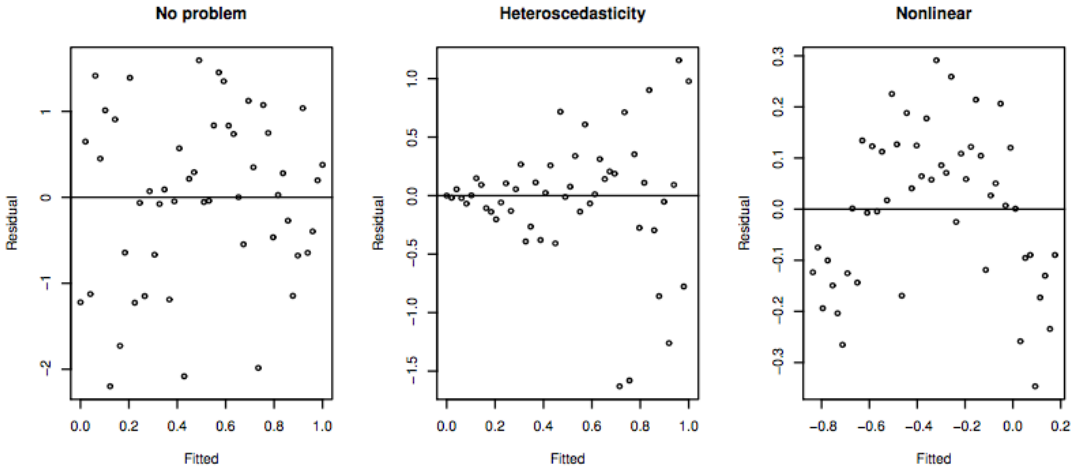
Skutočné errory $\epsilon$ nepozorujeme ani nikdy pozorovať nebudeme, vieme ich len odhadnúť pomocou reziduí $\hat{\epsilon}$. Pripomeňme, že $\hat{y} = X\hat{\beta} = X(X^TX)^{-1}X^Ty = Hy$. Pre reziduá máme

$$\hat{\epsilon} = y - \hat{y} = (I - H)y = (I - H)X\beta + (I - H)\epsilon = (I - H)\epsilon.$$

Teda $var(\hat{\epsilon}) = var((I - H)\epsilon) = (I - H)\sigma^2$, a vidíme, že residuály nemusia mať rovnakú varianciu a nemusia byť nekorelované. Tento rozdiel je však zvyčajne malý a preto robíme diagnostiku na reziduách.

### 5.1.1 Konštantná variancia

Chceme zistiť, či majú reziduá konštantnú varianciu. Môžeme si ich nakresliť, niektoré budú malé iné veľké ale samotnej variancii nám to nepovie nič. Potrebujeme zistiť, či náhodou nie sú reziduá systematicky väčšie alebo menšie pri niektorých prediktoroch alebo nafitovaných hodnotách odozvy. Teda či náhodou nie sú funkciou čohosi = nie sú konštantné.




Obr. 9: Graf reziduálov vs. fitovanej odozvy. (Zdroj: [Far14])

Môžeme pozorovať oblak bodov, to znamená nevidíme žiaden systematický vzťah medzi reziduálmi a $\hat{y}$ čo je fajn. Môžeme pozorovať, že variancia reziduálov sa mení s narastajúcimi hodnotami $\hat{y}$, to vrhá

tieň podozrenia na predpoklad konštantnej variancie alebo môžeme vidieť akúsi nelineárnu závislosť, čo naznačuje, že štrukturálna časť modelu ($y = X\beta + \epsilon$) nie je vhodne zvolená. Ak sa nám zdá, že oblak bodov je dostatočne rovnomerný môžeme zdvojiť rozlíšenie a namiesto $\hat{\epsilon}$ zobraziť $\sqrt{|\hat{\epsilon}|}$.

Ako štatistický test môžeme urobiť regresiu, kde vysvetľujeme $\sqrt{|\hat{\epsilon}|}$ pomocou napríklad $\hat{y}$. Teda testujeme, či *lineárny* vzťah medzi týmito dvomi veličinami je štatisticky významný. Takýto test je dosť špecifický a teda nemusí detekovať nelineárnu závislosť medzi týmito dvomi veličinami.

Okrem obrázka $\hat{y}$ voči $\hat{\epsilon}$ nás môže zaujímať, ako sa mení/nemení variancia reziduí s meniacimi sa prediktormi $x_i$, či už s tými použitými v regresii alebo s tými, čo nie sú použité v regresii.

Obrázok vs. test? Výhoda obrázku je, že môžeme objaviť štruktúru o ktorej by sme sa inak nedozvedeli. Na to aby sme z obrázku mohli detekovať nejaký problém s chybami treba skúsenosť. Pomocou počítača je však ľahké sa "natrénovať".

Čo môžeme urobiť, keď uvidíme v grafe nekonštantnú varianciu reziduí alebo nelineárnu závislosť??

- pretransformovať závislú premennú (pozor, lebo tým zmeníme distribúciu $Y$)
- pridať alebo pretransformovať prediktory (asi najlepšia rada)
- (ak len nekonštantná variancia) zmeniť estimátor na efektívnejší (o tomto viac neskôr)

Ako pretransformovať $y$ tak aby sme mali konštantnú varianciu? Uvažujme transformáciu $h(.)$, ktorú zvolíme tak aby $var(h(y))$ bola konštantná:

$$h(y) = h(E(y)) + (y - E(y))h'(E(y)) + ...$$

$$var(h(y)) = 0 + (h'(E(y)))^2 \, var(y) + ...$$

Teda potrebujeme aby $h'(E(y))$ bola proporciálna $(var(y))^{-1/2}$.

Ak je variancia funkciou strednej hodnoty $var(y) = g(\mu)$, potrebujeme

$$(h'(\mu))^2 \, g(\mu) = const.$$

preto

$$h(\mu) = \int \frac{d\mu}{\sqrt{g(\mu)}}.$$

Preto ak $var(y)$ rastie s druhou mocninou $E(y)$, teda ak $var(y) = var(\epsilon) \propto (E(y))^2$ potom $h(y) = \log(y)$ zastabilizuje varianciu a ak $var(\epsilon) \propto E(y)$, potom $h(y) = \sqrt{y}$ zastabilizuje varianciu. Odozva typu "počet"môže byť často modelovaná Poissonovým rozdelením, pre ktoré platí $E(y) = var(y)$ a teda $h(y) = \sqrt{y}$ je vhodné. Vždy však môžeme vyskúšať inú transformáciu a pozrieť sa ako vhodná je.

### 5.1.2 Normalita

Väčšina testov je založená na normalite reziduí. Normalitu vieme graficky posúdiť pomocou QQ-plotu, ktorý je preferovaný oproti histogramu alebo jadrovému odhadu hustoty, pretože pri histograme obrázok záleží od šírky stĺpikov a pri odhade hustoty od veľkosti vyhladzovacieho parametra. Na detekciu treba vidieť mnoho QQ-plotov aby sa človek naučil ako vyzerajú rôzne porušenia normality (naklonenosť, ťažké chvosty, ľahké chvosty).

Ak chyby nie sú normálne, potom odhad metódou MNŠ už nemusí byť nutne optimálny. Stále bude najlepší spomedzi lineárnych nevychýlených odhadov ale nelineárne odhady môžu byť efektívnejšie. Ak máme veľkú dátovú vzorku, nemusíme sa moc starať o normalitu, ktorá bude zabezpečená vďaka centrálnej limitnej vety.

Reakcia na nenormálne chyby záleží aj od toho akým spôsobom sú nenormálne:

- Pre ľahkochvosté distribúcie sú zväčša dôsledky mierne a tento problém môžeme ignorovať.
- Pre naklonené distribúcie častokrát pomôže transformácia $y$.
- Pre ťažkochvosté distribúcie môžeme napríklad použiť bootstrap.

A čo formálne testy normality? Nie sú moc užitočné: pre malé dátové vzorky sú príliš slabé na to aby detekovali porušenie normality a pre veľké dátové vzorky zamietnu aj veľmi malú výchylku od normality a kedže naše dáta nie sú perfektne normálne, toto nie je užitočné. Preto sa formálne testy normality odporúča používať len v spojitosti s grafickými metódami.

### 5.1.3 Korelovanosť chýb

Korelovanosť chýb sa ťazko kontroluje, lebo existuje veľmi veľa rôznych spôsobov ako môžu byť chyby korelované, resp. čo môže túto koreláciu spôsobovať.

Chyby môžu byť korelované v čase a na detekciu takejto korelácie sa dá použiť Durbinov-Watsonov test. Jeho štatistika vyzerá nasledovne

$$DW = \frac{\sum_{n=2}^{n}(\hat{\epsilon}_i - \hat{\epsilon}_{i-1})^2}{\sum_{n=1}^{n}\hat{\epsilon}_i^2}.$$

Korelácia reziduí môže byť spôsobená aj chýbajúcimi prediktormi. Niekedy sa dá korelácie zbaviť zmenou štrukturálnej časti modelu ale inokedy ju musíme brať do úvahy a zostrojiť efektívnejší estimátor (o "zovšeobecnenej metóde najmenších štvorcov", ktorá berie do úvahy korelačnú štruktúru chýb neskôr neskôr).

## 5.2 Nezvyčajné pozorovania

Niektoré pozorovania skrátka nemodelujeme dobre, voláme ich autlájeri (**outliers**). Potom sú ešte pozorovania, ktoré výrazne ovplyvňujú fit modelu, tieto sa nazývajú **vplyvné**. Potom sú ešte pozorovania, ktoré majú potenciál výrazne ovplyvniť fit modelu ale nemusí tomu tak byť, tieto sa nazývajú **pákové**.

### 5.2.1 Pákové pozorovania

Diagonálny člen matice $H = X(X^TX)^{-1}X$, teda $h_i = H_{ii}$ sa nazýva páka (**leverage**). Variancia reziduí $var(\hat{\epsilon}_i) = \sigma^2(1 - h_i)$, teda reziduál pozorovania $i$, ktoré má veľkú hodnotu $h_i$ bude mať malú varianciu a preto $\hat{y}_i$ bude blízko $y_i$. Suma pák je rovná počtu parametrov $\sum_i h_i = p$ a preto je priemerná hodnota $p/n$. Poznamenajme, že pákovosť daného pozorovania vôbec nezávisí od $y$. Čo nás môže zaujímať je, že ktoré pozorovania sú dôležité, ktoré výrazne ovplyvňujú náš fit. Na to aby sme detekovali výrazne veľké páky používame half-normal plot, teda QQ plot normálneho rozdelenia s dvojitým rozlíšením. Pozeráme na vzťah medzi $h_i$ a $|u|$, kde $u \sim N(0, 1)$. Half-normal plot pre rad $x_i$ skonštruujeme nasledovne:

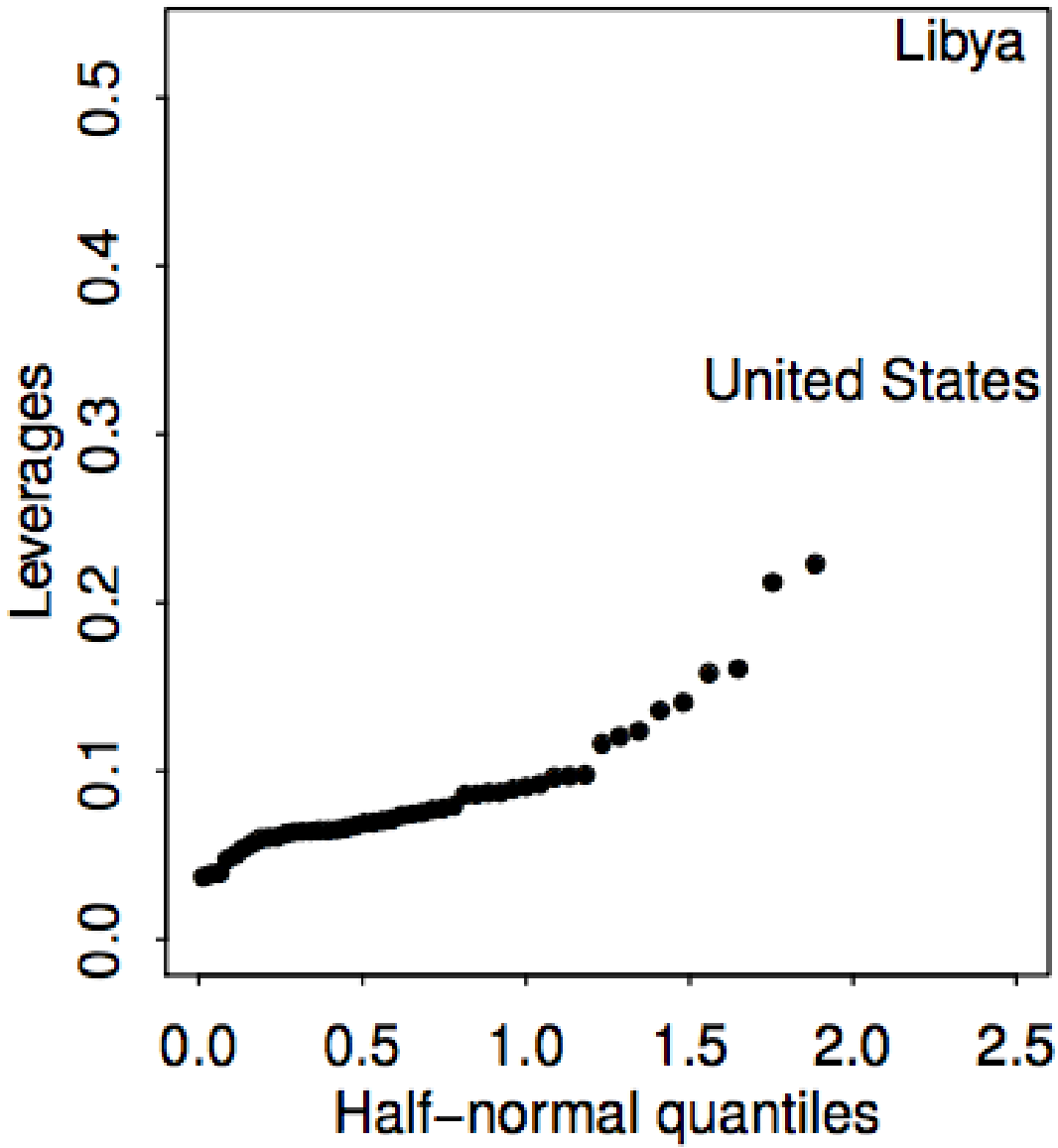
- Usporiadame dáta: $x_{[1]} \leq x_{[2]} \leq x_{[3]} \leq ...x_{[n]}$
- Vypočítame $u_i = \Phi^{-1}\left(\frac{n+i}{2n+1}\right)$
- Nakreslíme $x_{[i]}$ oproti $u_i$.

Nečakáme, že naše páky budú normálne rozdelené ale dávame si pozor na pozorovania, ktoré sú výrazne nad čiarou lineárneho trendu.

Páky $h_i$ sú užitočné aj na reškálovanie reziduí. Nakoľko vieme, že variancia reziduí $\hat{\epsilon}_i$ nemusí byť konštantá aj keď variancia chýb $\epsilon_i$ je, môžeme sa zbaviť tohoto problému ak podelíme reziduá odmocninou z odhadu jej variancie:

$$r_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}\sqrt{1 - h_i}}.$$

Takéto reziduá sa nazývajú **štandardizované** a sú preferované na diagnostiku. Neochránia nás však pred potenciálnou nekonštantnosťou variancie chýb. Veľké rozdiely v používaní môžeme badať ak je v dátovej vzorke veľa pákových pozorovaní.




Obr. 10: Detekcia pákových pozorovaní pomocou half-normal qq-plotu. (Zdroj: [Far14])

### 5.2.2 Outlieri

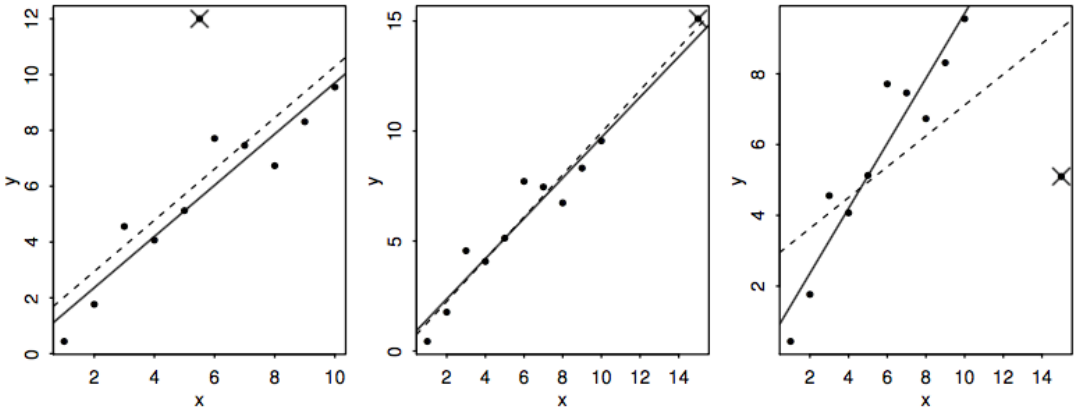
Outlier je pozorovanie, ktoré nie je dobre popísané modelom. Test na detekciu outlierov je užitočný lebo rozlišuje bod, ktorý nie je dobre popísaný modelom (skutočný outlier) a bod, ktorému zodpovedá veľký rezdiuál ale nie významne.

Na detekciu toho, či je nejaký bod taký nebezpečný outlier ako v ľavom obrázku alebo nie, ho dáme preč a odhadneme model. Keď dáme $i$-te pozorovanie preč, potom označme zodpovedajúce estimátory ako $\hat{\beta}_{(i)}$ a $\hat{\sigma}^2_{(i)}$. Potom nech $\hat{y}_{(i)} = x_i^T \hat{\beta}_{(i)}$ a pozorovanie $i$ je outlier ak je hodnota $\hat{y}_{(i)} - y_i$ veľká. Túto hodnotu potrebujeme ešte vhodne naškálovať tak, aby sme vedeli posúdiť variabilitu. Nakoľko $\hat{var}(y_i - \hat{y}_{(i)}) = \hat{\sigma}^2_{(i)}(1 + x_i^T (X_{(i)}^T X_{(i)})^{-1} x_i)$, zadefinujeme si studentizované reziduá ako

$$t_i = \frac{y_i - \hat{y}_{(i)}}{\hat{\sigma}_{(i)}(1 + x_i^T (X_{(i)}^T X_{(i)})^{-1} x_i)^{1/2}} \sim t_{n-p-1}.$$

Toto, zdá sa, vyžaduje odhadnút $n$ lineárnych modelov, našťastie však existuje vzťah medzi $t_i$ a $r_i$, teda štandardizovanými reziduami, spomenutými skorej:

$$t_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1 - h_i}} = r_i \left( \frac{n - p - 1}{n - p - r_i^2} \right)^{1/2}.$$




Obr. 11: Vľavo - outlier, ktorý nezmenil smernicu krivky. V strede, pákový bod, ktorý neovplyvnil fit. Vpravo, outlier, ktorý výrazne ovplyvnil fit. Zahrnutie tohoto pozorovanie výrazne zvýšilo reziduály aj ostatných pozorovaní. (Zdroj: [Far14])

Môžeme ísť cez všetky pozorovania a testovať jedno po druhom, či je hodnota $t_i$ veľká, teda testujeme, či je $i$-te pozorovanie outlier ($H_A$) alebo nie je ($H_0$). Tu je však problém s tým, že v 5% prípadov sa budeme mýliť, takže v piatich prípadoch zo 100 nájdeme outliera aj keď tam žiaden nie je. Potrebujeme nejak vziať do úvahy fakt, že testujeme naraz veľa hypotéz. Ak by sme chceli test na hladine významnosti $\alpha$. Chceli by sme aby pravdepodobnosť chyby prvého druhu bola ohraničená aj keď testujeme veľakrát. $P(\text{vsetky testy nezamietnu } H_0) = 1 - P(\text{jeden alebo viac testov zamietne H}_0) \geq 1 - \sum_i P(\text{test i zamietne H}_0) = 1 - n\alpha$ Takže keď robíme $n$-testov potrebujeme nastaviť chybu prvého druhu na menšiu ako $\alpha/n$ a potom budeme mať zaručené, že celkový test bude nesprávne zamietať nulovú hypotézu menej ako v $\alpha \cdot 100\%$ prípadov. Tomuto triku sa hovorí *Bonferroniho korekcia*, problémom je, že takýto test je konzervatívny, takže detekuje menej outlierov ako by pri danej fixnej chybe prvého druhu mohol. Čím väčšie $n$, tým konzervatívnejší test je. Konzervatívnosť je spôsobená nerovnosťou vyššie, tá je rovnosťou len ak sú testy nezávislé, čo nie sú, lebo sú počítané na veľmi podobných datasetoch (v každom z nich odhodíme iné pozorovanie).

Veci, ktoré treba brať do úvahy

- Jeden outlier môže zamaskovať druhého.
- To či je nejaké pozorovanie outlierom alebo nie záleží aj od modelu, transformácie prediktora alebo odozvy môže zmeniť či je nejaké pozorovanie outlierom.
- Niekedy distribúcia chýb nie je normálna a preto môžeme očakávať väčšie reziduá.
- Jeden outlier zväčša nie je problém, problémom môže byť zhluk outlierov.

Čo s tým môžeme urobiť?

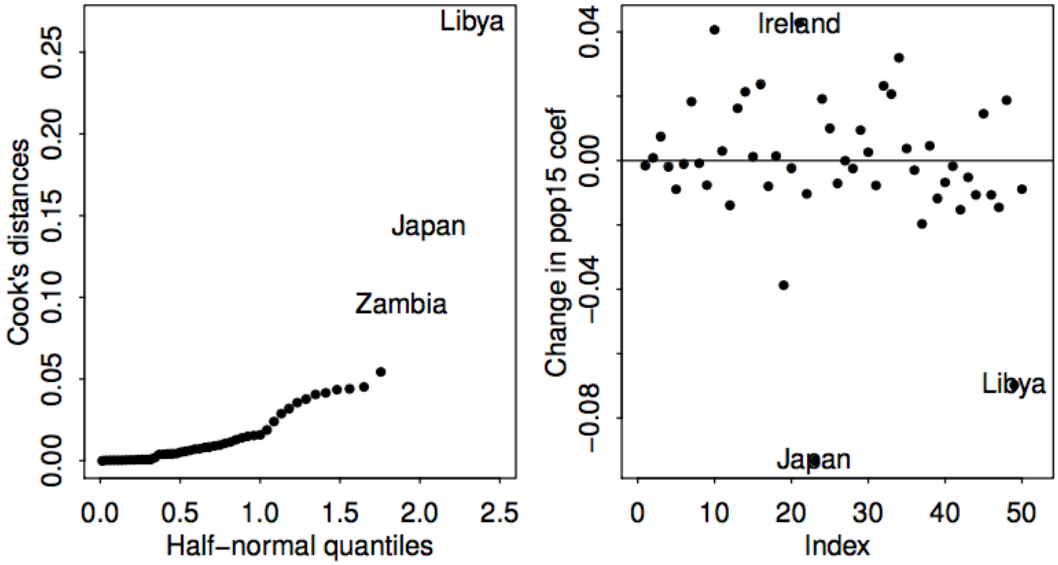
- Skontrolujeme, či dáta nie sú zle zadané, či nenastal napríklad preklep. Toto je veľmi časté.
- Porozmýšľame nad tým prečo tam ten outlier je, niekedy je outlier zaujímavý sám o sebe.
- Dáme ho preč a pozrieme sa ako sa model zmení. Toto môže spôsobiť, že náš model alebo parameter, ktorý nás zaujíma je štatisticky významný. Preto je dôležité to v analýze reportovať. Nereportovanie je považované za výrazne nevedecký prístup.
- Je veľmi nebezpečné vyhadzovať outlierov automaticky. Vďaka tomu sme objavili porušenie ozónovej vrstvy o niekoľko rokov neskôr ako sme mohli.

### 5.2.3 Vplyvné pozorovania

Vplyvné pozorovania sú také, ktoré výrazne ovplyvňujú fit modelu. Môžu ale nemusia byť outliermi. Ako ich detekovať? Môžeme sa pozrieť na zmenu $\hat{y} - \hat{y}_{(i)}$ ale to musíme pozerať $n$-krát na vektory dĺžky $n$. Trošku si to vieme zjednodušiť tým, že môžeme pozerať na zmeny $\hat{\beta}_{(i)}$, teda na $n \cdot p$ hodnôt. Existuje však Cookova štatistika, ktorá zhrnie všetku túto informáciu do jedného čísla:

$$D_i = \frac{(\hat{y} - \hat{y}_{(i)})^T(\hat{y} - \hat{y}_{(i)})}{p\hat{\sigma}^2} = \frac{1}{p} r_i^2 \frac{h_i}{1 - h_i}.$$

Použitím half-normal plotu môžeme identifikovať pozorvania, ktoré majú výrazne vyššiu Cookovu štatistiku ako ostatné. Užitočné je nazrieť ako veľmi sa mi menia odhadnuté parametre ak vynechám nejaké pozorovanie.




Obr. 12: Vľavo - half-normal plot Cookovych štatistík. Vpravo - citlivosť odhadnutého parametra na vynechanie $i$-teho pozorovania. (Zdroj: [Far14])

## 5.3 Štrukturálna časť modelu

Azda najpodstatnejšou časťou modelu je jeho štrukturálna časť, teda $E(y) = X\beta$. Pozeráme sa na reziduá a hľadamé nejaké systematické závislosti. Graf závislostí $\hat{\epsilon}$ voči $\hat{y}$ a $x_i$ nám pomohol detekovať nekonštantnosť variancie reziduí a tá sa dala vyriešiť transformáciou odozvy alebo prediktorov a táto nutne zmení aj štrukturálnu časť modelu.

Chceli by sme vedieť ako vyzerá efekt $x_i$ na $y_i$, či je vskutku lineárny alebo nie. Avšak $x_i$ je korelované s inými prediktormi, takže dostávame zmiešaný pohľad na vec. Chceli by sme izolovať efekt $x_i$ na $y_i$ v rámci modelu.

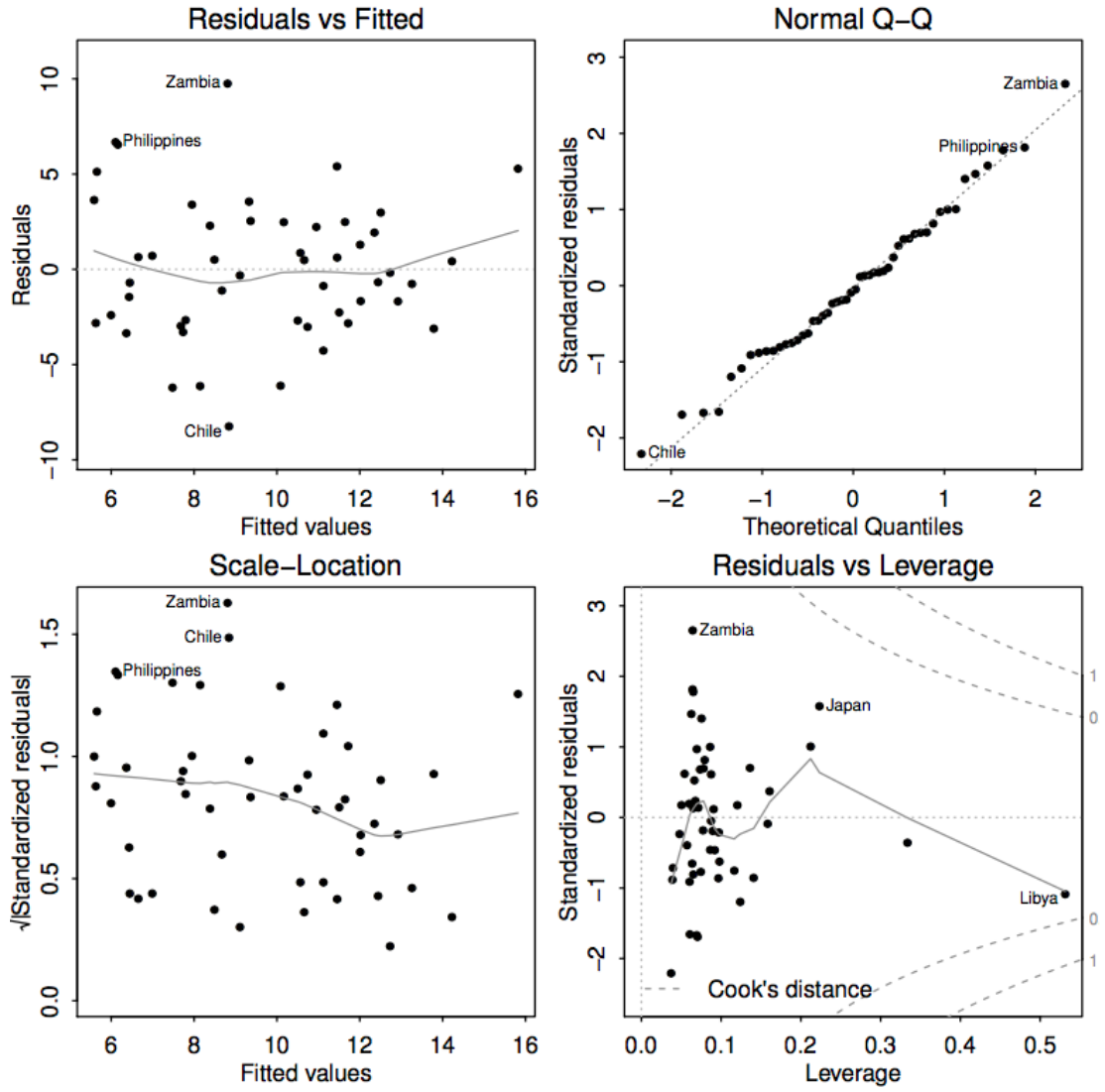
### 5.3.1 Parciálna regresia

- $\hat{\delta}$ dostaneme ako reziduá z regresie kde $y$ vyvetľujeme pomocou všetkých $x_j$ okrem $x_i$. Teda $\hat{\delta}$ zachytáva variáciu v $y$, ktorá je nevysvetlená ostatnými prediktormi a len čaká na to kým ju príde vysvetliť $x_i$.
- $\hat{\gamma}$ dostaneme ako reziduá z regresie kde $x_i$ vysvetľujeme pomocou ostatných $x_j$. Teda $\hat{\gamma}$ zachytáva variáciu $x_i$, ktorá je nevysvetlená ostatnými prediktormi a teda nám hovorí ako veľmi rozširuje prediktor $x_i$ priestor kde môžeme zobrazovať $y$.
- Zaujíma nás, či je závislosť $\hat{\delta}$ od $\hat{\gamma}$ lineárna. Ak nie je, zamyslíme sa nad vhodnou transformáciou. Na signifikantnosť lineárnej závislosti môžeme použiť signifikantnosť koeficientu $\alpha_1$ pri premennej $\hat{\gamma}$ v regresii $\hat{\delta} = \alpha_0 + \alpha_1\hat{\gamma} + \epsilon$.

### 5.3.2 Parciálne reziduá

Alternatívou k parciálnej regresii je graf parciálnych reziduí voči $x_i$. Parciálne reziduum pre prediktor $i$ je

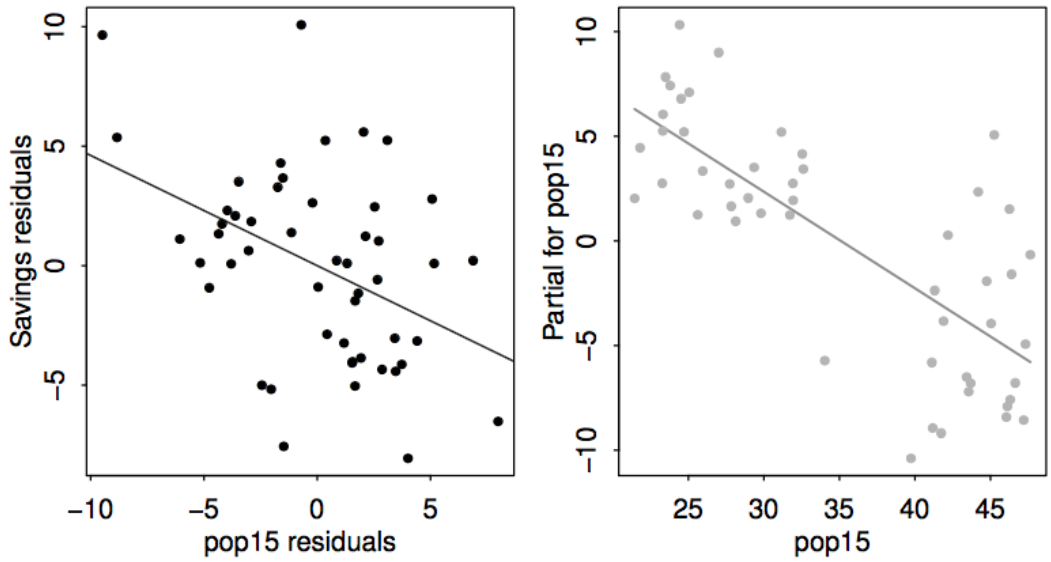
$$y - \sum_{j \neq i} x_j\hat{\beta}_j = \hat{y} + \hat{\epsilon} - \sum_{j \neq i} x_j\hat{\beta}_j = x_i\hat{\beta}_i + \hat{\epsilon}$$



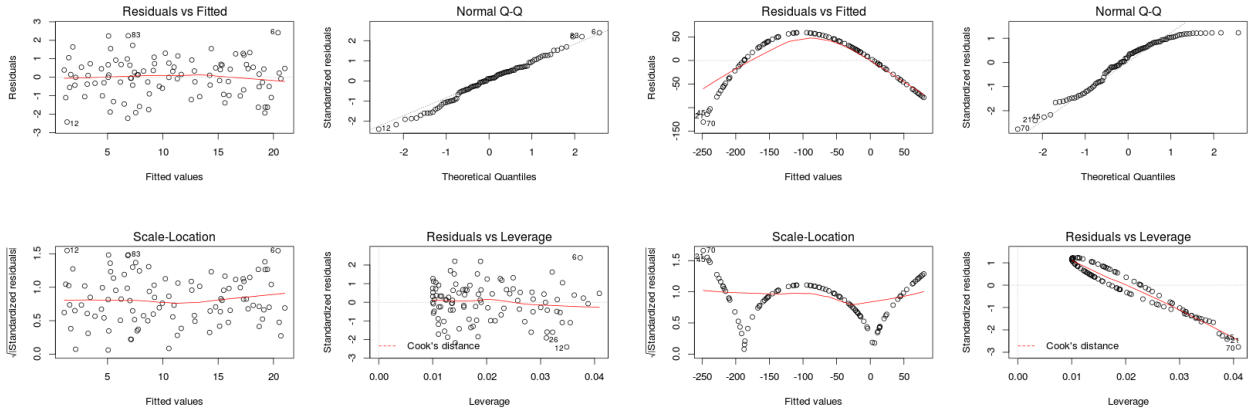

Obr. 13: (Sumár diagnostických grafov. Vľavo hore a dole: nevidíme žiadne výrazne funkčné závislosti, Vpravo hore: porozovania vyzerajú normálne. Vpravo dolu:Cookova vzdialenosť je funkciou pákovosti a štandardizovaných reziduí, preto môžeme zobraziť kontúry tejto funkcie. Vidíme, že Lýbiu môžeme považať za najvplyvnejšie pozorovanie. Zdroj: [Far14])

Teda vysvetľujeme čo zostalo z odozvy potom ako sme odobrali všetky ostatné prediktory. Tento graf je lepší na detekciu nelinearity ako parciálna regresia, ktorá je lepšia na detekciu outlierov.

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




Obr. 14: Vľavo- parciálna regresia, vpravo graf parciálnych reziduí voči $x_i$. Zdroj: [Far14])




```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. 15: Diagnostika: vľavo korektne špecifikovaný model, vpravo nie, zdroj: [glmb].

# 6 Problémy s prediktormi

V tejto kapitole sa budeme zaoberať tromi rôznymi typmi problémov, ktoré môžeme mať s prediktormi. Prvým problémom je ak sú prediktory pozorované s nejakou chybou. Zdroje tejto chyby môžu byť rôzne, napríklad nepresnosť meracieho prístroja. Potom to bude efekt preškálovania prediktorov, toto je zaujímavé pretože preškálovanie môže pomôcť s porozumením hodnôt odhadov parametrov. Rovnako nám to pomôže "férovo" porovnávať efekt rôznych prediktorov na odozvu.

## 6.1 Chyby v prediktoroch

Prečo sú premenné pozorované s chybou?

- nepresnosť meracieho prístroja (spôsobená napr. nekalibráciou)
- rozlíšenie meracieho prístroja
- chybné zaznamenanie spôsobené ľudským faktorom
- nekompletná definícia (meranie vykonané rôznym spôsobom)
- meniace sa podmienky prostredia

Napriek tomu, že budeme hovoriť o náhodných chybách v tom ako sú prediktory namerané, nič to nemení na tom, že na prediktory sa počas regresie pozeráme ako na fixné hodnoty. Označme pozorované hodnoty ako $y_i^O$ a $x_i^O$ (O ako observed) a skutočné ale nepozorované hodnoty ako $y_i^A$ a $x_i^A$ (A ako accurate), kde $y_i^O = y_i^A + \epsilon_i$ a $x_i^O = x_i^A + \delta_i$. Uvažujme situáciu, že vektory chýb $\epsilon$ a $\delta$ sú nezávislé a nech skutočné hodnoty sú v lineárnom vzťahu $y_i^A = \beta_0 + \beta_1 x_i^A$. Pre pozorované veličiny teda dostávame

$$y_i^O = \beta_0 + \beta_1 x_i^O + (\epsilon_i - \beta_1 \delta_i).$$

Chceli by sme odhadnúť parametre $\beta_0$ a $\beta_1$ pomocou metódy najmenších štvorcov. Nech $E(\epsilon_i) = E(\delta_i) = cov(\epsilon_i, x_i^A) = 0$ a nech $var(\epsilon_i) = \sigma_\epsilon^2$ a $var(\delta_i) = \sigma_\delta^2$. Nech naviac $\sigma_x^2 = \sum_{i=1}^n (x_i^A - \bar{x}^A)^2/n$ a $\sigma_{x\delta} = cov(x^A, \delta)$.

Nakoľko $\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})}$ dostávame $\lim_{n\to\infty} E(\hat{\beta}_1) = \beta_1 \frac{\sigma_x^2 + \sigma_{x\delta}}{\sigma_x^2 + \sigma_\delta^2 + 2\sigma_{x\delta}}$. Teda vo všeobecnosti je odhad MNŠ **vychýlený**.

(1) Ak neexistuje žiaden systematický vzťah medzi $x^A$ a $\delta$ a $\sigma_{x\delta} = 0$, potom

$$E(\hat{\beta}_1) = \beta_1 \frac{1}{1 + \frac{\sigma_\delta^2}{\sigma_x^2}},$$

teda odhad smernice lineárneho odhadu bude vychýlený smerom k nule, teda bude sčapenejší.

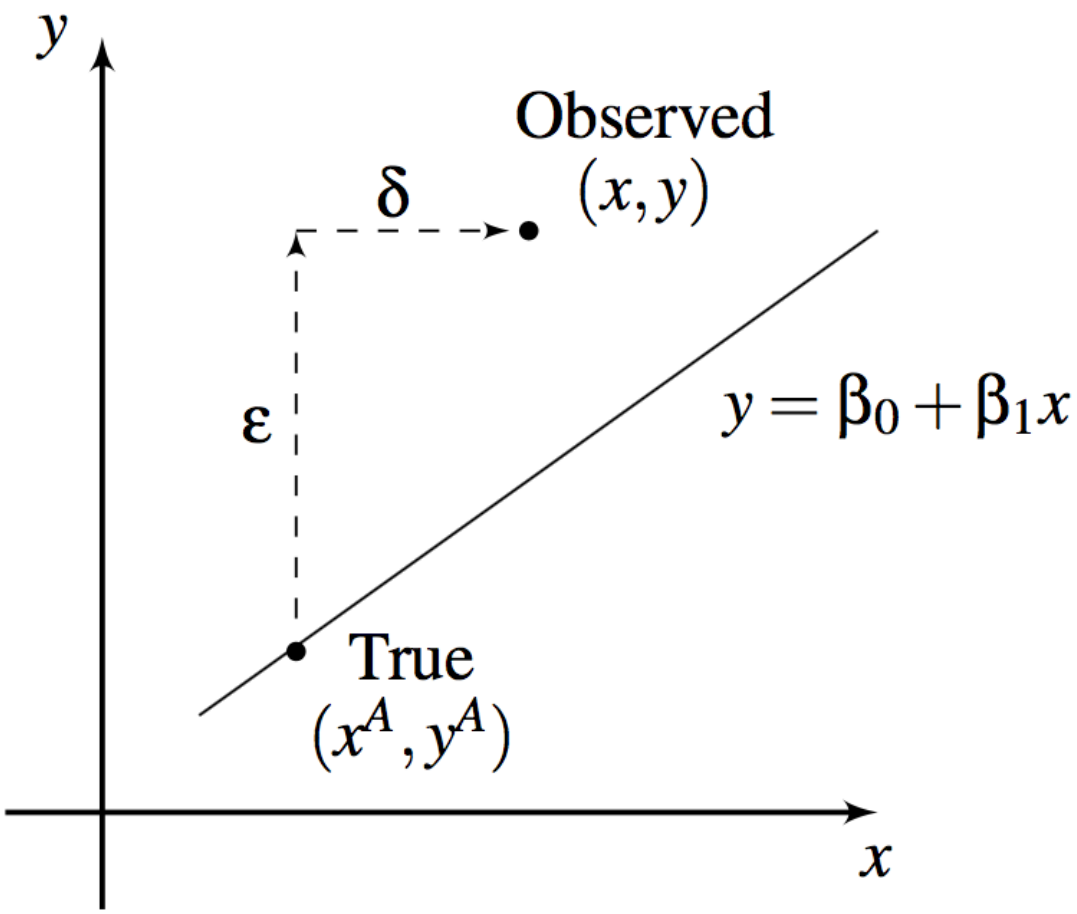
(2) Ak máme kontrolovaný experiment musíme odlíšiť dve situácie

(1) Existuje nejaké fixné $x_A$, my však meriame len $x^O$. Ak však budeme merať znova znova dostaneme nejaké $x^O$, avšak to pravdivé $x^A$ zostáva stále rovnaké.

(2) Fixujeme $x^O$, napríklad namiešame stále tú istú koncentráciu, aj keď tá skutočná koncentrácia $x^A$ bude každý raz iná. V tomto prípade máme $\sigma_{x\delta} = cov(x^O - \delta, \delta) = -\sigma_\delta^2$, teraz však dostaneme nevychýlený odhad lebo úlohy $x^A$ a $x^O$ sú vymenené.

Teda vidíme, že chyby v prediktoroch sú problémom. Náš estimátor dokonca ani nebude nevychýlený. No ale čo ak vieme aká je variancia chýb v prediktoroch? Nevieme nejak využiť túto informáciu, aby sme dostali nevychýlený odhad $\beta_1$?

Môžeme, a to nasledovne. Budeme pridávať šum s väčšou a väčšou varianciou a budeme sa pozerať ako sa mení odhad parametra $\beta_1$. Pre každé pridanie rôzne veľkého šumu vygenerujeme veľa datasetov a odhadneme priemernú hodnotu $\beta_1$. Dostaneme tak krivku kde veľkosť parametra je funkciou šumu. Z tejto krivky extrapolujeme do situácie kedy nie je v prediktore žiaden šum.




Obr. 16: Chyby v meraniach (Zdroj: [Far14]).

## 6.2 Zmena škály prediktorov

Niekedy sme v situácii, že paramter pri nejakom prediktore je veľmi malý, napr. 0.00000000245, človek sa ľahko stratí v toľkých nulách. Preto je užitočné, preškálovať prediktor, napríklad z príjmu v dolároch urobiť príjem v miliónoch dolárov. Inokedy, ak sú prediktory výrazne inak škálované tak odhadovanie môže viesť k numerickým nestabilitám.

Čo sa stane ak $x_i \to \frac{x_i+a}{b}$? Testy založené na t-štatistike a F-štatistike ako aj $\hat{\sigma}^2$ a $R^2$ sa nezmenia a $\hat{\beta}_i \to b\hat{\beta}_i$

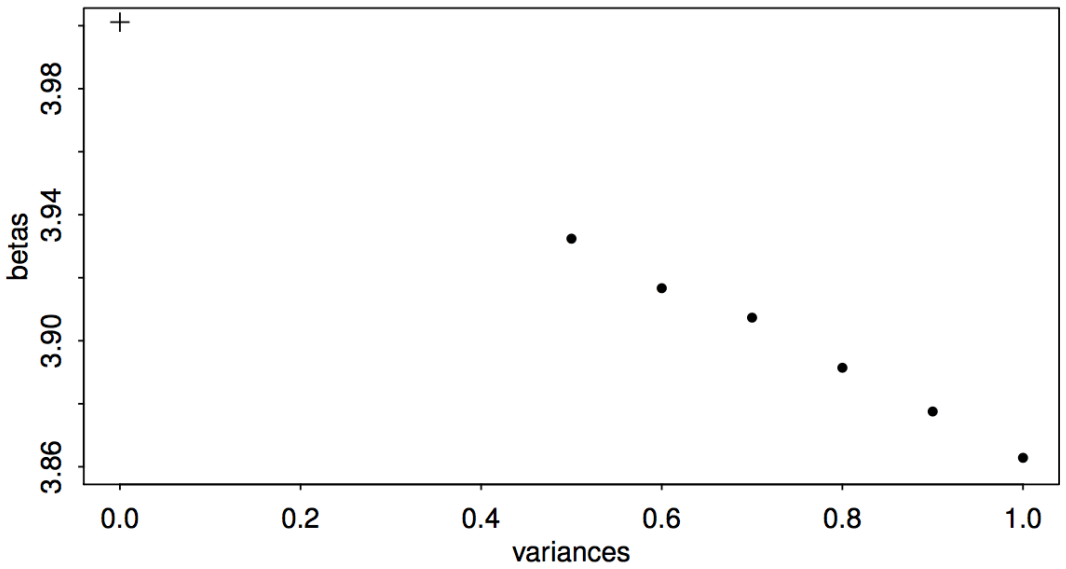
Čo sa stane ak $y_i \to \frac{y_i+a}{b}$? Testy založené na t-štatistike a F-štatistike ako aj $R^2$ sa nezmenia a $\hat{\beta}_i \to b\hat{\beta}_i$ a $\hat{\sigma}^2 \to b^2\hat{\sigma}^2$.

Keď chceme porovnávať efekt prediktorov na našu odozvu je užitočné preškálovať si všetky prediktory tak, aby mali rovnakú strednú hodnotu 0 a varianciu 1, toto urobíme tak, že odčítame od premennej odhad strednej hodnoty a podelíme odhadom smerodajnej odchýlky. Takto porovnávame jablká s jablkami.

Ak však máme binárny prediktor, tak pri 50% - 50% rozložení núl a jednotiek má táto premenná smerodajnú odchýlku len 0.5, preto aby to bolo spravodlivé, podelím ostatné prediktory dvojnásobkom odhadu smerodajnej odchýlky.

## 6.3 Kolinearita v prediktoroch

Ak sú nejaké prediktory lineárme korelované, potom je aj matica $X^TX$ singulárna a neinvertovateľná a preto neexistuje jednoznačný - najmenšie štvorce minimalizujúci vektor parametrov. Ak sú prediktory úplne korelované, riešenie je zväčša jednoduché, a to odobrať lineárne závislé prediktory. Problémom však je aj keď sú prediktory blízke k úplne korelovaným, teda keď je ich korelácia blízka +1 alebo -1. Toto sa nazýva kolinearita a vedie k nepresným odhadom parametrov. Aj veľmi malá zmena $y$ vedie k




Obr. 17: Ako odhadnúť $\beta_1$? Skutočné $\beta_1$ zodpovedá situácii keď je variancia chýb nulová. Vieme, že náš prediktor je porozovaný s chybou, ktorej variancia je 0.1. Pridávame šum veľkosti 0.1 až 0.5 a kreslíme priemerné odhady $\beta_1$ z veľa simulácií. Cez tieto body môžeme odhaliť závislosť medzi úrovňou biasu a varianciou chýb. Toto použijeme na extrapoláciu do situácie keď je variancia chýb nulová (Zdroj: [Far14]).

výrazne rozdielnym odhadom parametrov. Ako zistiť či máme problém s kolinearitou?

(1) Pozrieme sa na korelačnú maticu $X$ a hľadáme čísla blízke +1 alebo -1.
(2) Ak vysvetľujeme regresor $x_i$ pomocou ostatných regresorov, miera lineárneho fitu $R_i^2$ blízka 1 indikuje problémy.
(3) Usporiadajme vlastné čísla matice $X^TX, \lambda_1 \geq \cdots \geq \lambda_p$. Číslo podmienenosti $\kappa = \sqrt{\frac{\lambda_1}{\lambda_p}} \geq 30$ indikuje problémy.

Efekt kolinearity vidíme keď sa pozrieme na varianciu odhadu parametra:

$$var(\hat{\beta}_j) = \sigma^2 \left( \frac{1}{1 - R_j^2} \right) \frac{1}{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2},$$

kde $\frac{1}{1-R_j^2}$ sa nazýva *variance inflation factor.*

Ak by sme mali možnosť nadizajnovať $X$, tak ortogonalitou prediktorov spôsobíme, že $R_i^2 = 0$ a snažíme sa maximalizovať $\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$, teda roztiahnuť $x_j$ čo najviac ako to ide. Túto druhú vlastnosť si zrejme uvedomoval už Thomas Mayer z úvodnej kapitoly.

Kolinearita prediktorov nevadí až tak predikcii ako odhadom parametrov. Aj keď pri kolinearite je priestor pokrytý stĺpcami matice $X$ menší než by sa zdalo, preto budú predikcie väčšou extrapoláciou a teda budú nepresnejšie, než by sa zdalo.

# 7 Nejaké problémy s chybami

Doteraz sme predpokladali, že v Gaussovskom lineárnom modeli je kovariančná matica chýb $\epsilon$ diagonálna matica (chyby sú nekorelované), ktorá má na diagonále rovnaké členy $\sigma^2$ (chyby sú homoskedastické):

$$var(\epsilon) = \sigma^2 I = \begin{pmatrix} \sigma^2 & 0 & ... & 0 \\ 0 & \sigma^2 & ... & 0 \\ \vdots & & \ddots & \\ 0 & 0 & ... & \sigma^2 \end{pmatrix}.$$

V prípade, že chyby sú závislé, môžeme použiť **zovšeobecnenú metódu najmenších štovrcov (GLS - Generalized Least Squares)**.

Ak sú chyby nezávislé ale majú rôznu varianciu, môžeme použiť **váženú metódu najmenších štvorcov (WLS - Weighted Least Squares)**.

Neskôr sa pozrieme na testovanie nedostatku fitu (lack of fit).

Ak chyby nie sú noromálne rozdelené, namiesto metódy najmenších štvorcov môžeme použiť **robustnú regresiu**.

## 7.1 Zovšeobecnená metóda najmenších štvorcov

Nech sú teraz v lineárnom modeli chyby závislé, a teda $var(\epsilon) = \sigma^2\Sigma$ a nech poznáme maticu závislostí $\Sigma$. Choleského dekompozíciou môžeme rozdeliť $\Sigma = SS^T$. Šikovnou transformáciou sa dostaneme do štandardného Gaussovského modelu s iid chybami.

$$\begin{aligned} y = & \quad X\beta + \epsilon \\ S^{-1}y = & \quad S^{-1}X\beta + S^{-1}\epsilon \\ y' = & \quad X'\beta + \epsilon' \end{aligned}$$

Variancia nových transformovaných chýb $\epsilon'$ je potom

$$var(\epsilon') = var(S^{-1}\epsilon) = S^{-1}var(\epsilon)S^{-T} = \sigma^2 I.$$

Teda aplikujeme MNŠ na $S^{-1}y$ a $S^{-1}X$. V tomto prípade minimalizujeme

$$(y' - X'\beta)^T(y' - X'\beta) = (y - X\beta)^T\Sigma^{-1}(y - X\beta),$$

čoho riešením je

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

Variancia tohoto odhadu je

$$var(\hat{\beta}) = \sigma^2(X^T\Sigma^{-1}X)^{-1}.$$

Praktickým problémom však je, že málokedy vieme $\Sigma$. Ak ho chceme odhadnúť, tak máme problém, pretože má veľa parametrov, a to $n(n-1)/2$ a v dispozícii na to máme len $n$ pozorovaní. Namiesto toho môžme uvažovať, že korelačná matica chýb má akúsi štruktúru.

- Napríklad v prípade seriálne korelovaných chýb to môže byť $\epsilon_{i+1} = \phi\epsilon_i + \delta_i$, kde $\delta_i \sim N(0, \tau^2)$.
- V prípade ak sú dáta zoskupené tak môžeme uvažovať, že korelácia v rámci skupinu je nejaké číslo $\phi$ inak 0.
- Chyby môžu byť priestorovo korelované.

## 7.2 Vážená metóda najmenších štvorcov

Teraz nech chyby sú nekorelované ale heteroskedastické, teda nech

$$var(\epsilon) = \begin{pmatrix} \sigma_1^2 & 0 & ... & 0 \\ 0 & \sigma_2^2 & ... & 0 \\ \vdots & & \ddots & \\ 0 & 0 & ... & \sigma_n^2 \end{pmatrix} = \begin{pmatrix} \frac{1}{w_1} & 0 & ... & 0 \\ 0 & \frac{1}{w_2} & ... & 0 \\ \vdots & & \ddots & \\ 0 & 0 & ... & \frac{1}{w_n} \end{pmatrix}.$$

V tomto prípade je váha pozorovania proporcionálna variancii, teda presnejšie meranie považujeme za dôveryhodnejšie a má väčšiu váhu. $var(\epsilon) = SS^T$, kde

$$S = \begin{pmatrix} \frac{1}{\sqrt{w_1}} & 0 & ... & 0 \\ 0 & \frac{1}{\sqrt{w_2}} & ... & 0 \\ \vdots & & \ddots & \\ 0 & 0 & ... & \frac{1}{\sqrt{w_n}} \end{pmatrix}$$

$$S^{-1} = \begin{pmatrix} \sqrt{w_1} & 0 & ... & 0 \\ 0 & \sqrt{w_2} & ... & 0 \\ \vdots & & \ddots & \\ 0 & 0 & ... & \sqrt{w_n} \end{pmatrix}.$$

Príklady:

- Ak $var(\epsilon_i) \propto x_i$, tak použijeme $w_i = x_i^{-1}$. Túto závislosť môžeme vyčítať z diagnostických grafov.
- Ak sú $y_i$ priemery z $n_i$ pozorovaní, potom podla ZVČ je variancia proporciálna $1/n_i$. Takže $var(y_i) = var(\epsilon_i) = \sigma^2/n_i$, teda $w_i = n_i$. Príklad: priemerná mzda v rôznych krajinách.
- Ak je presnosť pozorovaní rôznej kvality, tak váhy nastavíme ako $w_i = 1/var(y_i)$.

## 7.3 Testovanie vhodnosti fitu

Model dobre fituje dáta ak je náš odhad variancie šumu $\hat{\sigma}^2$ blizky skutočnej hodnote $\sigma^2$. Ak je odhad príliš veľký potom sú naše reziduá príliš veľký a náš model nefituje dáta dobre.

Môžeme náš odhad $\hat{\sigma}^2$ porovnať s nejakým $\sigma^2$ z nejakého iného modelu (pomocou F-testu). Toto by však značilo, že preferujeme určitý model ale či ten fituje dáta to nevieme.

Potrebovali by sme variáciu v $y$ pre rôzne kombinácie prediktorov. Môžeme napríklad merať tlak tomu istému pacientovi viacekrát. Toto je však nepostačujúce, potrebujeme variáciu $y$ pre **rôznych** ľudí. Nech $y_{ij}$ je $i-$te pozorvovanie v skupine ľudí $j$, ktorý majú rovnaké prediktory. Odhad $\sigma^2$, ktorý nezáleží od konkrétneho modeluj je

$$\hat{\sigma}^2_{\text{free}} = \sum_j \sum_i (y_{ij} - \bar{y}_j)^2/df,$$

kde $\bar{y}_j$ je priemer v $j$ tej skupine a počet stupňov voľnosti je $df = n - \#\text{groups}$, teda počet pozorovaní mínus počet skupín.

Takéto $\hat{\sigma}^2_{\text{free}}$ sa ľahko zráta. Treba len každej kombinácii prediktorov $j$ priradiť jeden parameter. Takýto model sa nazýva *saturovaný*. Porovnaním saturovaného modelu s naším modelom pomocou F-testu nám dáva spôsob ako otestovať vhodnosť fitu. Treba si dať pozor na interpretáciu $R^2$ pretože ak máme pre nejakú kombináciu prediktorov viacero rôznych odoziev, potom maximálne $R^2$ nie je 1 ale menej. Štatistické testovanie fitu je len pomôckou, nie cieľom pre výber modelu. Vieme, že model, ktorý výborne fituje dáta vie veľmi zle predikovať budúce hodnoty.

## 7.4 Robustná regresia

Ak sú chyby normálne rozdelené, MNŠ je odhad maximum likelihood, teda asymptoticky efektívny, teda v istom zmysle najlepší. Ak však nie sú, môžeme sa popozerať po alternatívach. Problémom je ak majú chyby ťažkochvosté rozdelenie a teda pravdepodobnosť veľkých reziduí je veľká a tieto vedia výrazne ovplyvniť fit, čo je nežiadúce.

Z minula vieme ako detekovať outlierov a vieme, že ich môžeme vyhodiť a odhadnúť parametre pomocou MNŠ na oklieštenej vzorke. Problémom však je ak je outlierov viacej a jeden maskuje druhého.

## 7.5 M - estimation

Ideme minimalizovať nejakú funkciu výchyliek, nie nutne štorce

$$\sum_{i=1}^{n} \rho(y_i - x_i^T \beta)$$

- $\rho(x) = x^2$ dostávame MNŠ
- $\rho(x) = |x|$ sa nazýva least absolute deviation - LAD alebo $L_1$ regresia
- $\rho(x) = \begin{cases} x^2/2 & \text{ak } |x| \leq c \\ c|x| - c^2/2 & \text{inak} \end{cases}$ sa nazýva Huberova metóda a je čosi medzi MNŠ a LAD. Parameter $c$ volíme ako nejaký robustný odhad $\sigma$. Hodnota propociálna mediánu $|\hat{\epsilon}|$ je vhodná.

Minimalizácia najmenších štvorcov vedie k riešeniu nasledovnej sústavy rovníc

$$X^T(y - X\hat{\beta}) = 0,$$

v prípade WLS máme

$$\sum_{i=1}^{n} w_i x_{ij}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j) = 0, \quad \forall j = 1, ..., p$$

Ak zdiferencujeme $\sum_{i=1}^{n} \rho(y_i - x_i^T\beta)$ podľa $\beta_j$ dostávame

$$\sum_{i=1}^{n} \rho'(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)x_{ij} = \sum_{i=1}^{n} \frac{\rho'(u_i)}{u_i} x_{ij}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j) = 0, \quad \forall j = 1, ..., p,$$

teda vidíme, že M-estimácia zodpovedá WLS s váhami $w(u) = \rho'(u)/u$.

- V prípade MNŠ je $w(u)$ konštantná
- Pri LAD: $w(u) = 1/|u|$. (Pri $u = 0$ máme problém, takže to musíme modifikovať)
- Huberova metóda: $w(u) = \begin{cases} 1 & \text{ak } |x| \leq c \\ c/|u| & \text{inak} \end{cases}$

Táto metóda potrebuje odhadnúť váhy na základe reziduí, preto postupujeme viackrokovo. V prvom kroku odhadneme reziduá a pomocou nich skonštruujeme váhy, ktoré použijeme v druhom kroku. Takto iterujeme až kým nedokovergujeme. Pozor $R^2$ v tomto prípade nedáva zmysel.

## 7.6 Least trimmed squares

Huberova aj $L_1$ metóda zlyhá ak máme zhluk outlierov. Least trimmed squares metóda sa pozerá len na určitý počet najmenších reziduálov, tie ostatné ignoruje.

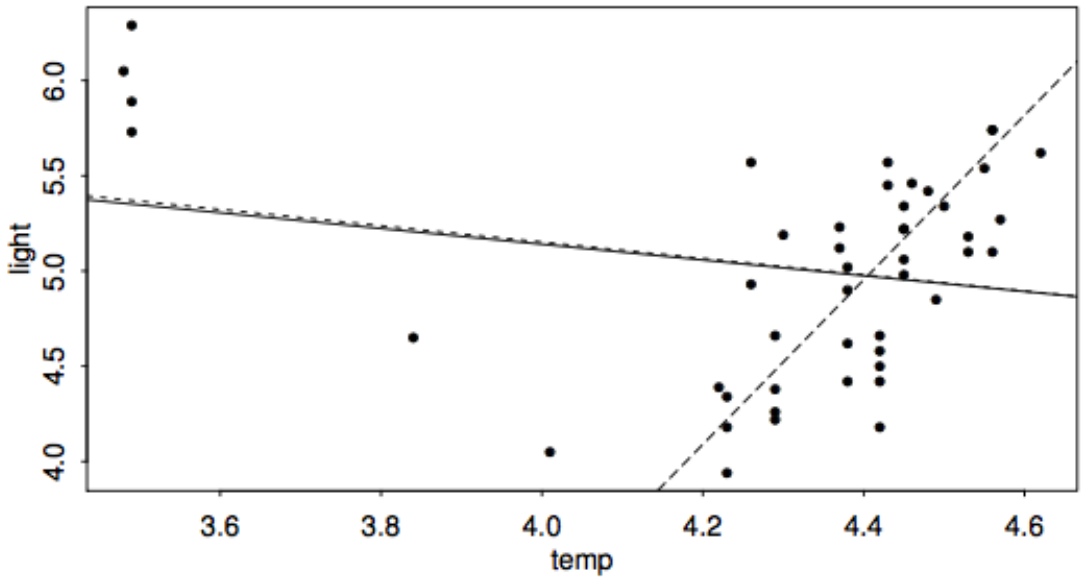
$$\arg\min \sum_{i=1}^{q} \hat{\epsilon}_{(i)}^2$$

Ako zvoliť $q$? Odporúča sa začať s $q = \lfloor n/2 \rfloor + \lfloor (p+1)/2 \rfloor$.

V tomto prípade nemáme asymptotické rozdelenie estimátora ako v prípade MNŠ a preto na konštrukciu konfidenčných intervalov použijeme bootstrap. Pripomíname reziduálny bootstrap

- Vygenerujeme $\epsilon^*$ - vektor dĺžky $n$ pomocou vyťahovania s opakovaním z $\hat{\epsilon}_1, ..., \hat{\epsilon}_n$
- Vytvoríme $y^* = X\hat{\beta} + \epsilon^*$
- Pomocou $(X, y^*)$ vypočítame $\hat{\beta}^*$

Toto zopakujeme mnohokrát a kvantily distribúcie vygenerovaných $\hat{\beta}^*$ použijeme ako konfidenčný interval.




Obr. 18: Porovnanie: MNŠ (plná čiara), Huberov estimátor (bodkovaná) a LTS (čiarkovaná). (Zdroj: [Far14]).

## 7.7 Zhrnutie

(1) Robustné estimátory nám pomôžu s dlhými chvostami distribúcie chýb avšak nie s korelovanosťou chýb.
(2) Na štatistickú inferenciu pri robustných estimátoroch používame bootstrap.
(3) Robustný estimátor môžeme použiť na detekciu problému s distribúciou chýb. Ak sa robustný estimátor a estimátor pomocou MNŠ príliš líšia, je to zdvihnutý varovný prst.
(4) Robustný estimátor je užitočný keď treba automaticky analyzovať dáta a outlieri nám môžu výrazne ublížiť.

# 8 Transformácie

V kapitole o diagnostike sme diskutovali aké máme v dispozícii riešenia v prípade ak niektoré predpoklady lineárneho Gaussovského modelu zlyhajú. V tejto kapitole sa podrobnejšie pozrieme na transformovanie premenných.

## 8.1 Transformácia odozvy

Ak je odozva transformovaná, tak model sa úplne zmenil. Parametre modelu majú úplne inú interpretáciu. To, ktorý model je vhodnejší závisí aj od toho, akým spôsobom v stupuje chyba do modelu a to čí

- aditívne - napr.: $y = \beta_0 + \beta_1 x + \epsilon$
- multiplikatívne - napr.: $\log(y) = \beta_0 + \beta_1 x + \epsilon \sim y = \exp(\beta_0 + \beta_1 x)\exp(\epsilon)$.

Toto v praxi samozrejme nevieme a preto je dobrou praxou vyskúšať viacej modelov a vybrať ten, ktorý má štrukturálnu formu správne. Distribúcia chýb je až druhoradá. Naučili sme sa spôsoby ako atakovať problémy nekonštatnej variancie alebo korelovanosti chýb.

### 8.1.1 Box-Coxova transformácia

Box-Coxova transformácia je spôsob ako pretransformovať kladnú odozvu ($y > 0$) tak, aby predpoklad normality v lineárneho modeli bol čo najvierohodnejší. Transformácia vyzerá nasledovne

$$y \to g_\alpha(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{ak } \lambda \neq 0 \\ \log(y) & \text{ak } \lambda = 0 \end{cases}.$$

teda uvažujeme akúsi triedu transformácií. Otázkou zostáva ako rozumne určiť parameter $\lambda$. Profilový log-likelihood v prípade lineárneho modelu s normálnymi chybami je

$$L(\lambda) = -\frac{n}{2}\log(RSS_\lambda/n) + (\lambda - 1)\sum \log(y_i).$$

Čo je to profilová log-vierohodnosť? Uvažujme log-likelihood $L(\beta, \sigma, \lambda)$ ktorý pre fixný parameter $\lambda$ je maximalizovaný v $(\beta^*(\lambda), \sigma^*(\lambda)) = \arg\max_{\beta,\sigma} L(\beta, \sigma, \lambda)$. Potom $L(\lambda) = L(\beta^*(\lambda), \sigma^*(\lambda), \lambda)$.

Maximalizáciu $L(\lambda)$ musíme robiť numericky. Dostaneme $\hat{\lambda}$ odhad ML, ktorý nám našepkáva ako môžeme pretransformovať odozvu. Nemusíme však robiť priamo $\frac{y^\lambda - 1}{\lambda}$ ale stačí $y^\lambda$, lebo násobok odozvy nám len ponásobí odhady parametrov a posun bude pohltený interceptom. A ak nám vyjde $\hat{\lambda} = 0.4893$, tak použije najbližšie rozumné číslo, teda v tomto prípade 0.5.

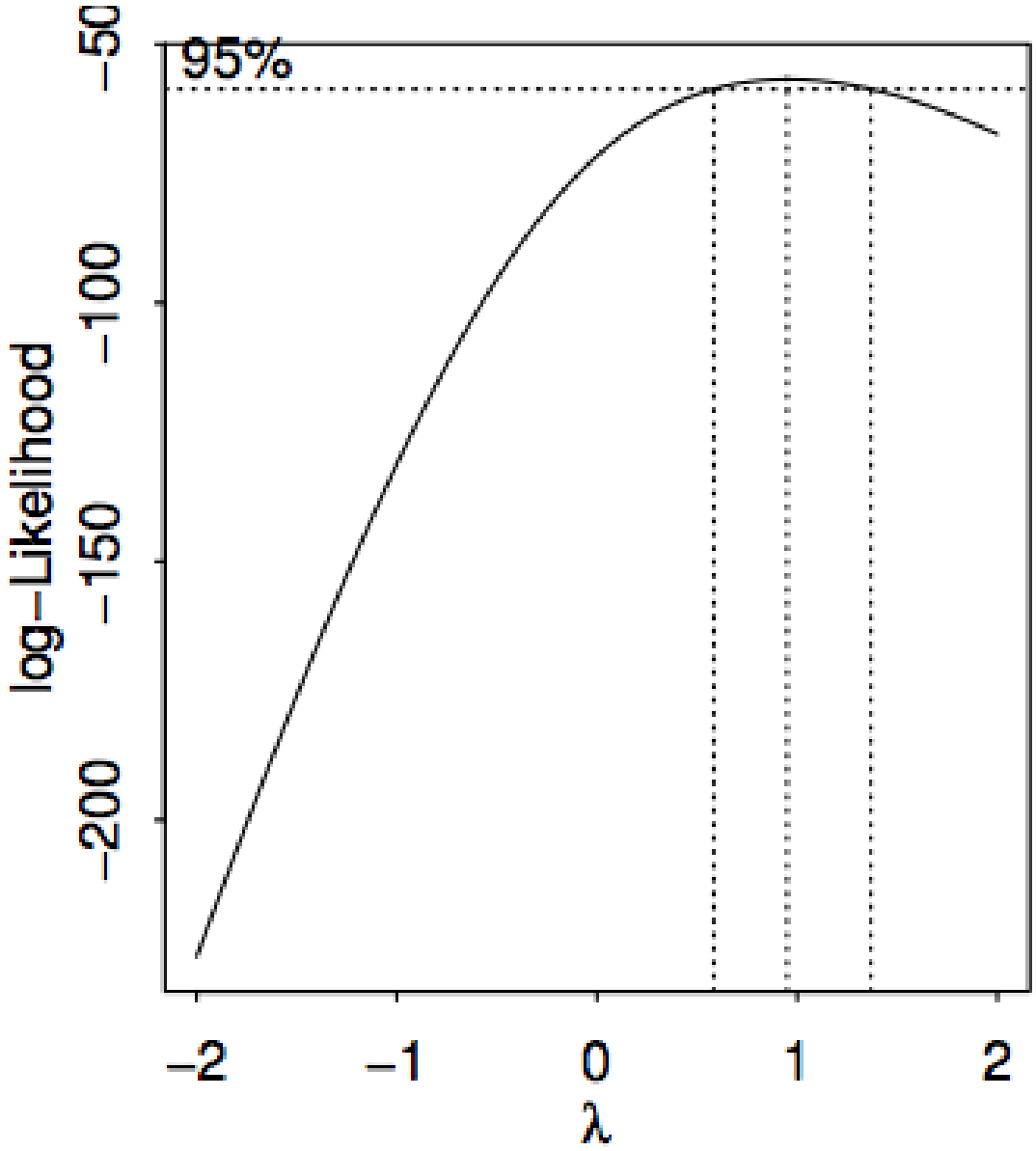
Testovať $H_0 : \lambda = \lambda_0$, môžeme pomocou pomeru vierohodností $2(L(\hat{\lambda}) - L(\lambda_0))$, ktorý má za platnosti $H_0$ asymptoticky $\chi^2_1$ rozdelenie. Zaujímavou je hodnota $\lambda_0 = 1$, teda test, či má vôbec zmysel transfromovať odozvu alebo nie.

Zopár poznámok ohľadom Box-Coxovej transformácie.

- Táto metóda je citlivá na outlierov, čo je indikované príliš vysokou hodnotou $\hat{\lambda}$.
- Ak sú nejaké $y_i < 0$, môžeme posunúť všetky $y_i$ o konštantu, ak nie je príliš veľká.
- Ak je $\max y / \min y$ malé, tak transformácia je skoro zbytočná lebo mocninová funckia je dobre aproximovaná lineárnou na krátkom intervale.

Sú aj iné metódy transformovanie $y$

- $\log(\alpha + y)$ kde $-\alpha$ je hodnota kedy začne $y$ prudšie rásť.
- $\log(y/(1-y))$ pre dáta typu proporcie
- $\frac{1}{2}\log((1+y)/(1-y))$ pre korelačné koeficienty




Obr. 19: Profile likelihood. V tomto prípade transformácia nie je užitočná, nulovú hypotézu $H_0 : \lambda = \lambda_0$ nevieme zamietnuť. (Zdroj: [Far14]).

## 8.2 "Hokejková"regresia

Predstavme si, že by sme pre rôzne časti dát chceli preložiť rôzne krivky. Napríklad krajiny s veľkým podielom populácie menej ako 15 ročných výrazne inak sporia ako krajiny z týmto nízkym podielom. Preloženie dát dvomi rôznymi krivkami vedia k nespojitosti, a to môže byť nežiadúce. Môžme chcieť preložit jednu krivku, ktorá však bude mať iné smernice pre rôzne časti dát.

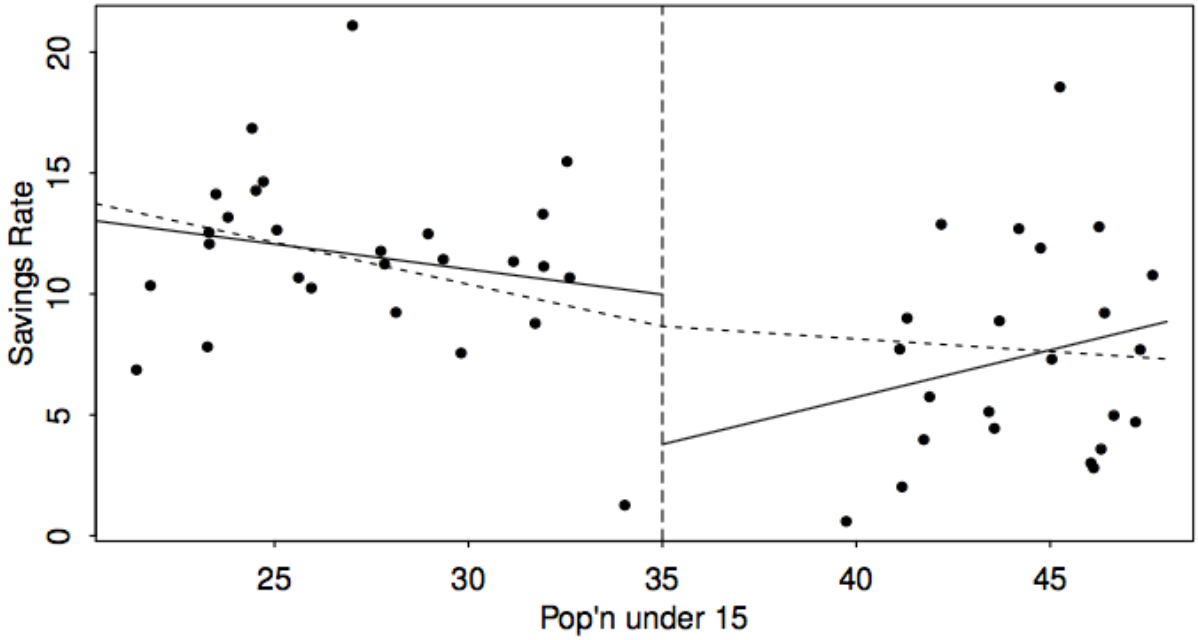
Ako to urobiť? Uvažujem dve hokejkové funkcie, kde konštanta $c$ je dopredu určená:

$$B_l(x) = \begin{cases} c - x & \text{ak } x < c \\ 0 & \text{inak} \end{cases} \qquad B_r(x) = \begin{cases} x - c & \text{ak } x > c \\ 0 & \text{inak} \end{cases}$$

Model, ktorý odhadneme bude nasledovný

$$y = \beta_0 + \beta_1 B_l(x) + \beta_2 B_r(x) + \epsilon.$$

ktorý má len 3 parametre, kdežto na preloženie dvoch priamok potrebujeme 4 parametre.




Obr. 20: Hokejková regresia. (Zdroj: [Far14]).

## 8.3 Polynómy

Lineárny model sa volá lineárny ale umožnuje nám odhadovať prudko nelineárne funkcie

$$y = \beta_0 + \beta_1 x + ... + \beta_d x^d + \epsilon.$$

Nejde ani o to, že by sme skutočne verili, že $y$ je naozaj akýmsi polynómom v závislosti od $x$ ale vie nám to pomôcť modelovať rôzne vlastnosti vzťahu medzi $x$ a $y$.

- kvadratickú funckiu používame ak sa nazdávame, že existuje nejaká optimálna hodnota prediktora pre odozvu. Napríklad ak je odozvou chuť chleba a prediktorom teplota. Nechceme totiž chlieb ani nedopečený ani spálený.
- Môžeme postupne pridávať členy vyššieho rádu kým sú signifikantné.
- Alebo postupne odoberať nesignifikantné členy.
- Odoberať členy nižšieho rádu vo všeobecnosti je nie dobrý nápad alebo ak to robíme, musíme mať na to dobrý dôvod.
- Ak napríklad odobereme lineárny člen ale kvadratický tam necháme, implikuje to, že optimum je v nule.
- Ak odoberieme intercept, tak regresná krivka musí ísť cez nulu.

Aby sme zabránili takýmto problémom, môžeme skonštruovať ortogonálne polynómy. Takže ak pridáme ďalší regresor (polynóm vyššieho rádu) parametre sa nám nezmenia. Ako bonus sú aj numericky stabilnejšie. Pre polynómy $z_i$

$$z_1 = a_1 + b_1 x$$
$$z_2 = a_2 + b_2 x + c_2 x^2$$
$$z_3 = a_3 + b_3 x + c_3 x^2 + c_4 x^3$$

nastavíme koeficienty $a, b, c...$ tak aby platilo $z_i^T z_j = 0$ pre $i \neq j$.

## 8.4 Splajny

Predošlé metódy mali nejaké žiadúce aj nežiadúce vlastnosti.

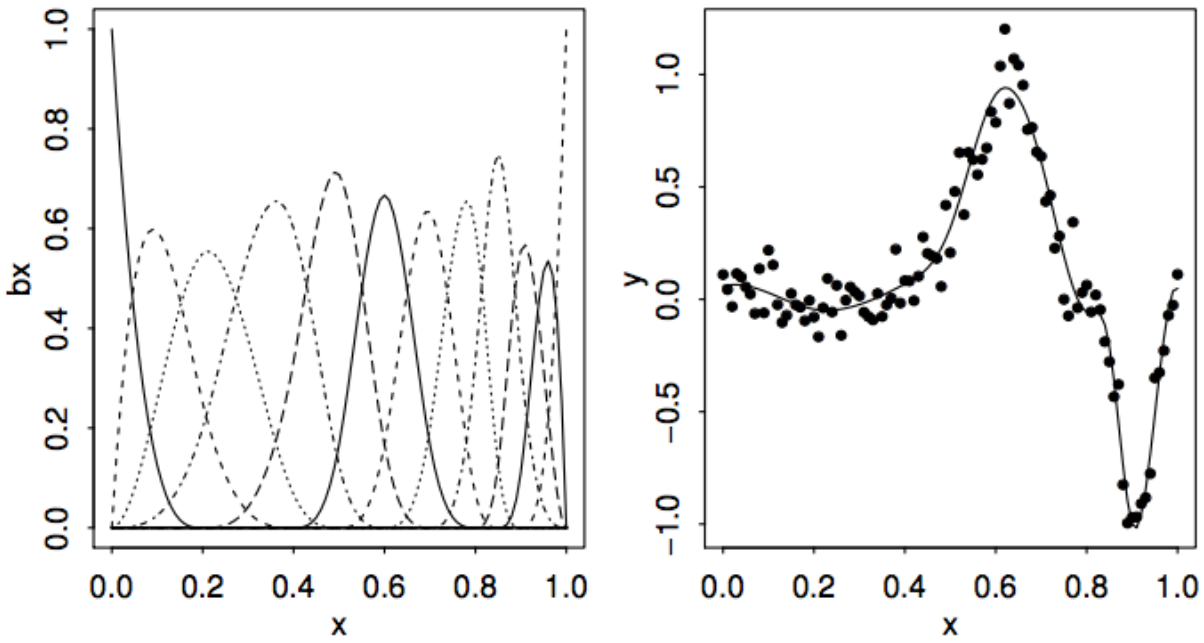
- polynómy boli hladké funkcie ale každý jeden bod ovplyvňuje celkový fit.

- hokejky mali vlastnosť "lokálnosti", body na jednej strane neovplyvňovali fit na druhej strane. Neboli však dosť hladké.

Splajny kombinujú výhody oboch týchto metód. Urobíme obyčajnú lineárnu regresiu, kde odozvu vysvetľujeme pomocou bázických funkcií.

Čo sú bázické funkcie kubického B-splajnu? Sú to funkcie na danom intervale $[a, b]$ a pri danej množine uzlov $t_1, t_2, ..., t_k$ ktoré spĺňajú nasledovné vlastnosti

- Každá bázická funkcia je nenulová na štyroch po sebe idúcich intervaloch zadefinovaných uzlami a všade inde je nulová. (Toto zabezpečí lokálnosť)
- Na každom intervale medzi dvoma po sebe idúcimi uzlami je bázická funkcia kubickým polynómom
- Každá bázická funkcia je spojitá a má spojité prvé a druhé derivácie v uzloch. (Toto zabezpečí hladkosť.)
- Integrál bázických funkcií je 1.




Obr. 21: Kubický B-splajn, vľavo bázické funkcie, vpravo dáta a fit. (Zdroj: [Far14]).

Alternatívou ku *regresných splajnov* spomenutým vyššie sú *vyhladzovacie splajny*. Namiesto minimalizácie štvorcov reziduí minimalizujú

$$\frac{1}{n}\sum(y_i - f(x_i))^2 + \lambda \int [f''(x)]^2 dx,$$

kde $\lambda$ je "cena" za nehladkosť. Parameter $\lambda$ môžeme zvoliť pomocou krížovej validácie.

## 8.5 Aditívne modely

Aditívne modely sú veľmi flexibilné. Naraz nám určia transformácie všetkých prediktorov.

$$y = \alpha + f_1(x_1) + ...f_p(x_p) + \epsilon$$

Funkcie $f_i$ môžu byť napríklad vyhladzovacie splajny.

# 9 Výber modelu

V tejto kapitole sa budeme baviť o výbere modelu. V dispozícii máme istú množinu modelov, napríklad v dispozícii nejaké regresory, ktoré môžu alebo nemusia vstupovať do regresie. Alebo môžeme mať záujem určovať stupeň polynómu.

Väčší model bude lepšie fitovať dáta, avšak pridaním nepotrebných premenných vnášame do modelu šum a ten zhorší predikciu. Nejde len o predikciu, princíp Occamovej britvy hovorí, že z rôznych vysvetlení nejakého fenoménu si vyberáme to jednoduchšie vysvetlenie.

- *'Among competing hypotheses, the one with the fewest assumptions should be selected.'*
- John Punch 1639: *'Entities must not be multiplied beyond necessity'*
- Aristoteles: *'We may assume the superiority ceteris paribus [other things being equal] of the demonstration which derives from fewer postulates or hypotheses.'*
- Ptolemaus: *'We consider it a good principle to explain the phenomena by the simplest hypothesis possible.'*
- Madhva: *'To make two suppositions when one is enough is to err by way of excessive supposition'*
- Isaac Newton: *'We are to admit no more causes of natural things than such as are both true and sufficient to explain their appearances. Therefore, to the same natural effects we must, as far as possible, assign the same causes.'*

Okrem toho, zbierať menej typov dát je aj lacnejšie. Výber modelu závisí aj od toho, čo s ním chceme robiť.

- Ak chceme predikovať, môžeme uprednostniť väčší model aj keď menší model môže byť estetickejší. Pripúšťame, že nejaký prediktor má vplyv na odozvu aj keď to v našej konkrétnej dátovej vzorke tak nevyzerá.
- Ak chceme vysvetľovať, preferujeme menšie modely, ale nedávame do regresie len tie prediktory, ktorých efekt nás zaujíma ale chceme kontrolovať aj efekt ostatných premenných.

V tejto kapitole sa budeme venovať dvom typom výberu modelu: jeden na základe testovania hypotéz a druhý na základe tzv. informačných kritérií.

## 9.1 Hierarchické modely

Niekedy môžeme v rámci priestoru modelov určiť akúsi prirodzenú hierarchiu. Napríklad keď modelujeme odozvu ako polynóm prediktora.

Predstavme si situáciu, že by sme mali model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

a vyšiel by nám, že prediktor $x$ nie je štatisticky významný. Ak by sme však vybrali miesto neho model

$$y = \beta_0 + \beta_2 x^2 + \epsilon,$$

tento model by mal takú vlastnosť, že ak posunieme $x$ o konštantu, teda ak $x$ nahradíme $x + a$, model sa zmení na

$$y = \beta_0 + \beta_2 a^2 + 2\beta_2 a x + \beta_2 x^2 + \epsilon.$$

Teda náš model neumožnuje posunutie $x$ o konštantu, a to je nežiadúca vlastnosť.

Iný príklad: ak modeluje odozvu ako polynóm dvoch prediktorov

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon,$$

a odoberieme interakčný člen, znamená to, že dostaneme model ktorý neumožňuje rotáciu priestoru prediktorov, táto by totiž znovu zaviedla interakčný člen.

## 9.2 Procedúry založené na testovaní hypotéz

- **Spätná eliminácia** - Začneme s modelom so všetkými prediktormi a postupne vynechávame prediktory, ktorých príslušná p-hodnota je väčšia ako nejaká hodnota $\alpha_{crit}$. Toto opakujeme až dokým nebudú mať všetky prediktory p-hodnotu menšiu ako $\alpha_{crit}$. Ak nás zaujíma predikcia, hodnota $\alpha_{crit}$ okolo 15-20% sa javí ako vhodná.
- **Postupný výber regresorov** - Začíname s malým modelom a pridávame do modelu prediktor, ktorý má najmenšiu p-hodnotu, ktorá je zároveň menšia ako $\alpha_{crit}$. Zastavíme, keď žiaden nový prediktor nemá p-hodnotu menšiu ako $\alpha_{crit}$.
- **Kroková regresia** - Je kombináciou predošlých dvoch. Sadou pravidiel môžeme v jednotlivých krokoch pridať alebo odoberať regresory.

Zopár poznámok:

- Tieto metódy nemusia vybrať optimálny model.
- p-hodnoty finálneho modelu nie sú platné, pretože sú výsledkom viackrokového testovania. Výsledky sú optimistickejšie ako skutočnosť.
- to, že sú premenné odstránené z modelu neznamená, že nesúvisia s odozvou, ale skôr, že ich predikčná sila je popri ostatných prediktoroch len malá.
- tieto procedúry majú tendenciu vyberať modely, ktoré sú príliš malé na predikciu. Predstavte si príklad, kedy uvažujeme či necháme sklon priamky v regresii alebo nie. Aj keď napríklad nie je štatisticky významný, nevieme, či v budúcnosti nebude vedieť dobre predikovať, nesignifikantnosť môže byť čisto len dôsledkom malej dátovej vzorky.
- keď máme príliš málo pozorovaní v porovnaní s počtom prediktorov, nájdeme aj šume nejaké závislosti.
- niekedy máme premenné, ktoré spoločne fungujú ako skupina. Napríklad dummy (0-1) premenné pre roky, ak modelujeme závislosť v čase. Tieto procedúry na to neprihliadajú a môžu napríklad vyhodiť dummy premenné pre niektoré roky a iné zasa nie. V tom prípade musíme tie ostatné do regresie manuálne pridať
- tieto metódy v žiadnom prípade **nenahrádzajú** expertnú skúsenosť alebo zdravý rozum a musia byť používané veľmi opatrne. Sú akýmsi veľkým mechanickým kladivom.
- tieto metódy vám neodhalia zázračne čo sa v dátach nenachádza. Ak máte málo alebo nekvalitných dát, tak vám nepomôže žiadna akokoľvek sofistikovaná metóda.

## 9.3 Procedúry založené na informačných kritériách

Informačné kritériá hodnotia fit modelu ale berú do úvahy aj jeho komplexitu. Ak majú dva modely rovnaký fit vyberáme si ten, ktorý je jednoduchší.

$$IC = -\left([\text{FIT MODELU}] - [\text{KOMPLEXITA MODELU}]\right)$$

Máme parametrický model $f(y|\theta)$, zatiaľčo dáta pochádzajú zo skutočnej, nám neznámej distribúcie $g(y)$. Ak je model korektne špecifikovaný, potom existuje hodnota $\theta_0$ taká, že $f(y|\theta_0) = f(y)$. Ak model nie je korektne špecifikovaný, potom takáto hodnota neexistuje. Keď sa už funkcia $g$ nemôže rovnať $f(\cdot|\theta)$ pre žiadnu hodnotu parametra $\theta$, chceli by sme vybrať aspoň takú hodnotu, aby boli $g$ a $f(\cdot|\theta)$ čo najbližšie pri sebe. Uvažujme nasledovnú vzdialenosť dvoch distribúcií, ktorá sa nazýva *Kuhlback-Leiblerova divergencia.*

$$KL\left(g, f(\cdot|\theta)\right) = \int g(y) \log\left(\frac{g(y)}{f(y|\theta)}\right) dy$$

Uvažujme maximum-likelihood odhad $\hat{\theta}$ parametra $\theta$. KL divergencia v tomto bode je

$$KL\left(g, f(\cdot|\hat{\theta})\right) = \int g(y) \log\left(\frac{g(y)}{f(y|\hat{\theta})}\right) dy.$$

Táto kvantita je náhodnou premennou a náhodnosť prichádza z $\hat{\theta}$. Nevychýlený odhad strednej hodnoty tejto náhodnej premennej je

$$E_g\left(KL\left(g, f(\cdot|\hat{\theta})\right)\right) = -L(\hat{\theta}) + p + \texttt{constant},$$

kde $p$ je počet parametrov. Všimnime si, že ak uvažujeme sadu modelov s rovnakým počtom parametrov, tak minimalizovanie strednej hodnoty KL-divergencie je to isté ako maximalizovanie log-likelihoodu. Teda minimalizovanie hore uvedeného kritéria je akýsi maximum likelihood, ktorý berie do úvahy veľkosť modelu.

**Akkaikeho informačné kritérium (AIC)** je

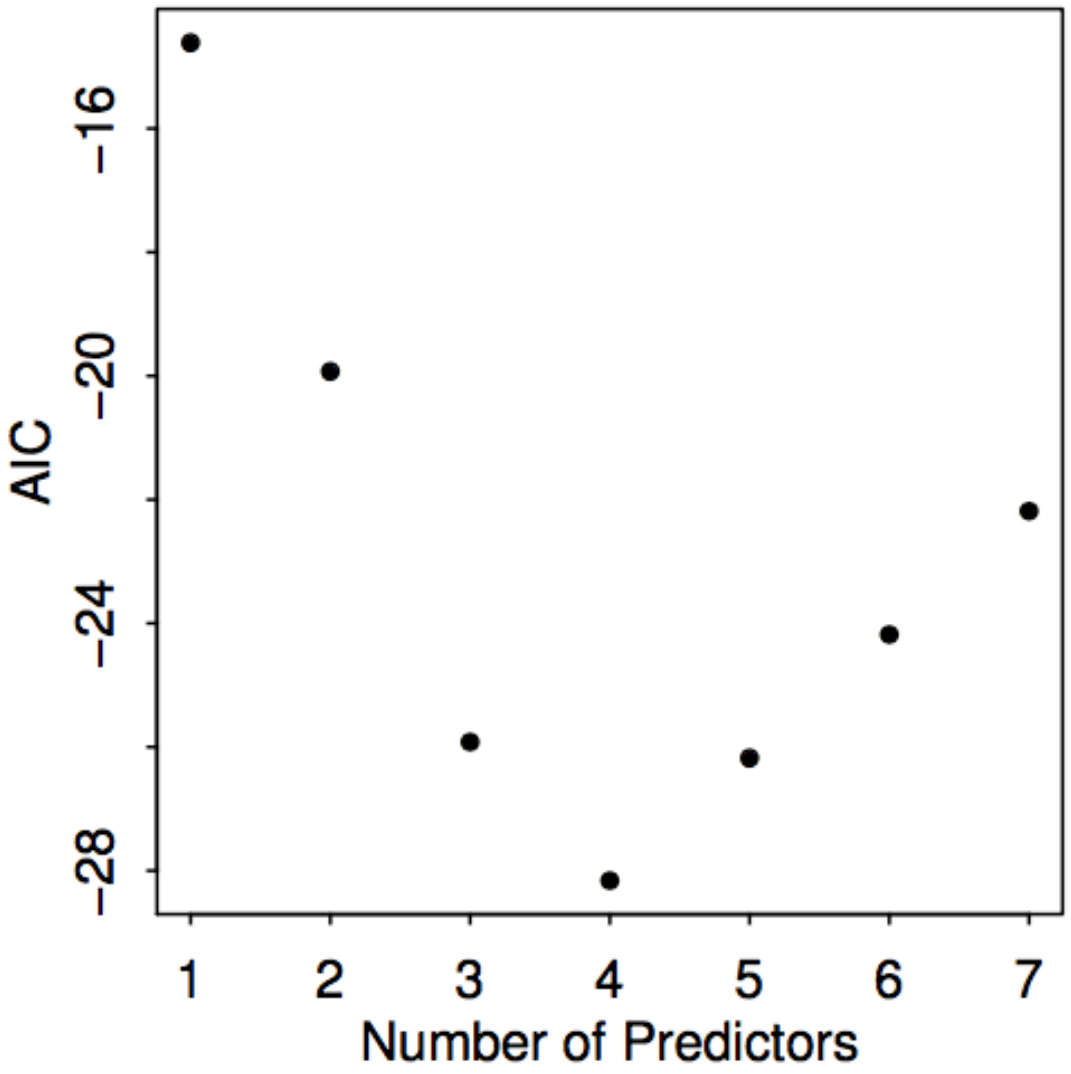
$$AIC = -2L(\hat{\theta}) + 2p,$$

a preferujeme modely s čo najnižšou hodnotou AIC. Prvý člen preferuje modely, ktoré fitujú dáta lepšie, druhý zas menšie modely.

Alternatívou k AIC je **Bayesovské informačné kritérium (BIC)**

$$BIC = -2L(\hat{\theta}) + 2p\log(n)$$

ktoré prísnejšie penalizuje veľké modely. Ktoré kritérium vybrať? Ak nás zaujíma prediktívna výkonnosť tak AIC, ak chceme malý parsimónny model, tak BIC.




Obr. 22: Najmenšie AIC pre rôzny počet prediktorov v lineárnej regresii. (Zdroj: [Far14]).

Ďalšie kritérium pre posudzovanie modelu je **upravené** $R^2$ (adjusted $R^2$ alebo $R_a^2$). Obyčajné $R^2 = 1 - \frac{RSS}{TSS}$ neberie do úvahy veľkosť modelu.

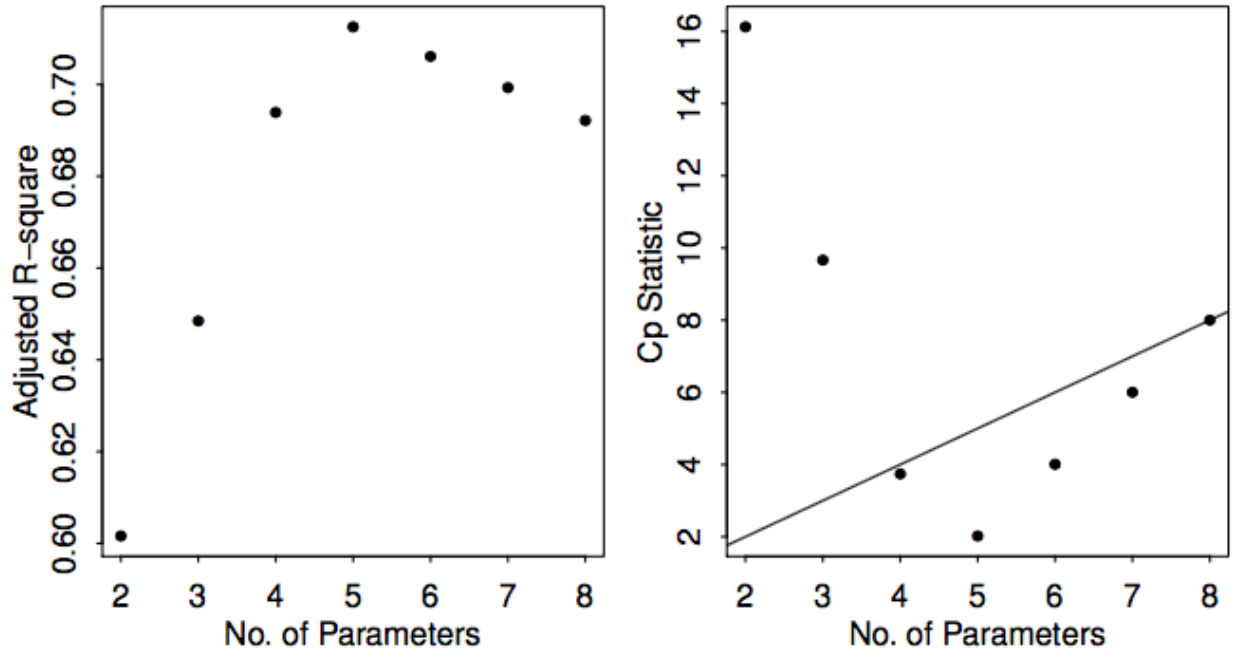
$$R_a^2 = 1 - \frac{RSS/(n-p)}{TSS/(n-1)} = 1 - \frac{\hat{\sigma}^2_{model}}{\hat{\sigma}^2_{null}}.$$

Kde $\hat{\sigma}^2_{model}$ je nevychýlený odhad variancie chýb v lineárnom regresnom modeli a $\hat{\sigma}^2_{null}$ je odhadom v modeli len s interceptom.

Posledným kritériom, ktoré spomenieme je **Mallow-ova $C_p$ štatistika.** Dobrý model by mal mať malú priemernú sumu štvorcov predikčných chýb $\frac{1}{\sigma^2}\sum_i E(\hat{y}_i - E(y_i))^2$ a táto kvantita sa dá odhadnúť pomocou

$$C_p = \frac{RSS_p}{\hat{\sigma}^2} + 2p - n,$$

kde $\hat{\sigma}^2$ je odhad variancie šumu z modelu so všetkými prediktormi, $RSS_p$ je reziduálna suma štvorcov z modelu s $p$ parametrami. Pre model so všetkými prediktormi sa $C_p = p$. Pre model s $p$ prediktormi $E(RSS_p) = (n-p)\sigma^2$, a preto $E(C_p) \approx p$. Chceli by sme model, kde $C_p$ bude menšie ako $p$.




Obr. 23: Najmenšie $R_a^2$ a $C_p$ pre rôzny počet prediktorov. (Zdroj: [Far14]).

## 9.4 Záver

Automatické metódy na výber modelu sú nástrojom, ktorý musí byť používaný opatrne a nenahradí expertnú skúsenosť. Metódy založené na testovaní hypotéz si opakovane vyberajú stále väčší a väčší model alebo menší a menší model. Kriteriálne metódy fungujú systematickejšie a sú lepšie teoreticky podkuté, preto sú vo všeobecnosti preferovanejšie.

Ak máme skupinu modelov, ktoré fitujú dáta podobne, tak na pomoc použijeme tieto otázky.

- Dávajú modely kvalitatívne podobné závery?
- Predikujú podobne?
- Ako náročné je meranie prediktorov?
- Ktorý model má najlepšie diagnostické grafy?

Ak sme v situácii, že modely sú podobné ale vedú k výrazne rozdielnym výsledkom, potom skrátka dáta nemusia vedieť jednoznačne odpovedať na otázku, ktorá nás zaujíma. Je intelektuálne poctivé oceniť túto neistotu, akokoľvek to môže byť nepraktické alebo nevýhodné.

# 10 Scvrkávacie metódy

Častokrát sme v situácii, že prediktorov je príliš veľa v porovnaní s veľkosťou dátovej vzorky. Predošlá kapitola hovorila o výbere modelu. Mohli sme z našich kandidátov na prediktorov vybrať nejakých napríklad pomocou informačného kritéria.

Táto kapitola bude o tom, ako rozumne zmenšiť zmenšiť dimenziu priestoru prediktorov.

## 10.1 Metóda hlavných komponentov (Principal Components Analysis)

Predstavme si situáciu, že máme mnoho prediktorov, veľmi korelovaných. Môžeme sa snažiť ortogonalizovať tento priestor. Ak máme 40 premenných a 100 porozovaní, tak máme 100 bodiek v 40 rozmernom priestore. Našou ambíciou bude teraz zmenšiť tento priestor. Naša matica prediktorov má tentkrát rozmer 100 krát 40 (pozor na preklep v knihe).

Začneme tým, že nájdeme 40 takých čísel $u_1$ (vektor rozmeru 40), že variancia $Xu_1$ je maximálna možná. Okrem toho budeme požadovať, aby $u_1^T u_1 = 1$, a to preto, aby bol vektor $u_1$ jednoznačne definovaný. Vektor $Xu_1$ budeme nazývať prvý hlavný komponent. Spomedzi všetkých možných lineárnych kombinácií prediktorov (teda kombinácií stĺpcov matice $X$) sme vybrali tú najvariabilnejšiu kombináciu, teda tú v ktorej je najviac informácie.

Druhý hlavný komponent dostaneme tak, že nájdeme vektor $u_2$ tak, aby $var(Xu_2)$ bola maximálna a zároveň aby $u_2$ bol kolmý na $u_1$, teda aby $u_2^T u_1 = 0$ a zároveň $u_2^T u_2 = 1$.

Týmto spôsobom pokračujeme a vieme reprezentovať každý bol v oblaku bodov matice $X$ ako jednoznačnú lineárnu kombináciu hlavných komponentov. Pointou je, že napríklad 5 hlavných komponentov môže vysvetliť až 99% variácie v $X$. To znamená, že nepotrebujeme 40 čisel na reprezentovanie jedného porozorania ale stačí nám napr. len 5.

- Hlavné komponenty sú ortogonálne čo je výhodné, ak ich používame ako prediktory v regresii, lebo pridaním ďalšieho hlavného komponentu sa nám nezmenia odhady parametrov. Okrem toho odhady sú numericky stabilnejšie.
- Hlavné komponenty môžu šetriť pamäť alebo miesto na disku.
- Hlavné komponenty môžu ale nemusia byť interpretovateľné. Niekedy tým, že vidíme akej lineárnej kombinácii $u_1$ zodpovedá prvý komponent, tak môžme pochopiť o čo ide, a to nám vie pridať vhľad do problematiky.
- Hlavné komponenty nám môžu pomôcť odhaliť zhluky podobných bodov. V 40 rozmeroch nás častokrát naša intuícia opúšťa.

Pre použítím PCA je častokrát vhodné normalizovať prediktory, to je kvôli tomu, aby analýza nezávisela na posune o konštantu alebo na zmene jednotiek.

PCA je založená na variancii a tá je citlivá na outlierov. Otázkou je ako nájsť outlierov v 40 rozmernom oblaku dát? Pomocou tzv. Mahalanobisovej vzdialenosti.

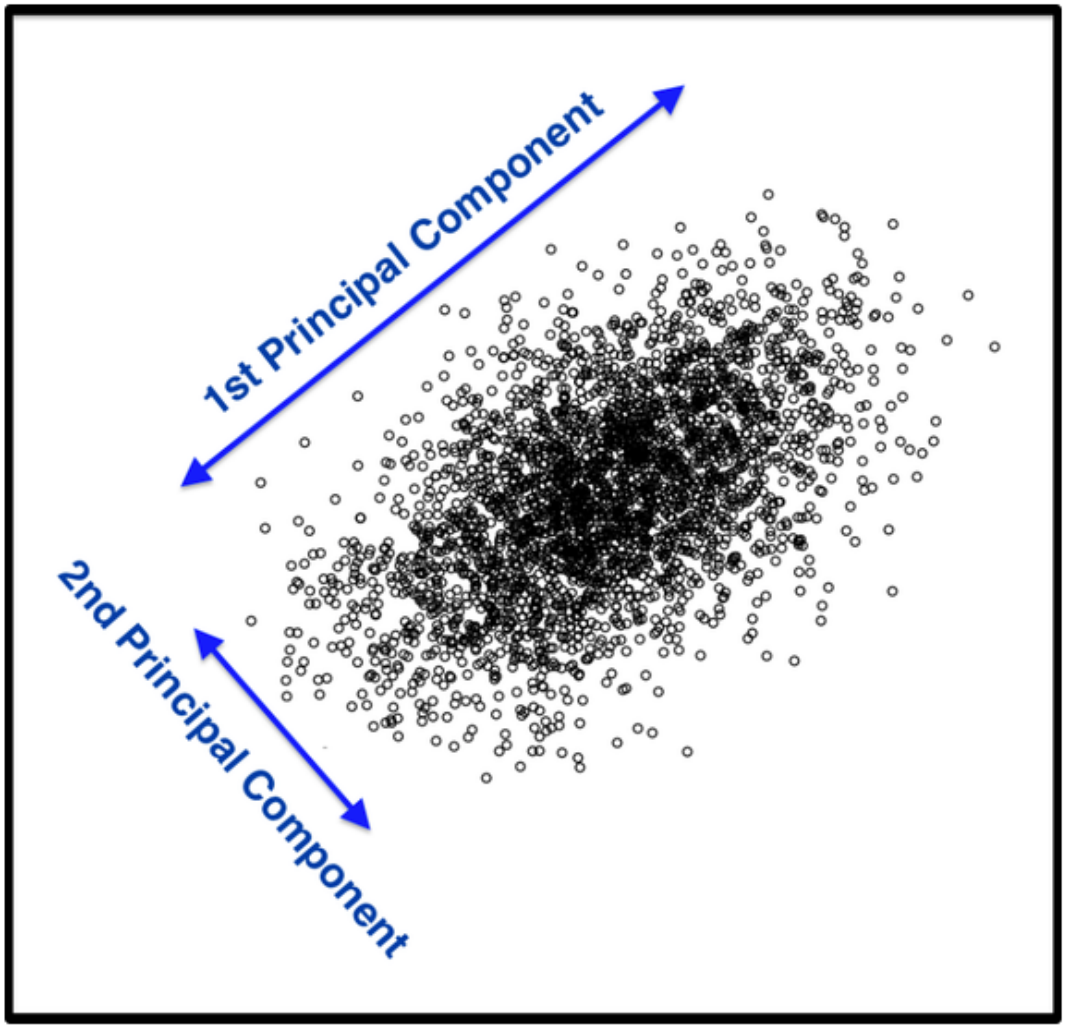
$$d_i = \sqrt{(x_i - \mu)^T \Sigma^{-1} (x_i - \mu)},$$

kde $\mu$ je miera centrality a matica $\Sigma$ je mierou kovariancie. Na odhad $\mu$ a $\Sigma$ je vhodné použiť robustné odhadovacie metódy.

Hlavné komponenty sa dajú použiť v regresii, akým spôsobom vyberieme počet hlavných komponentov vstupujúcich do regresie? Pomocou krížovej validácie. To ako dobre model predikuje budeme merať pomocou RMSE= root mean squared error.

$$RMSE = \sqrt{\left( \sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n} \right)}.$$

Keď už chceme vedieť ako dobre naša metóda predikuje, najlešie je to na nových dátach. Alebo môžeme rozdeliť naše pôvodné dáta na dve vzorky - trénovaciu a testovaciu. Na trénovacej odhadneme model a na testovacej otestujeme ako dobre sa modelu darí predikovať.




Obr. 24: (Zdroj: `https://www.quora.com/What-is-an-intuitive-explanation-for-PCA`

## 10.2 Odvodenie PCA

Majme $p$-rozmerný náhodný vektor $x = (x_1, x_2, ..., x_p)^T$, ktorého stredná hodnota je $(0, ..., 0)^T$ a kovariančná matica $(p \times p)$ je $\Sigma$, ktorej $(i, j)$-ty člen je $\Sigma_{i,j} = cov(x_i, x_j)$. Variancia $u^T x$ je $var(x^T u) = u^T \Sigma u$. K nasledujúcemu minimalizačnému problému

$$\max u^T \Sigma u \quad \text{subject to} \quad u^T u = 1$$

je príslušná Lagrangeova funkcia $u^T \Sigma u - \lambda(u^T u - 1)$. Ak je nejaký vektor $u$ riešením tohoto maximalizačného problému, potom musí spĺňať podmienky prvého rádu

$$\frac{\partial}{\partial u}\left(u^T \Sigma u - \lambda(u^T u - 1)\right) = 2\Sigma u - 2\lambda u = 0,$$

a preto musí byť $u$ vlastné číslo kovariančnej matice $\Sigma$ a Lagrangeov mulitplikátor $\lambda$ je príslušné vlastné číslo.

$$var(u^T x) = u^T \Sigma u = \lambda u^T u = \lambda,$$

preto prvý hlavný komponent je $x^T u_1$, kde $u_1$ je vlastný vektor, ktorému zodpovedá najväčšie vlastné číslo matice $\Sigma$.

Druhý hlavný komponent dostaneme podobne, riešime problém

$$\max u^T \Sigma u \quad \text{subject to} \quad u^T u_1 = 0, u^T u = 1,$$

kde podmienka $u^T u_1 = 0$ zabezpečí, že $cov(u^T x, u_1^T x) = u^T \Sigma u_1 = \lambda_1 u^T u_1 = 0$. Príslušná Lagrangeova funkcia je $u^T \Sigma u - \lambda(u^T u - 1) - \phi u^T u_1$. Pomocou podmienok prvého rádu

$$\frac{\partial}{\partial u}\left(u^T \Sigma u - \lambda(u^T u - 1) - \phi u^T u_1\right) = 2\Sigma u - 2\lambda u - \phi u_1 = 0$$

Ponásobením poslednej rovnice zľava $u_1^T$ dostávame $2u_1^T \Sigma u - 2\lambda u_1^T u - \phi u_1^T u_1 = 0 - 0 - \phi 1 = 0$, preto $\phi = 0$ a podmienka prvého rádu sa redukuje na

$$\Sigma u - \lambda u = 0.$$

Preto znova aj pre druhý hlavný komponent musí platiť, že je to lineárna kombinácia $x^T u_2$ taká, že $u_2$ je vlastný vektor matice $\Sigma$. Vyberieme vlastný vektor, ktorému zodpovedá druhé najväčšie vlastné číslo.

Tento postup opakujeme a $k$-ty hlavný komponent je je $x^T u_k$, kde $u_k$ je vlastný vektor matice $\Sigma$ zodpovedajúci $k$-temu najväčšiemu vlastnému číslu-

Kovariančná matica $\Sigma$ je neznáma ale vieme ju konzistentne odhadnúť ako (poznamenávame, že $x$ má nulovú strednú hodnotu)[4]

$$S = \hat{\Sigma} = \frac{1}{n-1} X^T X,$$

kde $X$ je $(n \times p)$ matica, ktorej riadky sú jednotlivé pozorovania.

[4]Ak $x$ nemá nulovú strednú hodnotu, tak stačí od každého prvku matice $X$ odpočítať jeho priemer cez stĺpce.

## 10.3 PCA príklady


Obr. 25: (Vľavo: rôzne kombinácie prediktorov dosiahnu rôzne veľkú variáciu v dátach. Tretia lineárna kombinácia obsahu vit.C vlákniny a tuku nám umožňuje rozlišovať medzi potravinami. Vpravo: výsledok PCA, zobrazenie jedál pomocou dvoch hlavných komponentov. Zdroj: `https://www.quora.com/What-is-an-intuitive-explanation-for-PCA`

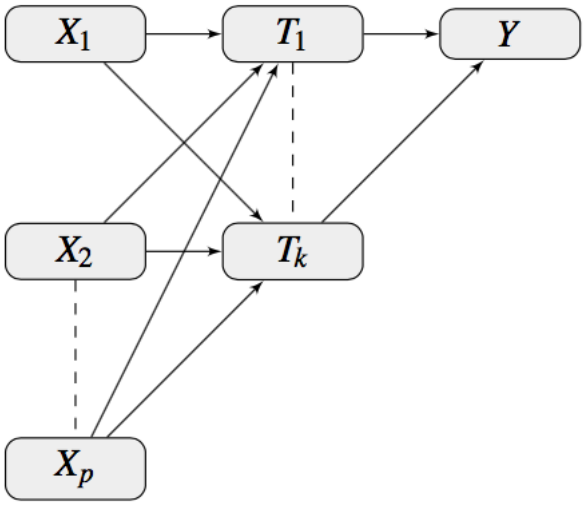

Obr. 26: Každých obrázok je dlhý vektor čísel, kde každé číslo zodpovedá sivosti daného pixelu. Vľavo originálne tváre. V strede: odhad tvárí pomocou 36 hlavných komponentov. Vpravo: 36 hlavných komponentov. Zdroj: `https://github.com/gbuesing/ pca/tree/master/examples`

## 10.4 Partial Least Squares

Našou úlohou je nájsť také ortogonálne kombinácie $T_1, ..., T_k$ prediktorov $X_1, ..., X_p$, že predikcia

$$\hat{y} = \beta_1 T_1 + ... + \beta_k T_k,$$

je čo najlešia možná. Na odhadnute $T_i$ sú rôzne algoritmy a ich počet sa vyberá pomocou krížovej validácie.




Obr. 27: Partial Least Squares - schematické znázornenie. Zdroj: Faraway(2014)

## 10.5 Ridge regression

Ako zostabilniť odhady parametrov? Tak, že im "zakážeme"príliš veľké hodnoty. Toto môžeme urobiť viacerými spôsobmi. Jedným z nich je hrebeňová regresia. Namiesto minimalizácie štvorcov, minimalizujeme

$$(y - X\beta)^T(u - X\beta) + \lambda \sum_j \beta_j^2$$

čo je ekvivalentné s

$$(y - X\beta)^T(u - X\beta) \quad \text{subject to} \quad \sum_j \beta_j^2 \leq t^2.$$

Odhad je $\hat{\beta} = (X^T X + \lambda I)^{-1} X^T y$. Toto je príklad penalizovanej regresie. Pre $\lambda \to 0$ dostávame $\hat{\beta} \to \hat{\beta}_{LS}$ a pre $\lambda \to \infty$ dostávame $\hat{\beta} \to 0$. Parameter $\lambda$ nastavíme pomocou krížovej validácie.

Odhad parametrov je vychýlený ale to je cena, ktorú platíme za stabilnejší, teda menej variabilný odhad.

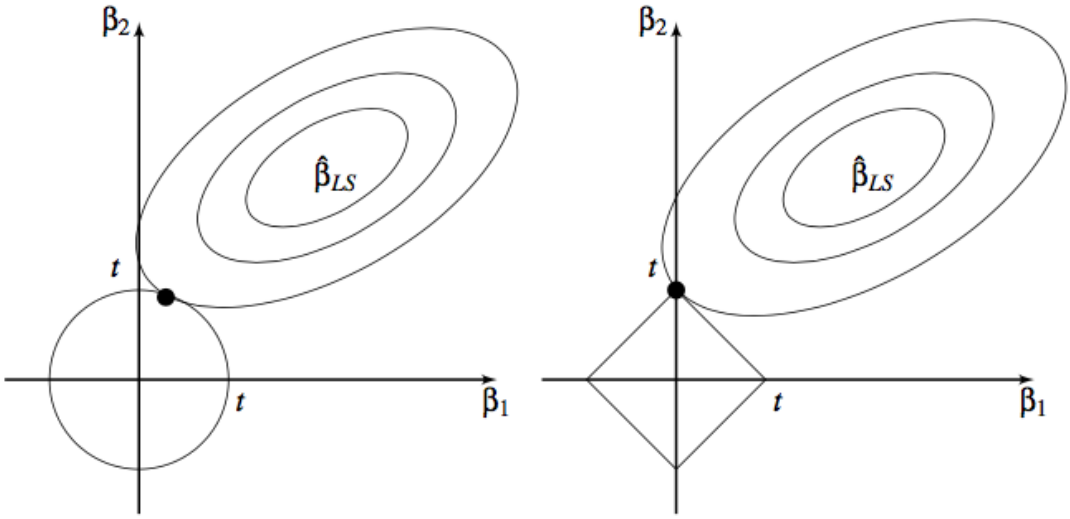
## 10.6 LASSO

$$(y - X\beta)^T(u - X\beta) + \lambda \sum_j |\beta_j|$$

čo je ekvivalentné s

$$(y - X\beta)^T(u - X\beta) \quad \text{subject to} \quad \sum_j |\beta_j| \leq t$$

Výhodou Lassa je, že vďaka tvaru penalty niektoré parametre priamo vynuluje. Preto akoby robil výber modelu aj odhadovanie modelu naraz! Lasso je rozumné používať ak sa domnievame, že existuje niekoľko silných efektov a veľa iných prediktorov nemá na odozvu žiaden vplyv.




Obr. 28: Stratová funkcia má minimum v $\hat{\beta}_{LS}$, tam sa dosahuje najlepší fit. Penalta $\lambda$ však posúva optimum bližie k nule (scvrkáva parametre), a to na kružnicu (vľavo ridge: $\sum_{j=1}^{p} \beta_j^2 = t^2$) alebo štvorec (vpravo LASSO : $\sum_{j=1}^{p} |\beta_j| = t$). Zdroj: Faraway (2014)

# 11 Kategorické prediktory

## 11.1 T-test

Začnime zopakovaním si klasického t-testu na rovnosť strdných hodnôt. Predstavme si, že sme v situácii, že chceme porovnať, či majú dve skupiny, skupina A a skupina B, rovnakú strednú hodnotu normálne rozdelenej skalárnej odozvy $y$. Pozorujeme dátové vzorky veľkosti $n_A$ a $n_B$ a prepokladáme, že $y_A$ a $y_B$ majú rovnaké variancie $\sigma_A^2 = \sigma_B^2$. Neznáme stredné hodnoty $\mu_A$ a $\mu_B$ vieme konzistentne odhadnúť pomocou aritmetického priemeru, ktorý má normálne rozdelenie. Ak by sme aj neuvažovali predpoklad normality rozdelenia $y_A$ a $y_B$ tak Centrálna limitná veta nám hovorí, že správne naškálovaný aritmetický priemer má asymptoticky normálne rozdelenie. Môžeme ale nemusíme uvažovať, že tieto naše dve skúmané skupiny majú rovnakú varianciu odozvy. Varianciu vieme odhadnúť konzistentne pomocou výberovej variancie, ktorá je za predpokladu normality rozdelená ako $\chi^2_(n-1)$ (alebo aspoň asymptoticky ak nepredpokladáme normalitu), kde $n$ je veľkosť dátovej vzorky.

Studentov t-test pre rovnosť stredných hodnôt, za predpokladu rovnakých variancií $\sigma_A^2 = \sigma_B^2$. $H_0 : \mu_A = \mu_B$ voči obojstrannej alternatívnej hypotéze $H_1 : \mu_A \neq \mu_B$.

Na testovanie $H_0$ môžeme použiť fakt, že nasledovná štatistika má $t$ rozdelenie s $(n_A + n_B - 2)$ stupňami voľnosti.

$$T = \frac{(\bar{y}_A - \bar{y}_B) - (\mu_A - \mu_B)}{S_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} \sim t_{n_A+n_B-2},$$

kde $S_p^2 = \frac{(n_A-1)S_A^2+(n_B-1)S_B^2}{n_A+n_B-2}$ a $S_A^2$ a $S_B^2$ sú výberové variancie.

## 11.2 Teraz naspäť k regresii

Uvažujme nasledovný lineárny regresný model bez interceptu s normálne rozdelenými homoskedastickými chybami

$$y_i = \beta_A d_{iA} + \beta_B d_{iB} + \epsilon$$

kde $d_{iA} = 1$ ak $i$-te pozorovanie patrí do skupiny A a 0 inak. Podobne $d_{iB} = 1$ ak $i$-te pozorovanie patrí do skupiny B a 0 inak.

Všimnime si, že pre skupinu A je na základe lineárneho regresného modelu odozva $y_A \sim N(\beta_A, \sigma^2)$ a $y_B \sim N(\beta_B, \sigma^2)$. Toto sú úplne identické predpoklady ako v prípade $t$-testu vyššie!

Pomocou metódy najmenších štvorcov môžeme odhadnúť parametre $\beta_A = \mu_A$ a $\beta_B = \mu_B$ a tie nebudú nič iné ako aritmetické priemery a teda $\hat{\beta}_A = \bar{y}_A$ a $\hat{\beta}_B = \bar{y}_B$. Z regresnej tabuľky hneď vieme vyčítať, či sú tieto priemery signifikantne rôzne od nuly. To je fajn ale nás teraz zaujíma rozdiel týchto stredných hodnôt.

Ak by sme pridali intercept to hore uvedeného modelu dostali by sme

$$y_i = \beta_0 + \beta_A d_{iA} + \beta_B d_{iB} + \epsilon$$

a mali by sme problém s multikolinearitou, a to preto lebo $d_{iA} + d_{iB} = 1$ a teda regresory sú perfektne korelované.

Môžme si zvoliť jednu skupinu za referenčnú, nech je to skupina B a potom odhadnúť nasledovný model

$$y_i = \beta_0 + \beta d_{iA} + \epsilon$$

Teraz ak $d_{iA} = 1$ tak $y_i \sim N(\beta_0 + \beta, \sigma^2)$ a ak $d_{iA} = 0$ tak $y_i \sim N(\beta_0, \sigma^2)$, takže koeficient $\beta$ má teraz inú interpretáciu, a to, že o koľko je v priemere odozva $y_A$ väčšia ako $y_B$.

Testovanie signifikantnosti tohoto parametra $\beta$ je teda úplne totožné s t-testom vyššie. Tu vidíme, aký bohatý je lineárny regresný model. T-test na testovanie rovnosti stredných hodnôt za predpokladu rovnakej variancie (preto predpoklad homoskedasticity v regresnom modeli) je len jeho špeciálnym prípadom.

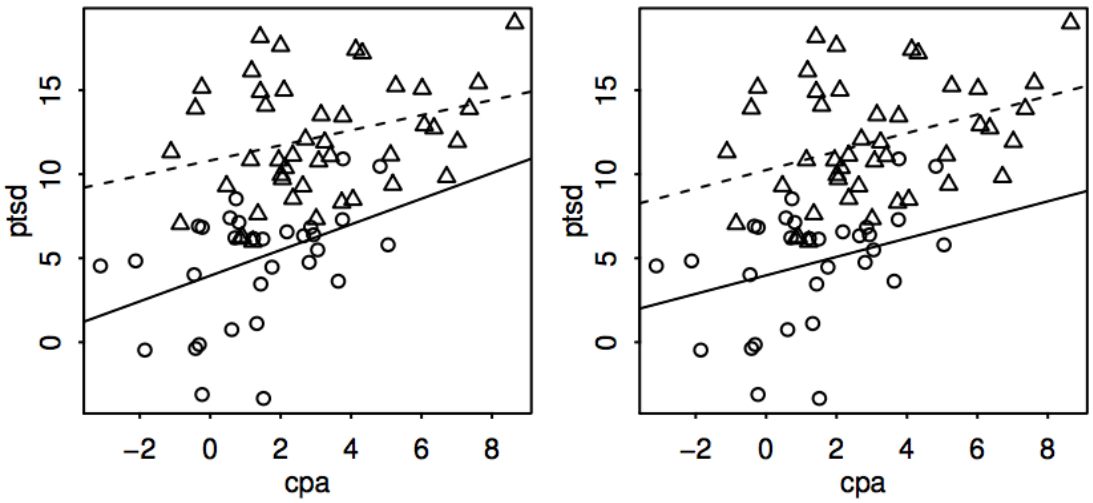
Uvažujme model, kde je okrem kategorickej premennej aj ordinálna

$$y_i = \beta_0 + \beta_A d_{iA} + \beta_X x_i + \epsilon$$

tu umožňujeme aby stredná hodnota závisela nielen od skupiny ale aj od hodnoty $x_i$. Skupina A a skupina B však majú rovnakú citlivosť strednej hodnoty na $x$. Flexibilnejšou alternatívou je model

$$y_i = \beta_0 + \beta_A d_{iA} + \beta_X x_i + \beta_{XA} x_i d_{iA} + \epsilon.$$

V tomto modeli $y_{iA} \sim N(\beta_0 + \beta_A + (\beta_X + \beta_{XA})x_i, \sigma^2)$ a $y_{iB} \sim N(\beta_0 + \beta_X x_i, \sigma^2)$.




Obr. 29: (Vľavo flexibilnejší model s interakčným členom, citlivosť pre dve skupiny môže byť rôzna. Vpravo model bez interakčného člena, stredná hodnota je posunutá o konštantu. Zdroj: Faraway 2014

V prípade, že chceme interpretovať parametre v modeli s interakčným členom je výhodné si vycentrovať ordinálnu premennú pomocou nejakej typickej hodnoty (či už stredná hodnota alebo medián). Koeficient pri dummy premennej sa niekedy ťažko interpretuje, nakoľko na interpretáciu potrebujeme, aby bola ordinálna premenná 0 a to niekedy nie je zmysluplné, napríklad, ak je ordinálnou premennou vek.

Pre kategorickú premennú, ktorá nadobúda viacej ako 2 hodnoty použijeme viacej dummy premenných, konkrétne počet hodnôt mínus jeden.

# 12 Čriepky z elementárnej pravdepodobnosti a štatistiky

## 12.1 Spojite rozdelený náhodný vektor

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

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \tag{12.1}$$

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

$$E(X) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_n) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} = \mu \tag{12.2}$$

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

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

teda jedná sa o maticu **čísiel** a nie náhodných premenných.

Pre lineárnu transformáciu náhodného vektora $Y = AX$ platí

$$\begin{aligned} E(Y) &= E(AX) = AE(X) = A\mu, \\ Var(Y) &= E\left[(Y-E(Y))(Y-E(Y))^T\right] = \\ &= E\left[(AX-A\mu)(AX-A\mu)^T\right] \\ &= E\left[A(X-\mu)(X-\mu)^T A^T\right] = \\ &= A\ E\left[(X-\mu)(X-\mu)^T\right] A^T = \\ &= AVar(X)A^T = A\Sigma A^T, \end{aligned} \tag{12.3}$$

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 $(\mu, \Sigma)$ má nasledovnú funkciu hustoty

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

kde $x$ označuje stĺpcový vektor.

## 12.2 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 $Y \equiv X_1^2 + \dots X_k^2$ 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 $\beta_i$ 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.

## 12.3 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 $\alpha$, 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 $\alpha$ 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 $\alpha$ 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 $\forall i : E(X_i) = \mu, Var(X_i) = \sigma^2 = 1$.
- 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: $\sqrt{n}\frac{X_1+\cdots+X_n}{n} = 0.353 \leq 1.96$ preto nezamietame $H_0$.
- P-hodnota: $1 - \Phi(0.353) = 1 - 0.638 = 0.362 \geq 0.05$ preto nezamietame $H_0$.
- Interval spoľahlivosti pre neznámy parameter $\mu$: $CI = (\frac{X_1+\cdots+X_n}{n} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \frac{X_1+\cdots+X_n}{n} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}) = (-0.567, 0.818)$. $0 \in CI$, preto nezamietame nulovú hypotézu.

## 12.4 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 $\beta_0$, ktorý vygeneroval dáta. Odhad ML je **konzistentný**, takže

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

teda konverguje podľa pravdepodobnosti ku skutočnej hodnote $\beta_0$, $\forall \epsilon > 0 : P(|\hat{\beta}^n_{ML} - \beta_0| < \epsilon) \to 0$ 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 $\theta_0$.

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 $\beta$. 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 $\beta_{ML}$ 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

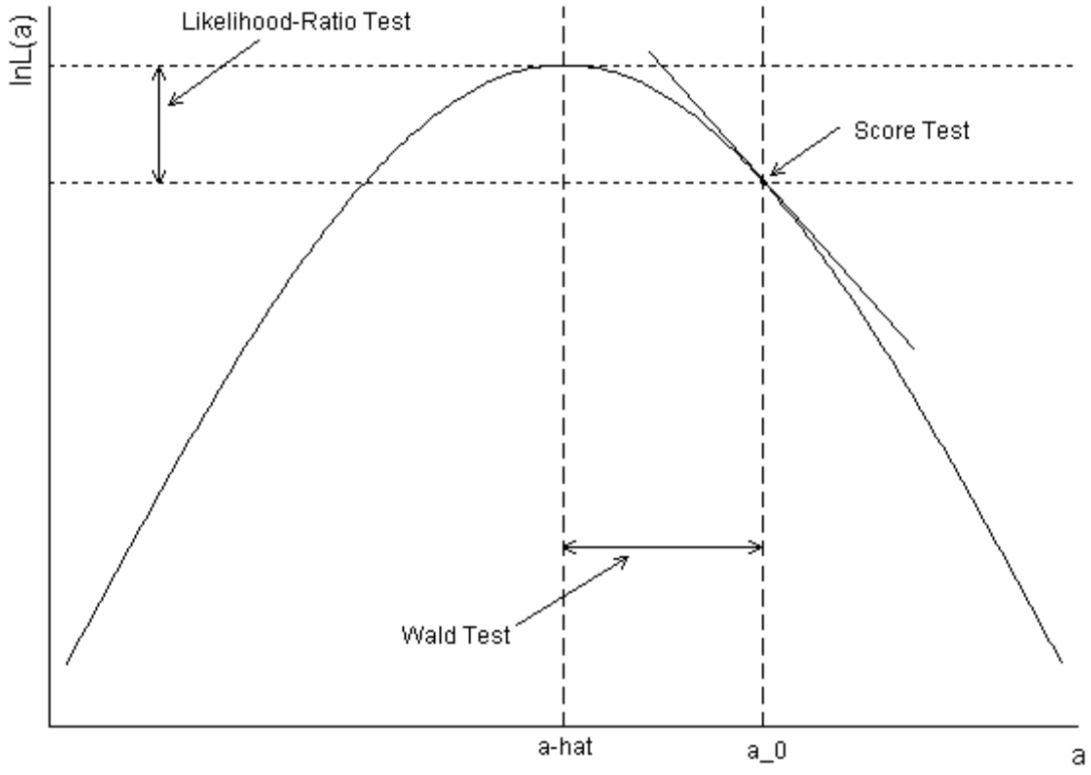
$$\sqrt{n}(\hat{\beta}_{ML} - \beta_*) \to_D N\left(0, H(\beta_*)^{-1} I(\beta_0) H(\beta_*)^{-1}\right),$$

kde $\beta_*$ je minimizátor Kuhlback-Leiblerovej divergencie.

Na základe likelihood funkcie máme **tri typy testov** na porovnávanie dvoch vnorených modelov, každý je asymptoticky rozdelený ako $\chi^2$. 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 $\chi^2$ 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. 30.




Obr. 30: Porovnanie testov založených na likelihoode, zdroj: [Fox97].

# Literatúra


[Don11] Rafe Donahue. *Fundamental Statistical Concepts in Presenting Data - Principles for Constructing Better Graphics.* 2011.

[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.

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

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

[HS33] Harold Hotelling and Resena De H Secrist. The triumph of mediocrity in business. *Journal of the American Statistical Association*, 28(184):463–465, 1933.

[SHR+34] Horace Secrist, Harold Hotelling, MC Rorty, Corrado Gini, and Willford I King. Open letters, 1934.

[Sti86] Stephen M Stigler. *The history of statistics: The measurement of uncertainty before 1900.* Harvard University Press, 1986.

[Tuk77] John W Tukey. Exploratory data analysis. 1977.

[Wic09] Hadley Wickham. *ggplot2: elegant graphics for data analysis.* Springer Science & Business Media, 2009.