Regrese

Dále se budeme zabývat odhadem funkční závislosti hodnot jedné veličiny na hodnotách druhé, tj. budeme hledat funkci f tak, aby pro pozorované hodnoty $Y_1, \ldots, Y_n$ platilo $Y_i = f(X_i) + e_i$, kde $X_1, \ldots, X_n$ jsou hodnoty nezávisle proměnné, e_i jsou náhodné odchylky pozorovaných dat od ideální křivky.

Statistika bohužel neumí navrhnout optimální tvar této křivky (tj. to bude přímka, polynom, exponenciela, ..), umí pouze k předem zvolenému obecnému tvaru f dopočítat optimální parametry, odhadnout chybu při položení křivky a porovnat, zda je jedna konkrétní křivka pro daná data lepší než jiná.

Lineární regresní model

máme vektor pozorování $\mathbb{Y} = (Y_1, \dots, Y_n)'$

matice daných čísel $\mathbb{X}$ typu $n \times k$ – budeme předpokládat, že hodnost $\mathbb{X}$ je k a že $n > k$

$\beta = (\beta_1, \dots, \beta_k)'$ vektor neznámých parametrů

$e = (e_1, \dots, e_n)'$ vektor náhodných chyb, pro který platí $Ee = 0$, $\operatorname{var} e = \sigma^2 \mathbb{I}$

(tj. e_i jsou nezávislé náhodné veličiny, $Ee_i = 0$, $\operatorname{var} e_i = \sigma^2$)

model. $\mathbb{Y} = \mathbb{X}\beta + e$

platí $E\mathbb{Y} = \mathbb{X}\beta$, $\operatorname{var} \mathbb{Y} = \sigma^2 \mathbb{I}$

například pro $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$ bude matice

$$\mathbb{X} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix} \quad \text{a} \quad \beta = (\beta_0, \beta_1, \beta_2)'$$

pak opravdu $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i$

kdybych naopak chtěla položit křivku $f(x) = \beta_0 + \beta_1 \log x$,

zvolila bych

$$\mathbb{X} = \begin{pmatrix} 1 & \log x_1 \\ 1 & \log x_2 \\ \vdots & \vdots \\ 1 & \log x_n \end{pmatrix} \quad \text{a} \quad \beta = (\beta_0, \beta_1)'$$

a obecný model by byl pořád stejný $\mathbb{Y} = \mathbb{X}\beta + e$

terminologie: pojem lineární model znamená lineární závislost na neznámých parametrech, ne na hodnotách x_i

v modelu je třeba odhadnout neznámé parametry β a σ^2

odhaduje se metodou nejmenších čtverců, aby výraz

$$S_e = \mathbf{e}'\mathbf{e} = (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta) \text{ byl minimální}$$

připíšeme $S_e = \mathbf{e}'\mathbf{e} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{k} x_{ij}\beta_j\right)^2$

položíme

$$\frac{\partial S_e}{\partial \beta_1} = -2 \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{k} x_{ij}\beta_j\right) x_{i1} = 0$$

$$\vdots$$

$$\frac{\partial S_e}{\partial \beta_k} = -2 \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{k} x_{ij}\beta_j\right) x_{ik} = 0$$

dostaneme

$$\sum_{i=1}^{n} Y_i x_{i1} = \sum_{i=1}^{n} \sum_{j=1}^{k} x_{ij}\beta_j x_{ik}$$

$$\vdots$$

$$\sum_{i=1}^{n} Y_i x_{ik} = \sum_{i=1}^{n} \sum_{j=1}^{k} x_{ij}\beta_j x_{ik}$$

to je maticově $\mathbf{X}'\mathbf{Y} = \mathbf{X}'\mathbf{X}\beta$ – tomu se říká soustava normálních rovnic

řešení je $\mathbf{b} = \hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ pokud $\mathbf{X}'\mathbf{X}$ je regulární, a to v našem případě je, protože $\mathbf{X}$ má plnou hodnost k

$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b}$ je odhad (aproximace) vektoru $\mathbf{Y}$

platí $E\mathbb{b}=\beta$, $\quad \operatorname{var} \mathbb{b} = \sigma^2(\mathbb{X}'\mathbb{X})^{-1}$, $\quad E\hat{\mathbb{Y}} = E\mathbb{X}\mathbb{b} = \mathbb{X}E\mathbb{b} = \mathbb{X}\beta = E\mathbb{Y}$

důkaz: $\quad E\mathbb{b} = E\left[(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{Y}\right] = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'E\mathbb{Y} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{X}\beta = \beta$

$$\operatorname{var} \mathbb{b} = \operatorname{var}\left[(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{Y}\right] = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}' \operatorname{var} \mathbb{Y}\, \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1} =$$

$$= (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\sigma^2\mathbb{I}\,\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1} = \sigma^2(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1} = \sigma^2(\mathbb{X}'\mathbb{X})^{-1}$$

předpokládejme navíc, že $\mathbb{Y} \sim N(\mathbb{X}\beta, \sigma^2\mathbb{I})$

pak $\mathbb{b} \sim N(\beta, \sigma^2(\mathbb{X}'\mathbb{X})^{-1})$

veličina $R = \hat{S}_e = (\mathbb{Y}-\mathbb{X}\mathbb{b})'(\mathbb{Y}-\mathbb{X}\mathbb{b})$ se nazývá <u>reziduální součet čtverců</u>

představuje odhad chyby, které se dopustíme, když vektor neznámých parametrů β nahradíme jeho odhadem $\mathbb{b}$

dá se dokázat, že $ER = (n-k)\sigma^2$

a tedy $s^2 = \dfrac{R}{n-k}$ je nestranný odhad rozptylu σ^2

a že $\dfrac{R}{\sigma^2} \sim \chi^2_{n-k}$

a navíc $\mathbb{b}$ a R jsou nezávislé

(dokazovat to nebudem, je k tomu potřeba pár věcí z teorie matic, bez nich to asi nejde)

veličina s^2 se nazývá <u>reziduální rozptyl.</u>

Dále se budeme všímat jednotlivých složek b_i odhadu $\mathbb{b}$.

jelikož $\mathbb{b} \sim N(\beta, \sigma^2(\mathbb{X}'\mathbb{X})^{-1})$, pak $b_i \sim N(\beta_i, \sigma^2 v_{ii})$

kde $v_{ii} = (\mathbb{X}'\mathbb{X})^{-1}_{ii}$ ... prvek matice z toho $U_i = \dfrac{b_i - \beta_i}{\sigma\sqrt{v_{ii}}} \sim N(0,1)$

dále $\chi = \dfrac{(n-k)s^2}{\sigma^2} = \dfrac{R}{\sigma^2} \sim \chi^2_{n-k}$

proto $T_i = \dfrac{U_i}{\sqrt{\dfrac{\chi}{n-k}}} \sim t_{n-k}$

dosadíme: $T_i = \dfrac{\dfrac{b_i - \beta_i}{\sigma\sqrt{v_{ii}}}}{\sqrt{\dfrac{s^2}{\sigma^2}}} = \dfrac{b_i - \beta_i}{\sqrt{s^2 v_{ii}}}$

na tyto veličiny se pak dají založit testy hypotéz o parametrech β_i, např. $H: \beta_i = 0$ proti $A: \beta_i \neq 0$

třeba v případě aproximace polynomem $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$ testem $H: \beta_2 = 0$ zjistím, jestli by nestačilo vzít přímku $f'(x) = \beta_0' + \beta_1' x$

Příklady

a) proložení přímky $f(x) = \beta_0 + \beta_1 x$

$$\mathbb{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \qquad \mathbb{X}'\mathbb{X} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}$$

$$(\mathbb{X}'\mathbb{X})^{-1} = \frac{1}{n\sum x_i^2 - (\sum x_i)^2} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix} \qquad \mathbb{X}'\mathbb{Y} = \begin{pmatrix} \sum Y_i \\ \sum x_i Y_i \end{pmatrix}$$

odhad $\mathbb{B} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{Y} = \dfrac{1}{n\sum x_i^2 - (\sum x_i)^2}\begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix}\begin{pmatrix} \sum Y_i \\ \sum x_i Y_i \end{pmatrix} =$

$$= \begin{pmatrix} \dfrac{\sum x_i^2 \sum Y_i - \sum x_i \sum x_i Y_i}{n\sum x_i^2 - (\sum x_i)^2} \\ \dfrac{n\sum x_i Y_i - \sum x_i \sum Y_i}{n\sum x_i^2 - (\sum x_i)^2} \end{pmatrix} = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix}$$

pro odhad rozptylu $s^2 = \dfrac{R}{n-2}$ použijeme vzorec $R = \mathbb{Y}'\mathbb{Y} - \mathbb{b}'\mathbb{X}'\mathbb{Y}$,

který se snadno dokáže, a máme

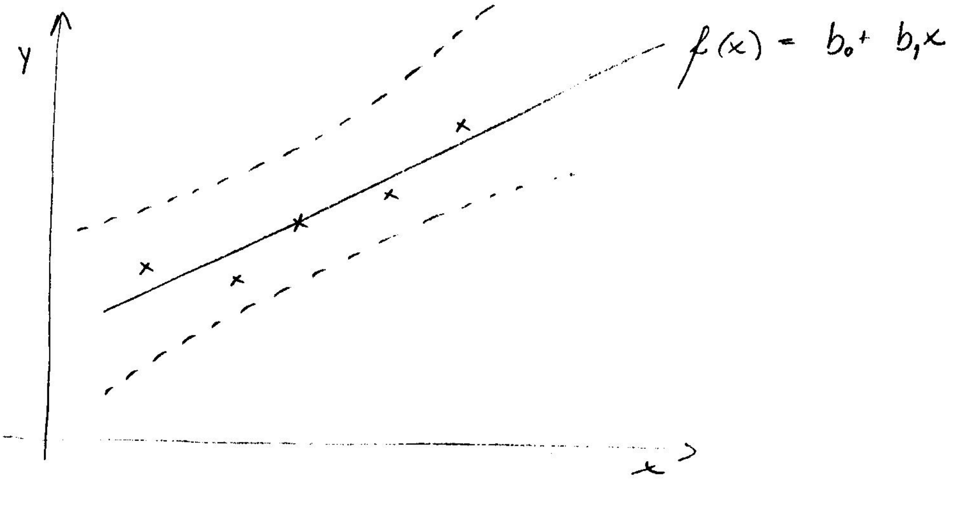
$$s^2 = \frac{1}{n-2}\left(\sum Y_i^2 - b_0 \sum Y_i - b_1 \sum x_i Y_i\right)$$

testové kritérium pro $H: \beta_1 = 0$ proti $A: \beta_1 \neq 0$

$$\frac{|b_1|\sqrt{\sum x_i^2 - n\bar{x}^2}}{\sqrt{s^2}} \geq t'_{n-2}(\alpha) \ \ldots \text{ kritická hodnota}$$

dal by se taky udělat interval spolehlivosti pro β_1, ale ten nás moc nezajímá

spíš by bylo zajímavé mít intervalový odhad pro $\beta_0 + \beta_1 x$, tj. pro každý bod na přímce, čímž by se vytvořil pás spolehlivosti



takhle nějak by vypadal

zavedu vektor $c = \begin{pmatrix} 1 \\ x \end{pmatrix}$

pak $\beta_0 + \beta_1 x = c'\beta$ a odhad $b_0 + b_1 x = c'b$

platí: $E c'b = c' E b = c'\beta$

$\operatorname{var} c'b = c'(\operatorname{var} b) c = \sigma^2 c'(X'X)^{-1} c$

$c'b$ má normální rozdělení

$c'b$ a s^2 jsou nezávislé

proto $$\frac{\dfrac{c'b - c'\beta}{\sqrt{\sigma^2 c'(X'X)^{-1} c}}}{\sqrt{\dfrac{s^2}{\sigma^2}}} = \frac{c'b - c'\beta}{\sqrt{s^2 c'(X'X)^{-1} c}} \sim t_{n-k}$$ (v našem případě t_{n-2})

intervalový odhad na hladině $1-\alpha$ pro $c'\beta$ tedy bude obecně

$$\left(c'b - t'_{n-k}(\alpha)\sqrt{s^2 c'(X'X)^{-1} c} \; ; \; c'b + t'_{n-k}(\alpha)\sqrt{s^2 c'(X'X)^{-1} c} \right)$$

a v našem případě

$$b_0 + b_1 x \pm t'_{n-2}(\alpha)\sqrt{s^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum x_i^2 - n\bar{x}^2}\right)}$$

b) proložení polynomu $f(x) = \beta_0 + \beta_1 x + \dots + \beta_k x^k$

vezme se matice

$$X = \begin{pmatrix} 1 & x_1 & \dots & x_1^k \\ 1 & x_2 & \dots & x_2^k \\ \vdots & \vdots & & \vdots \\ 1 & x_n & & x_n^k \end{pmatrix}$$

a tím je to teoreticky hotovo, prakticky ovšem nastávají potíže s rostoucím stupněm polynomu začíná být matice $X'X$ špatně podmíněná (má skoro nulový determinant, blíží se singulární matici)

výpočet $(X'X)^{-1}$ tedy může být značně nepřesný %

místo přímé inverze se doporučuje řešit soustavu normálních rovnic $X'X\beta = X'Y$ pomocí numerických metod, nebo se bez nich zcela obejít a řešit minimalizaci S od začátku numericky:
ortogonální transformace

Gramm-Schmidtova metoda

Householderova metoda

Givensova metoda

rozklad na singulární hodnoty

dají se najít např. v Ralstonovi (Základy numerické matematiky) nebo Anděl - Vorlíčková: Vybrané metody statistické analýzy dat

přímá inverze se nedoporučuje asi od stupně polynomu 7

<u>Problém určení optimálního stupně polynomu</u>

možnosti: 1) začít stupněm $p = 0$ a postupně ho zvyšovat
přitom každý přidaný koeficient testovat, zda je různý od nuly
přestat v okamžiku, kdy přidaný člen je statisticky nevýznamný

2) taktéž od konce - začít s maximálním možným p a postupně snižovat

%. Příklady špatně podmíněných soustav.

a) $2x + 6y = 8$

$2x + 6{,}00001y = 8{,}00001$

řešení: $x = 1$

$y = 1$

b) $2x + 6y = 8$

$2x + 5{,}99999y = 8{,}00002$

řešení: $x = 10$

$y = -2$

Soustavy vypadají téměř totožné, přesto se řešení značně liší.
Je třeba počítat s maximální přesností, nebezpečné zaokrouhlování vede k nesmyslným výsledkům.

V případě a) má matice soustavy A determinant

$$\det A = 2 \cdot 6{,}00001 - 2 \cdot 6 = 0{,}00002$$

inverzní matice je

$$A^{-1} = \frac{1}{\det A} \begin{pmatrix} 6{,}00001 & -6 \\ -2 & 2 \end{pmatrix} = \begin{pmatrix} 300\,000{,}5 & -300\,000 \\ -100\,000 & 100\,000 \end{pmatrix}$$

řešení:

$$\begin{pmatrix} x \\ y \end{pmatrix} = A^{-1} \begin{pmatrix} 8 \\ 8{,}00001 \end{pmatrix} = \begin{pmatrix} 2\,400\,004 - 2\,400\,003 \\ -800\,000 + 800\,001 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

kdybychom zaokrouhlili $8{,}00001$ na 8:

$$\begin{pmatrix} x \\ y \end{pmatrix} = A^{-1} \begin{pmatrix} 8 \\ 8 \end{pmatrix} = \begin{pmatrix} 2\,400\,004 - 2\,400\,000 \\ -800\,000 + 800\,000 \end{pmatrix} = \begin{pmatrix} 4 \\ 0 \end{pmatrix}$$ značná chyba

Stupně polynomu

%. Odhad se dá např. založit na hodnotách S_k^2 počítaných postupně v modelech

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \dots + \beta_k X^k$$

pro $k = 1, 2, \dots$

Ve správně zvoleném modelu je S^2 nestranným odhadem σ^2.

Správný stupeň polynomu k_0 se určí tak, aby platilo

$ES_k^2 > \sigma^2$ pro $k < k_0$ a $ES_k^2 = \sigma^2$ pro $k \geq k_0$.

Zhruba se dá poznat z grafu hodnot S_k^2, který od k_0 začíná být konstantní.

Není vhodné volit stupeň polynomu vyšší, než je třeba, protože se tím do modelu vnáší příliš mnoho neznámých parametrů, a to nebývá dobré. Také je třeba si uvědomit, že měření Y_i mohou být doprovázena šumem, takže my je vlastně ani nechceme vystihnout absolutně přesně, protože tím bychom ten šum vlastně zabudovali do modelu.

Jaký vliv bude mít porušení těchto předpokladů na výsledky?

Prakticky bude použitelný jenom odhad parametru β, tj.

$\mathbb{b} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{Y}$, a budeme vědět, že je nestranný.

Až sem se žádné další předpoklady nepoužily.

Ostatní výsledky (odhad rozptylu, testy hypotéz...) mohou být zkreslené.

(V případě analýzy složitosti algoritmů bude nejčastěji porušen předpoklad o konstantním rozptylu, protože rozptyl doby výpočtu s rostoucím rozsahem výběru obvykle taky roste.)

Literatura: K. Zvára: Regresní analýza

Academia, Praha 1989

Při porušení předpokladu normality chyb – existují robustní metody odhadu.

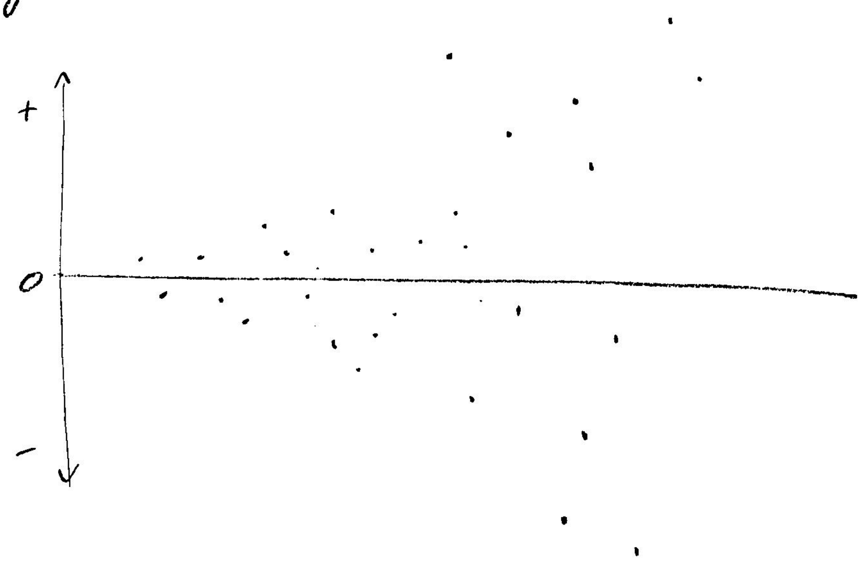
Ty se používají i v případech, kdy se v pozorováních objevují nějaké extrémy, vybočující hodnoty, odlehlá pozorování.

Při nekonstantním rozptylu se dá někdy použít metoda vážených nejmenších čtverců, kde $\operatorname{var} e = \sigma^2\mathbb{K}$ $(\neq \sigma^2\mathbb{I})$. Umíme-li matici $\mathbb{K}$ nějak odhadnout, pak minimalizujeme $(\mathbb{Y}-\mathbb{X}\beta)'\mathbb{K}^{-1}(\mathbb{Y}-\mathbb{X}\beta)$

a řešení bude ve tvaru $\hat{\beta} = (\mathbb{X}'\mathbb{K}^{-1}\mathbb{X})^{-1}\mathbb{X}'\mathbb{K}^{-1}\mathbb{Y}$.

Není-li matice $\mathbb{K}$ diagonální, vyřeší se tím zároveň i porušení předpokladu nezávislosti pozorování.

O kvalitě regrese dále vypovídají samotná residua, která by teoreticky měla odpovídat chybám e_i, o kterých se předpokládalo, že jsou to navzájem nezávislé náhodné veličiny s rozdělením $N(0,\sigma^2)$. To se dá buď statisticky otestovat, nebo alespoň vizuálně posoudit z grafu residuí. Ten může vypadat např. takto:



Co se z toho dá poznat? Třeba to, že chyby mají skutečně nulovou střední hodnotu, ale rozptyl není konstantní, roste.

Další podezřelé věci by mohly být třeba:

odlehlé (extrémní) hodnoty

nějaký typ závislosti nebo trend v residuích

převaha kladných (nebo záporných) hodnot, špatné střídání znamének

nějaká náhlá změna v grafu

Další poznámky k regresi:

Zatím jsme jako měřítko kvality regresního modelu měli reziduální součet čtverců $RSS = \sum_i (y_i - \hat{y}_i)^2$, čím je menší, tím je model lepší. Jak se ale pozná, jestli je RSS dost malé?

Jinou mírou kvality je koeficient determinace

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

který by měl být naopak blízký 1

Vysvětlení: $\sum_i (y_i - \bar{y})^2$ je celková variabilita vektoru Y

$\sum_i (y_i - \hat{y}_i)^2$ je část variability způsobená chybou regrese

zbytek je variabilita objasněná regresním modelem

podíl $\frac{RSS}{\sum_i (y_i - \bar{y})^2}$ tedy říká, jaký díl celkové variability je způsoben chybou

R^2 udává, jaký díl celkové variability je vysvětlený použitým regresním modelem

R^2 se obvykle vyjadřuje v procentech

Dalším měřítkem kvality modelu je tzv. jeho predikční schopnost. Data se rozdělí na dvě poloviny $(\underbrace{Y_1, \dots, Y_{n/2}}_{M_1})$ a $(\underbrace{Y_{n/2+1}, \dots, Y_n}_{M_2})$. Odhady se vypočtou z té první poloviny, pak se zkontroluje, jestli vyhovují i pro tu druhou polovinu. Např. podle kritéria

$$K = \frac{RSS}{\sum_{i \in M_1} \left(y_i - f(x_i, b(M_2))\right)^2 + \sum_{i \in M_2} \left(y_i - f(x_i, b(M_1))\right)^2}$$

kde $f(x_i, b(M_j))$ je hodnota regresní fce v bodě x_i při parametrech odhadnutých z části M_j.

Hodnota K by se v ideálním případě měla blížit 1.

Konečně kritériem je i shoda s praktickými požadavky, např. aby doba výpočtu nepřesahovala rozumnou mez.

Detekce vlivných pozorování v lineární regresi

Jak poznat mezi daty neočekávané hodnoty?

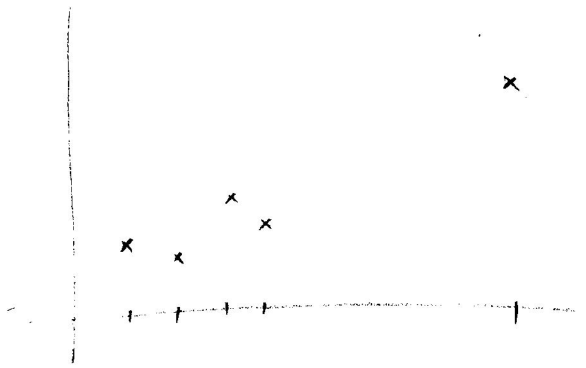
Jak rozhodnout, zda mohou významně ovlivnit statistickou analýzu a jakým způsobem?

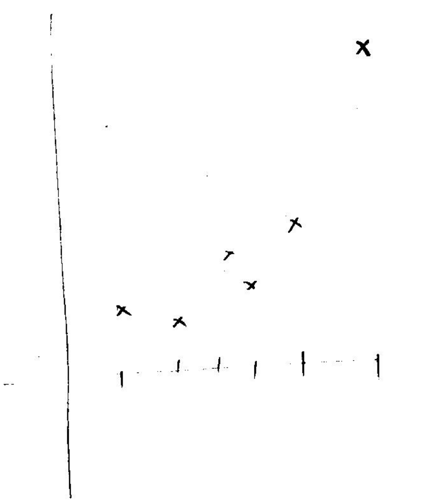
Rozeznáváme tzv. odlehlá pozorování – to jsou neočekávané hodnoty ve směru Y

vybočující body – neočekávané body ve směru X

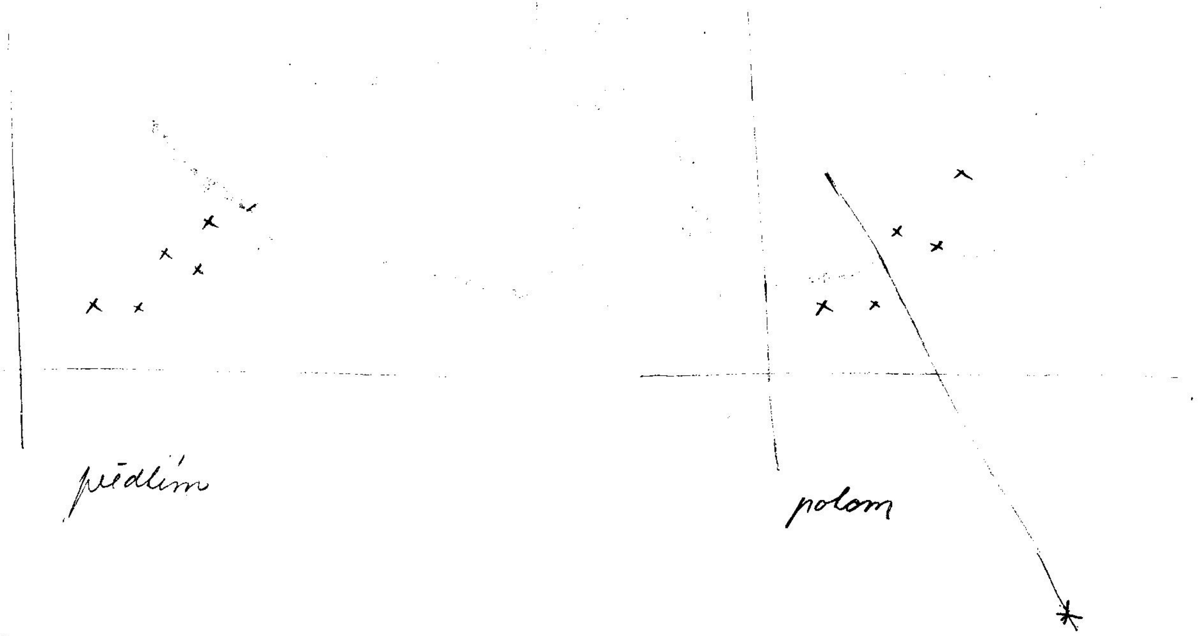
(někdy může jít o oba typy zároveň).

Jejich výskyt někdy neovlivní analýzu dat, protože "zapadají do modelu", např.





někdy je určení podstatné, např.



V tomto případě je nazýváme vlivnými body.

Detekce odlehlých pozorování

a) vizuálně z grafu nebo z grafu reziduí

b) statistickými testy, např.

nechť r_i jsou rezidua, $s^2 = \frac{S_e}{n-p}$ (r_i jsou prvky $r = Y - \hat{Y}$)

(S_e je reziduální součet čtverců, p je počet neznámých parametrů)

h_{ii} je prvek matice $H = X(X'X)^{-1}X'$ (projekční matice)

$t_i = \frac{r_i}{s\sqrt{1-h_{ii}}}$ jsou tzv. standardizovaná rezidua

$t_i^* = \frac{r_i}{s_{(i)}\sqrt{1-h_{ii}}}$, kde $s_{(i)} = \frac{Y_{(i)}'(I - X_{(i)}(X_{(i)}'X_{(i)})^{-1}X_{(i)}')Y_{(i)}}{n-p-1}$

jsou tzv. studentizovaná rezidua

($X_{(i)}$ je matice X po vypuštění i-tého řádku, $Y_{(i)}$ je vektor Y po vypuštění i-tého prvku)

Platí: t_i^* mají t-rozdělení o $n-p-1$ st. volnosti.

Nechť hypotéza H_0 : i-té pozorování odpovídá lineár. modelu jako ostatní

H_0 se zamítá, když $|t_i^*| > t_{n-p-1}(\alpha)$.

Mezi t_i^* a t_i platí vztah

$$t_i^* = t_i \sqrt{\frac{n-p-1}{n-p-t_i^2}}$$

Detekce vybočujících bodů

Je založena na hodnotách h_{ii}, pro které vždy platí $0 \leq h_{ii} \leq 1$.

Ví se, že pokud mají jednotlivá pozorování Y_i přispívat k odhadu stejnoměrně, měla by být $h_{ii} \doteq \frac{p}{n}$.

Hodnoty h_{ii} blízké k 1 zpravidla odpovídají vybočujícím bodům.

Detekce vlivných bodů

Jednou z používaných měr je <u>Cookova vzdálenost</u>

$$D_i = \frac{(\hat{Y} - \hat{Y}_{(i)})'(\hat{Y} - \hat{Y}_{(i)})}{p s^2} = \frac{(\hat{\beta} - \hat{\beta}_{(i)})' X'X (\hat{\beta} - \hat{\beta}_{(i)})}{p s^2} =$$

$$= \frac{t_i^2}{p} \cdot \frac{h_{ii}}{1-h_{ii}}$$

Velké hodnoty D_i ukazují výrazný vliv i-tého pozorování na odhad $\hat{\beta}$. V praxi se většinou D_i srovnává s kritickými hodnotami rozdělení $F_{p,n-p}$, i když to není úplně přesné.

Zobecnění pro měření vlivu více pozorování na odhad β:

$$D_{i_1,\dots,i_k} = \frac{(\hat{\beta}-\hat{\beta}_{(i_1,\dots,i_k)})' \mathbb{X}'\mathbb{X} (\hat{\beta}-\hat{\beta}_{(i_1,\dots,i_k)})}{p s^2}$$

kde $\hat{\beta}_{(i_1,\dots,i_k)}$ je odhad β metodou nejmenších čtverců po vypuštění pozorování $i_1,\dots,i_k$.

Hodnoty se opět porovnávají s kritickými hodnotami $F_{p,n-p}$.

<u>Robustní Cookova vzdálenost</u> – ve vzorci pro D_i se s^2 nahradí nahradí nějakým robustním odhadem rozptylu σ^2.

<u>Welschova – Kuhova vzdálenost</u>

$$DFITS_i = \frac{|\mathbb{x}_i'(\hat{\beta}-\hat{\beta}_{(i)})|}{s_{(i)}\sqrt{h_{ii}}} = |t_i^*| \sqrt{\frac{h_{ii}}{1-h_{ii}}}$$

kde $\mathbb{x}_i'$ je i-tý řádek matice $\mathbb{X}$

Velké hodnoty (větší než $\sqrt{\frac{4p}{n}}$) svědčí o významném vlivu i-tého pozorování.

Tato míra odráží vliv i-tého pozorování nejen na odhad β, ale i na odhad σ^2.

Existují další metody, které třeba měří vliv na jednotlivé složky vektoru β apod.

Po detekci "podezřelých" hodnot se musíme rozhodnout co s nimi, zda je vyloučit (mohou to být chyby) nebo je zahrnout do modelu (v tom případě je dobré použít nějaké robustní metody) nebo se jim věnovat zvlášť (mohou ukazovat na velmi zajímavé jevy).