Univerzita Pavla Jozefa Šafárika v Košiciach,
Prírodovedecká fakulta

Strojové učenie
Lineárne modelovanie, maximálna dôveryhodnosť

Gabriela Andrejková, Ľubomír Antoni

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA

Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje

www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Úvod

- Lineárna regresia – základný algoritmus strojového učenia, jednoducho analyzuje naše údaje pomocou preloženej nadroviny s najmenšou chybou.
- Uvidíme lineárnu regresiu z inej perspektívy, a síce z hľadiska pravdepodobnosti, najmä pomocou odhadu maximálnej dôveryhodnosti - Maximum Likelihood (ML).
- Budeme explicitne modelovať **šum** (chyby medzi modelom a pozorovaniami) v dátach pomocou **náhodnej premennej**.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Pripomenutie lineárnej regresie

Sú dané dáta – dvojice vstupných a výstupných hodnôt $(x_1, t_1), (x_2, t_2), \ldots, (x_N, t_N)$. Chyba modelu pre jednu dvojicu je

$$E_i(t_i; f(x_i; k, q)) = (t_i - f(x_i; k, q))^2 \tag{1}$$

Pre všetky dvojice budeme používať kvadratickú chybovú funkciu

$$E = \frac{1}{N} \sum_{i=1}^{N} E_i(t_i; f(x_i; k, q)) \tag{2}$$

Najlepší modelom je taký, pre ktorý je chyba (2) najmenšia.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA

Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje

www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Pripomenutie lineárnej regresie

Úpravami dostaneme výrazy pre k a q

$$k = \frac{\overline{xt} - \overline{x}\,\overline{t}}{\overline{x^2} - \overline{x}^2}, \quad q = \overline{t} - k\overline{x}, \tag{3}$$

$\overline{x}, \overline{t}, \overline{xt}, \overline{x^2}, \overline{x}^2$ sú priemery.

Výpočtom druhej derivácie sa presvedčíme, že ide o minimum.

Ak položíme $w_0 = q, w_1 = k, \mathrm{x}_n = [1, x_n]^T$, model vyjadríme

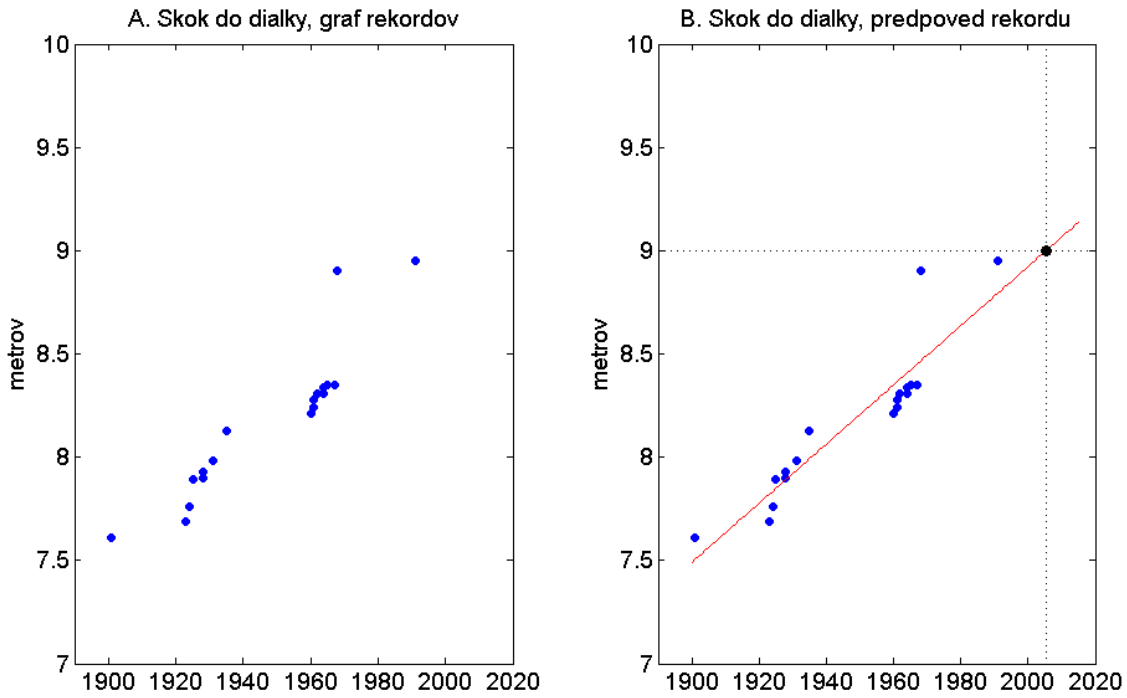
$$f(x_n; \mathrm{w}) = \mathrm{w}^T \mathrm{x}_n \tag{4}$$

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA

Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje

www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

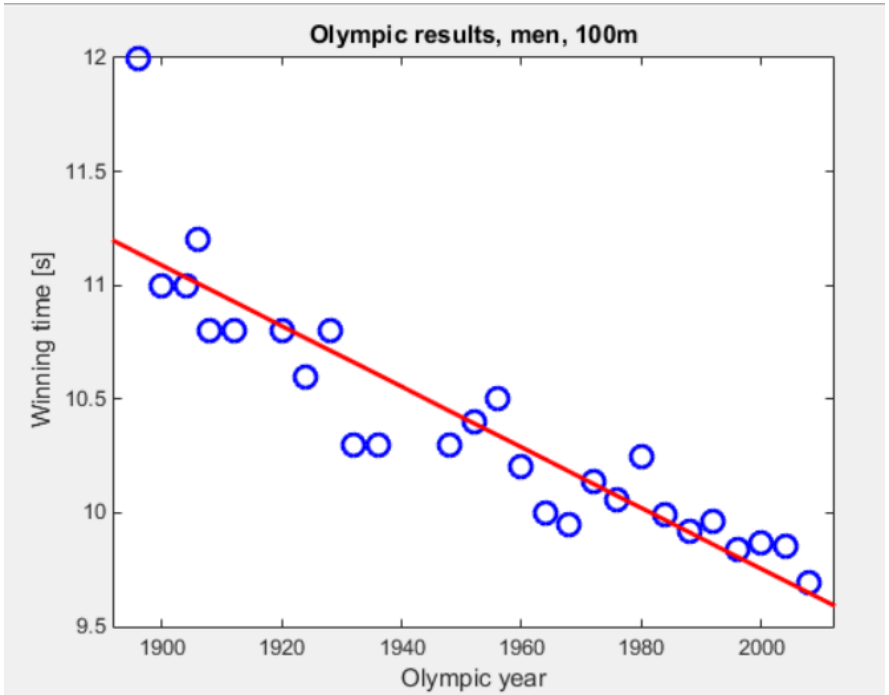
Rekordy v skoku do diaľky mužov



Obr.: Rekord 9 metrov bol očakávaný v roku 2008

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje

Iný príklad - beh mužov na 100m, olympijské časy



Obr.: Graf časov behu mužov, $f(x; w_0, w_1) = 36.416 - 0.013x$

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje

Chyby ako šum

- Pri vytváraní lineárneho modelu sme predpokladali lineárny vzťah medzi nezávislými premennými a závislou premennou.
- Tento model zachytáva všeobecný trend v dátach, ale ignoruje to, že v niektorých bodoch sa dopúšťa veľkej chyby.
- Z hľadiska modelu je ťažké tieto chyby obhájiť. Vieme, že chyby v dátach budú, ale mali by sme ich pri tvorbe modelu zohľadniť.
- Je možné s modelom experimentovať, modifikovať jeho parametre, ale je tu ešte iný prístup.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Chyby ako šum

Niektoré vlastnosti chýb (majme na mysli náš príklad rekordov):

- Chyby sú rôzne v každom roku (bode x). Niekedy sú pozitívne, niekedy negatívne a majú rôzne hodnoty.
- Nie je žiadny vzťah medzi veľkosťou chýb v rokoch (bodoch x). Chyby nie sú funkciou roka (bodu x).
- Nájdený model sa dá dosiahnuť aj iným datasetom.

Budeme považovať náš modelovací problém za **generatívny**, to znamená, že môžeme vytvárať model, ktorý bude použitý na vytvorenie (generovanie) podobného datasetu.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA

Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje

www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Chyby ako šum

- Použijeme náhodnú premennú ϵ_n v každom bode a náš model teraz bude vyzerať:

$$t_n = \mathbf{w}^T \mathbf{x}_n + \epsilon_n \tag{5}$$

- Na doplnenie definície modelu potrebujeme rozhodnúť o pravdepodobnostnom rozdelení premennej ϵ_n.
- ϵ_n je spojitá náhodná premenná.
  Rozdiel medzi modelom a konkrétnymi dĺžkami skokov je spojitá veličina (model je spojitý, medzi skokmi ...).

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Chyby ako šum

- Keď pracujeme so spojitými náhodnými premennými, potrebujeme spojitú analógiu pravdepodobnostného rozdelenia - funkcia hustoty pravdepodobnostného rozdelenia (probability density function – pdf) $p(x)$ pre premennú x.
- Vypočítať pravdepodobnosť, že X leží v intervale $x_1 \leq X \leq x_2$, znamená

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} p(x)dx$$

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Chyby ako šum

Ak náhodná premenná nadobúda hodnoty len v intervale $x_1 \leq X \leq x_2$, pravdepodobnosť je rovná 1, teda $\int_{x_1}^{x_2} p(x)dx = 1$

Normálne rozdelenie pre ϵ, t.j.

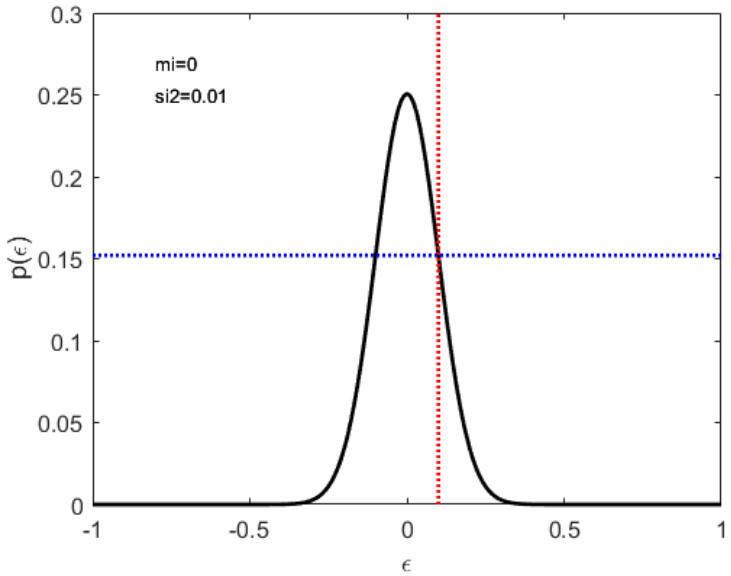
$$p(\epsilon|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} exp(-\frac{1}{2\sigma^2}(\epsilon - \mu)^2) \tag{6}$$

Použijeme označenie pre ϵ, t. j. $\mathcal{N}(\mu, \sigma^2) = p(\epsilon|\mu, \sigma^2)$.
Pre príklad s nastaveniami $\mu = 0, \sigma^2 = 0.05$ dostávame nasledujúci obrázok

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Vykreslenie pravdepodobnosti chyby

$$p(\epsilon) = \mathcal{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2*0.05*\pi}} exp(-\frac{1}{2*0.05}\epsilon^2) = \frac{1}{\sqrt{0.314}} exp(-10 * \epsilon^2)$$



Generovaním možnej chyby dostaneme realistickejší dataset.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA

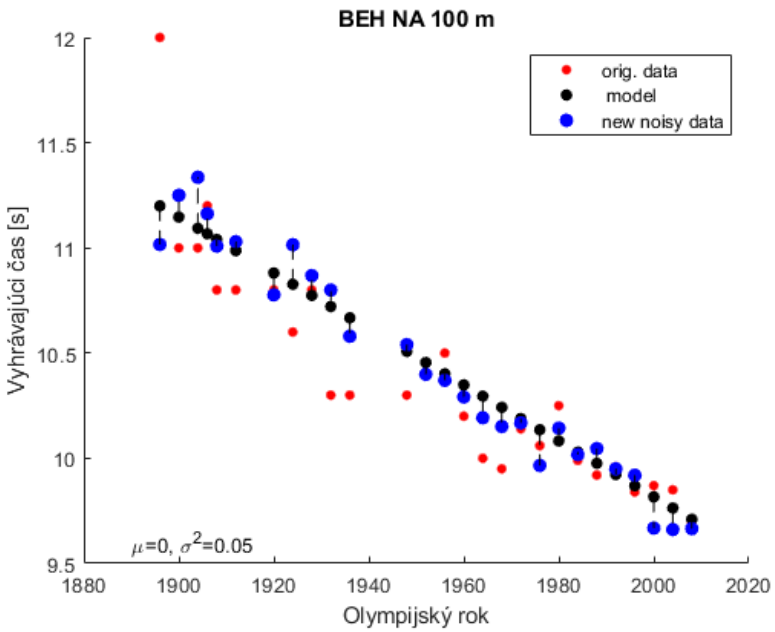
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje

www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Generované dáta so šumom - čas behu mužov

noise_var = 0.01;

noisy_t = model_t + randn(size(model_t)).*sqrt(noise_var);



EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje

Chyby ako šum

Náš model teraz pozostáva z dvoch komponentov v každom bode:

- deterministický komponent ($w^T x_i$), ktorý sa nazýva **trend** alebo **drift**,
- náhodný komponent (ϵ_i), ktorý sa nazýva **šum**.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Chyby ako šum

- Tiež si uvedomíme, že nemáme jedinu náhodnú premennú, ale máme pre každé x_i jednu.
- Môžeme predpokladať, že tieto hodnoty sú nezávislé. A teda pre pravdepodobnosť platí

$$p(\epsilon_1, \ldots, \epsilon_N) = \Pi_{i=1}^{N} p(\epsilon_i) \tag{7}$$

- Pripomíname, že rozdiel medzi modelom a dĺžkami skokov je predpokladaná spojitá veličina. Preto ϵ_i je je spojitá náhodná premenná.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Dôveryhodnosť (hodnovernosť)

- Náš model je teraz v nasledujúcom tvare

$$t_i = f(\mathrm{x}_i; \mathrm{w}) + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{8}$$

- Hľadáme optimálnu hodnotu pre w, ktorú označíme $\hat{\mathrm{w}}$.
- V regresnom modeli strata bola meraná pomocou rozdielu pozorovanej hodnoty a hodnoty, ktorú predikoval model.
- Efekt pridania náhodnej premennej do modelu je, že výstup modelu, t, je teraz tiež náhodná premenná. Inak povedané neexistuje len jedna hodnota t_i pre partikulárne x_i. Teda nie je možné použiť len stratu pre optimalizáciu w a σ^2.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Dôveryhodnosť (hodnovernosť)

Poznámka: Pridanie konštanty ($\mathrm{w}^T\mathrm{x}_i$) do Gaussovej náhodnej premennej je ekvivalentné inej Gaussovej náhodnej premennej s priemerom posunutým o uvedenú konštantu:

$$y = a + z, \quad p(z) = \mathcal{N}(m, s), \quad p(y) = \mathcal{N}(m + a, s)$$

Preto náhodná premenná t_i má nasledujúcu funkciu rozdelenia

$$p(t_i|\mathrm{x}_i, \mathrm{w}, \sigma^2) = \mathcal{N}(\mathrm{w}^T\mathrm{x}_i, \sigma^2) \tag{9}$$

t_i závisí od hodnôt x_i a w (určujú priemer) a od σ^2 (varianciu).

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

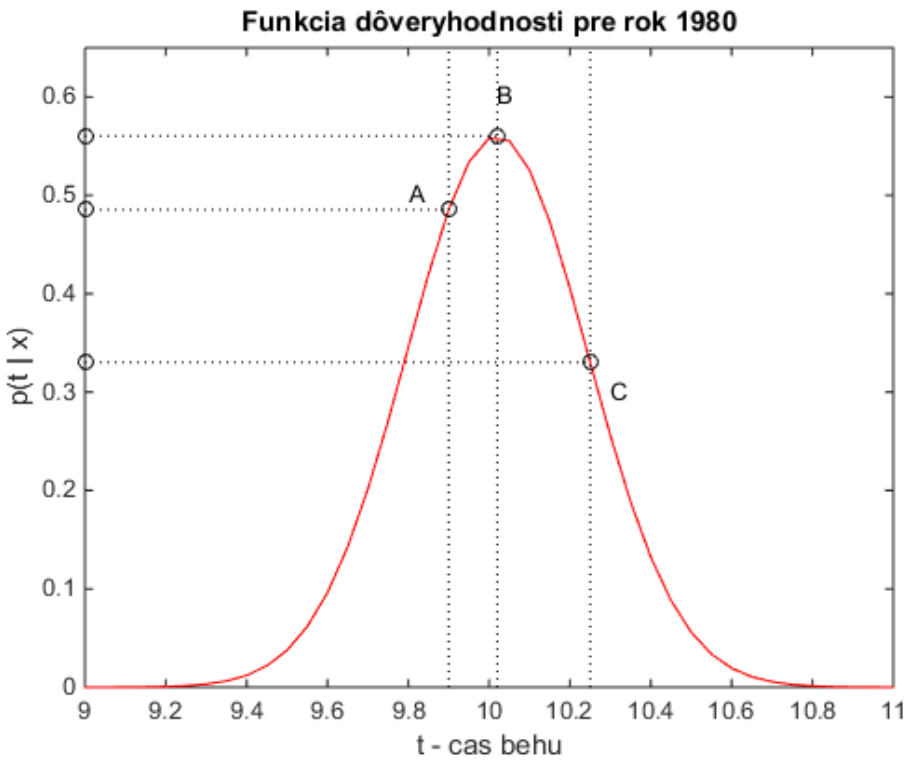
Dôveryhodnosť (hodnovernosť)

Aby sme videli, ako môžeme toto použiť na hľadanie optimálnej hodnoty w a σ^2 v úlohe Beh na 100 m, vyberme jeden rok z datasetu behov – 1980. Pre tento rok dostaneme

$$p(t_n|\mathrm{x}_n = [1, 1980]^T, \mathrm{w} = [36.416, -0.0133]^T, \sigma^2 = 0.05), \quad (10)$$

pri Gaussovom rozdelení $\mu = 36.416 - 0.0133 * 1980 = 10.02$ a $\sigma^2 = 0.05$

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk



Obr.: Graf pre 1980, x-ová os t, y-ová os $p(t|x)$, $\mu = 10.02$ a $\sigma^2 = 0.05$

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

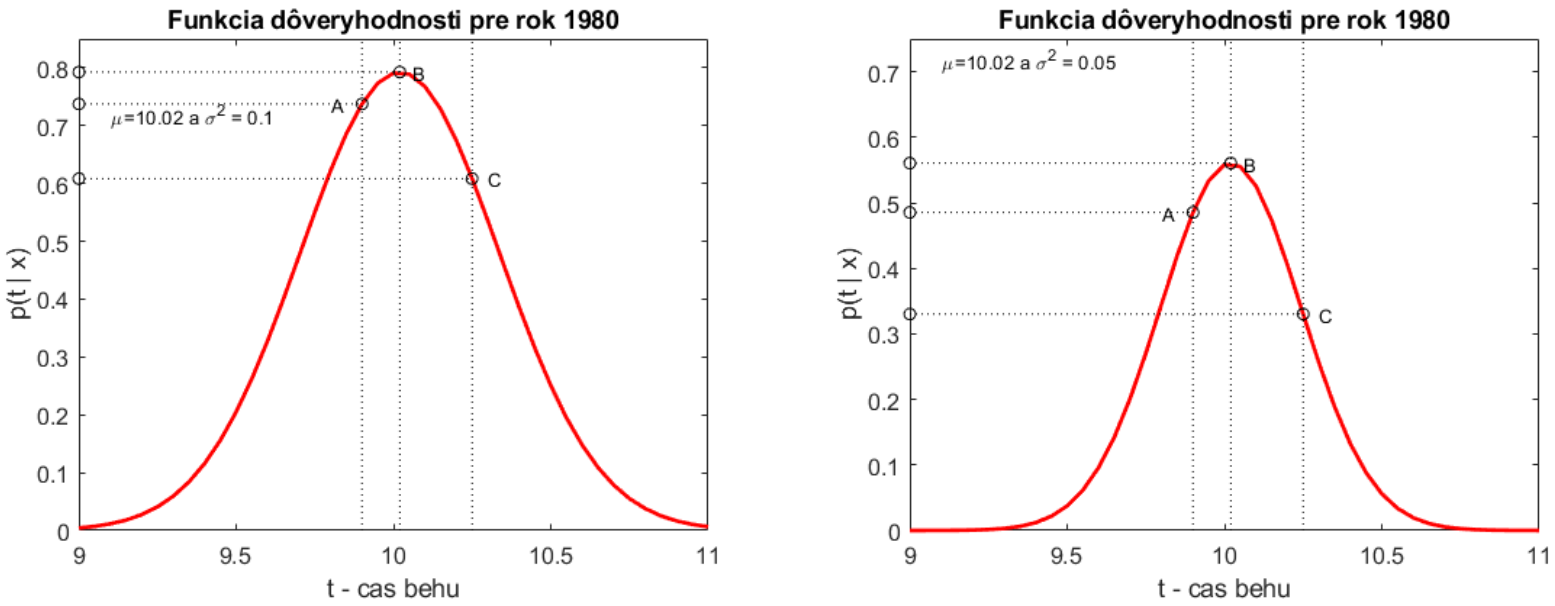
Dôveryhodnosť (hodnovernosť)

Pripomeňme, že pre spojitú premennú t, $p(t)$ nemôže byť interpretovaná ako pravdepodobnosť. Prečo?
Výška krivky v partikulárnej hodnote t môže byť interpretovaná nasledovne:

Hodnovernosť (dôveryhodnosť) možnej hodnoty t pre daný rok $x = 1980$.
Najhodnovernejšia hodnota v roku 1980 by bola 10.02 sekúnd (pre Gaussovo rozdelenie najhodnovernejší - najvyšší bod zodpovedá priemeru). Na obrázku máme 3 príklady časov - A, B, C. Z nich B je najhodnovernejší, a C najmenej hodnoverný.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Dôveryhodnosť (hodnovernosť)



Obr.: Graf pre 1980, x-ová os t, y-ová os $p(t|x)$, $\mu = 10.02$

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje

Dôveryhodnosť (hodnovernosť)

- Skutočný nameraný čas v roku 1980 je C (10.25 sekúnd). $p(t_i|\mathrm{x}_i, \mathrm{w}, \sigma^2)$ vyhodnotené v bode $t_i = 10.25$ je dôležitá veličina, ktorá je známa ako **dôveryhodnosť (hodnovernosť)** i-tého dátového bodu.
- Nemôžeme zmeniť $t_n = 10.25$ (máme to v dátach), ale môžeme zmeniť w a σ^2 a skúšať, posúvať a hľadať čo najväčšiu dôveryhodnosť.
- Idea hľadania parametrov, ktoré maximalizujú dôveryhodnosť týmto spôsobom, je kľúčový koncept v strojovom učení.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Hodnovernosť datasetu

Musíme uvažovať o celom datasete (nestačí jeden bod). Označme

$$\mathrm{X} = [\mathrm{x}_1^T \mathrm{x}_2^T \dots \mathrm{x}_N^T]^T, \mathrm{t} = [t_1 \, t_2 \dots t_N]^T \tag{11}$$

$$\mathrm{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix}$$

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Hodnovernosť datasetu

- V skutočnosti sa nezaujímame o dôveryhodnosť jedného dátového bodu ale o dôveryhodnosť celého datasetu.
- Keď máme N bodov zaujíma nás podmienená pravdepodobnosť

$$p(t_1, \ldots, t_N | \mathrm{x}_1, \ldots, \mathrm{x}_N, \mathrm{w}, \sigma^2) = p(\mathrm{t} | \mathrm{X}, \mathrm{w}, \sigma^2).$$

- **Vyhodnotenie tejto pravdepodobnosti (hustoty) v pozorovaných dátových bodoch dáva jednu hodnotu dôveryhodnosti pre celý dataset**, ktorú môžeme optimalizovať zmenou parametrov w a σ^2.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Hodnovernosť datasetu

- Predpoklad, že šum v každom dátovom bode je nezávislý, $(p(\epsilon_1, \ldots, \epsilon_N) = \Pi_n p(\epsilon_n))$, nám umožňuje použiť zaujímavé odvodenia.
- Táto spojená (joint) podmienená pravdepodobnosť (hustota) môže byť rozdelená do N separatných výrazov, jeden pre každý dátový objekt:

$$L = p(\mathrm{t}|\mathrm{X}, \mathrm{w}, \sigma^2) = \Pi_{i=1}^{N} p(t_i|\mathrm{x}_i, \mathrm{w}, \sigma^2) = \Pi_{i=1}^{N} \mathcal{N}(\mathrm{w}^T \mathrm{x}_i, \sigma^2) \quad (12)$$

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Hodnovernosť datasetu

- Pripomíname, že t_i hodnoty sú navzájom úplne nezávisle v našich úvahách.
- Ale sú v podstate v čase klesajúce, a teda ponúkajú jasnú štatistickú závislosť. Bez toho by sme nemohli vytvárať model.
- V skutočnosti sú ale **podmienene nezávislé** – k daným hodnotám w (deterministická časť modelu) t_i sú nezávislé; inak nie sú celkom nezávislé.
- Predstavme si, že nám chýbajú dáta z jedného roku, napríklad 1960. Ostatné roky sú v poriadku. X, t neobsahujú údaje o roku 1960.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Hodnovernosť datasetu

Z daných dát by sme chceli vedieť niečo o roku 1960. Teda nás zaujíma $p(t_{1960}|x_{1960}, X, t)$

Z definície podmienenej pravdepodobnosti dostaneme

$$p(t_{1960}|x_{1960}, X, t) = \frac{p(t_{1960}, t|x_{1960}, X)}{p(t|X)}$$

Za prepokladu, že t sú nezávislé, t_{1960} závisí len od x_{1960}:

$$p(t_{1960}|x_{1960}, X, t) = \frac{p(t_{1960}|x_{1960})\Pi_i p(t_i|x_i)}{\Pi_i p(t_i|x_i)} = p(t_{1960}|x_{1960})$$

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA

Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje

www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Hodnovernosť datasetu

- Pre náš model toto nie je použiteľné, t_{1960} musí byť v nejakom zmysle závisle od ostatných dát. Táto závislosť je obsiahnutá v parametroch w.
- Ak poznáme w, tak všetko, čo máme, sú chyby medzi pozorovanými dátami a $\mathrm{w}^T \mathrm{x}_i$.
- Predpokladáme, že tieto chyby sú celkom nezávislé. Vzhľadom na w pozorovania sú nezávislé (vo w je obsiahnutá závislosť).
- Bez modelu pozorovania nie sú nezávislé, ale keď berieme do úvahy model (w), pretože je tu aj chyba, pozorovania berieme ako nezávislé.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Maximálna dôveryhodnosť datasetu

Hľadáme parametre modelu, ktoré urobia náš model najdôveryhodnejší. Z analytických dôvodov budeme maximalizovať prirodzený logaritmus dôveryhodnosti L (12) (logaritmus je rastúca funkcia a teda je to možné urobiť).

$$\log L = \Sigma_{i=1}^{N} \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} exp\{-\frac{1}{2\sigma^2}(t_i - f(\mathrm{x}_i; \mathrm{w}))^2\} \right)$$

$$= \sum_{i=1}^{N} \left(-\frac{1}{2}\log(2\pi) - \log\sigma - \frac{1}{2\sigma^2}(t_i - f(\mathrm{x}_i; \mathrm{w}))^2 \right)$$

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA

Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje

www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Maximálna dôveryhodnosť

$$= -\frac{N}{2}\log 2\pi - N\log\sigma - \frac{1}{2\sigma^2}\Sigma_{i=1}^{N}(t_i - f(\mathrm{x}_i;\mathrm{w}))^2$$

Substitúciou $f(\mathrm{x}_i;\mathrm{w}) = \mathrm{w}^T\mathrm{x}_i$ dostaneme

$$\log L = -\frac{N}{2}\log 2\pi - N\log\sigma - \frac{1}{2\sigma^2}\Sigma_{i=1}^{N}(t_i - \mathrm{w}^T\mathrm{x}_i)^2 \qquad (13)$$

Parciálne derivácie podľa w

$$\frac{\partial \log L}{\partial \mathrm{w}} = \frac{1}{\sigma^2}\Sigma_{i=1}^{N}\mathrm{x}_i(t_i - \mathrm{w}^T\mathrm{x}_i) = \frac{1}{\sigma^2}\Sigma_{i=1}^{N}\left(\mathrm{x}_i t_i - \mathrm{x}_i\mathrm{x}_i^T\mathrm{w}\right) = 0$$

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Maximálna dôveryhodnosť

V maticovom tvare dostávame

$$\frac{\partial \log L}{\partial \mathrm{w}} = \frac{1}{\sigma^2}(\mathrm{X}^T\mathrm{t} - \mathrm{X}^T\mathrm{X}\mathrm{w}) = 0$$

Riešením tejto rovnice je

$$\mathrm{w} = (\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T\mathrm{t} \tag{14}$$

Teda maximálna dôveryhodnosť (z hľadiska w) je pre

$$\hat{\mathrm{w}} = (\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T\mathrm{t} \tag{15}$$

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Maximálna dôveryhodnosť

Pre σ^2 dostávame

$$\frac{\partial \log L}{\partial \sigma} = -\frac{N}{\sigma} + \frac{1}{\sigma^3}\Sigma_{i=1}^{N}(t_i - \mathrm{x}_i^T \hat{\mathrm{w}})^2 = 0 \tag{16}$$

$$\hat{\sigma^2} = \frac{1}{N}\Sigma_{i=1}^{N}(t_i - \mathrm{x}_i^T \hat{\mathrm{w}})^2 \tag{17}$$

Tento výraz je perfektný - variancia je jednoducho priemer kvadratickej chyby. Platí

$$\Sigma_{i=1}^{N}(t_i - \mathrm{x}_i^T \hat{\mathrm{w}})^2 = (\mathrm{t} - \mathrm{X}\hat{\mathrm{w}})^T(\mathrm{t} - \mathrm{X}\hat{\mathrm{w}})$$

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA

Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje

www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Maximálna dôveryhodnosť

V maticovom označení dostávame

$$\hat{\sigma^2} = \frac{1}{N}(\mathrm{t} - \mathrm{X}\hat{\mathrm{w}})^T(\mathrm{t} - \mathrm{X}\hat{\mathrm{w}}) \tag{18}$$

$$= \frac{1}{N}(\mathrm{t}^T\mathrm{t} - \mathrm{t}^T\mathrm{X}\hat{\mathrm{w}} - \hat{\mathrm{w}}^T\mathrm{X}^T\mathrm{t} + \hat{\mathrm{w}}^T\mathrm{X}^T\mathrm{X}\hat{\mathrm{w}})$$

Toto zjednodušíme substitúciou $\hat{\mathrm{w}}^T = \mathrm{t}^T\mathrm{X}(\mathrm{X}^T\mathrm{X})^{-1}$

$$= \frac{1}{N}(\mathrm{t}^T\mathrm{t} - \mathrm{t}^T\mathrm{X}\hat{\mathrm{w}} - \hat{\mathrm{w}}^T\mathrm{X}^T\mathrm{t} + \mathrm{t}^T\mathrm{X}(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T\mathrm{X}\hat{\mathrm{w}})$$

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Maximálna dôveryhodnosť

$$= \frac{1}{N}(t^T t - t^T X\hat{w} - \hat{w}^T X^T t + t^T X\hat{w})$$

$$= \frac{1}{N}(t^T t - t^T X\hat{w} - \hat{w}^T X^T t + (\hat{w}^T X^T t)^T)$$

Matica $\hat{w}^T X^T t$ je typu 1x1, preto jej transponovaná matica je rovnaká.
$\hat{w}^T$ je typu 1x2, X^T je typu 2xN, súčin je typu 1xN. t je typu Nx1. Súčin je typu 1x1.
Dostaneme

$$\hat{\sigma^2} = \frac{1}{N}(t^T t - t^T X\hat{w}) \tag{19}$$

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Maximálna dôveryhodnosť

Pre rekordy v behu dostávame

$$\hat{w} = [36.4165, -0.0133]^T, \hat{\sigma^2} = 0.0503$$

$\hat{w}$ je to to isté ako pri pri lineánej regresii, $\hat{\sigma^2}$ predstavuje zmenenú hodnotu Gaussovho rozdelenia – šum, ktorý sme predpokladali v dátach.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Charakteristika riešenia maximálnej dôveryhodnosti

Aby sme dokázali, že je to maximum, mali by sme vypočítať druhé derivácie. Vo vektorovom tvare dostaneme

$$\frac{\partial^2 \log L}{\partial \mathrm{w} \partial \mathrm{w}^T} = -\frac{1}{\sigma^2} \mathrm{X}^T \mathrm{X} \tag{20}$$

Ľavá strana predstavuje maticu všetkých druhých derivácií funkcie $\log L$, ktorú nazývame **Hessian**.
Teda riešenie bude maximom, ak táto matica bude **negatívne definitná**.
Matica H je **negatívne definitná**, ak platí $\mathrm{x}^T \mathrm{H} \mathrm{x} < 0$ pre všetky reálne vektory x.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA

Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje

www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Charakteristika riešenia maximálnej dôveryhodnosti

To znamená, že potrebujeme dokázať, že

$$-\frac{1}{\sigma^2}\mathrm{z}^T\mathrm{X}^T\mathrm{X}\mathrm{z} < 0, \quad \text{po úprave} \quad \mathrm{z}^T\mathrm{X}^T\mathrm{X}\mathrm{z} > 0$$

pre ľubovoľný reálny vektor $\mathrm{z}^T = [z_1, z_2]$.
Ak

$$\mathrm{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix}$$

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA

Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje

www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Charakteristika riešenia maximálnej dôveryhodnosti

$$\mathrm{X}^T\mathrm{X} = \begin{pmatrix} N & \sum_{i=1}^N x_i \\ \sum_{i=1}^N x_i & \sum_{i=1}^N x_i^2 \end{pmatrix}$$

$$\mathrm{z}^T\mathrm{X}^T\mathrm{X}\mathrm{z} = [Nz_1 + z_2\sum_{i=1}^N x_i, z_1\sum_{i=1}^N x_i + z_2\sum_{i=1}^N x_i^2]\mathrm{z} =$$

$$= [Nz_1^2 + 2z_1z_2\sum_{i=1}^N x_i + z_2^2\sum_{i=1}^N x_i^2]$$

Potrebujeme dokázať, že $Nz_1^2 + z_2^2\sum_{i=1}^N x_i^2 > 2z_1z_2\sum_{i=1}^N x_i$

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA

Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje

www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Charakteristika riešenia maximálnej dôveryhodnosti

Položíme $y_i = z_2 x_i$, dostaneme

$$N z_1^2 + \sum_{i=1}^{N} y_i^2 > 2 z_1 \sum_{i=1}^{N} y_i$$

$$\sum_{i=1}^{N} (z_1^2 + y_i^2 - 2 z_1 y_i) = \sum_{i=1}^{N} (z_1 - y_i)^2 = \sum_{i=1}^{N} (z_1 - z_2 x_i)^2 > 0$$

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Charakteristika riešenia maximálnej dôveryhodnosti

Aby sme ukázali, že náš výraz pre $\hat{\sigma^2}$ zodpovedá maximálnej dôveryhodnosti, potrebujeme vypočítať druhú deriváciu vzhľadom na σ

$$\frac{\partial^2 log\ L}{\partial \sigma^2} = \frac{N}{\sigma^2} - \frac{3}{\sigma^4}(\mathrm{t} - \mathrm{X}\hat{\mathrm{w}})^T(\mathrm{t} - \mathrm{X}\hat{\mathrm{w}})$$

$$\frac{\partial^2 log\ L}{\partial \sigma^2} = \frac{N}{\hat{\sigma^2}} - \frac{3}{\hat{\sigma^2}^2} N\hat{\sigma^2} = -\frac{2N}{\hat{\sigma^2}} < 0$$

čo zodpovedá maximu.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Charakteristika riešenia maximálnej dôveryhodnosti

Aby sme dokázali, že je to maximum, mali by sme vypočítať druhú deriváciu vo vypočítanom bode pre $\hat{\sigma^2}$. Hodnota $\log L$ v maxime je

$$\log L = -\frac{N}{2}\log 2\pi - N\log\hat{\sigma^2} - \frac{1}{2\hat{\sigma^2}}N\hat{\sigma^2} \tag{21}$$

$$= -\frac{N}{2}(1+\log 2\pi) - \frac{N}{2}\log\hat{\sigma^2}$$

L bude rásť, keď $\hat{\sigma^2}$ bude klesať. Toto ale ovplyní w, teda deterministický model (výber komplikovanejšieho modelu).

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Čo sme ukázali?

- Uvažovali sme, že v dátach je šum (chyby merania),
- Za predpokladu, že chyby majú Gaussovo rozdelenie pravdepodobnosti, sme ukázali, že vieme vypočítať dôveryhodnsť, ktorá popisuje ako dôveryhodné sú naše dáta ako funkciu parametrov nášho modelu,

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Vplyv šumu na odhad parametrov

Chceme ukázať benefity modelovania šumu:

- schopnosť kvantifikovať neurčitosť v parametroch, a
- schopnosť vyjadriť neurčitosti v našich predikciách.

Máme odvodené

$$\hat{w} = (X^T X)^{-1} X^T t \tag{22}$$

$$\hat{\sigma^2} = \frac{1}{N}(t^T t - t^T X \hat{w}) \tag{23}$$

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA

Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje

www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Vplyv šumu na odhad parametrov

Uvažujme

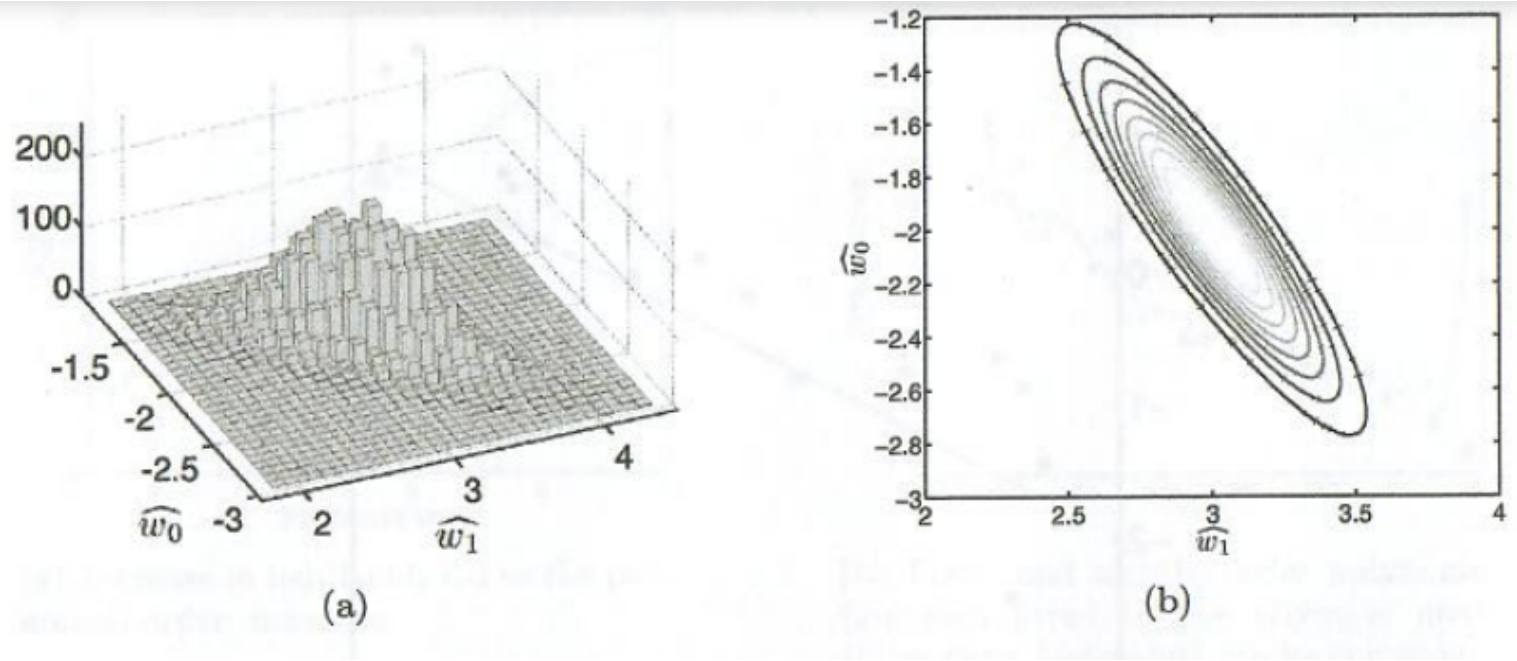
$$t_n = w_0 + w_1 x_n + \epsilon_n, \epsilon_n \sim \mathcal{N}(0, \sigma^2) \tag{24}$$

Predpokladajme, že $w_0 = -2, w_1 = 3$, hustota variancie $\sigma^2 = 0.5^2$. Vieme generovať množstvo syntetických datasetov. Pre každý dataset vieme vypočítať $\hat{\mathrm{w}}$.

Zaujíma nás variabilita w_0, w_1.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

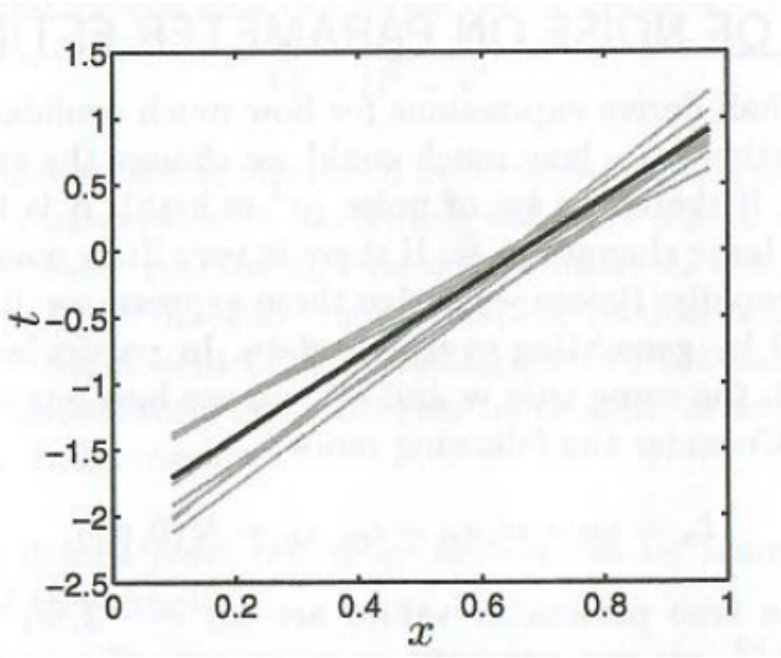
Vplyv šumu na odhad parametrov



Obr.: Variabilita **w** pre 10000 datasetov vygenerovaných modelom (24)

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Vplyv šumu na odhad parametrov



Obr.: Pôvodný model čierny, ďalšie vygenerované modely sivé.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Úlohy:

1. Predpokladajme, že dataset N binárnych hodnôt $x_1, \ldots, x_N$ bol generovaný Bernoulliho rozdelením. Vypočítajte odhad maximálnej dôveryhodnosti pre Bernoulliho parameter. Bernoulliho rozdelenie: $P(X = x) = q^x(1-q)^{1-x}$
2. Analyzujte skoky do diaľky u žien použitím šumu. Vygenerujte podobnú tréningovú vzorku.
3. Nasledujúci výraz vyjadruje váženú priemernú stratovú funkciu lineárneho modelu, v ktorom α_i sú konštanty

$$\mathcal{L} = \frac{1}{N}\Sigma_{i-1}^{N}\alpha_i(t_i - \mathrm{w}^T\mathrm{x}_i)^2$$

Odvoďte optimálne hodnoty parametrov $\hat{\mathrm{w}}$ pri klasickej chybe.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA

Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje

www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk