# Statistika

# (D360P03Z, D360P03U)

Karel Zvára

20. prosince 2004

# obecné předpoklady pro regresní model

- **tvar závislosti**: známe jak vysvětlovaná veličina závisí na vysvětlujících
- **homoskedasticita**: pro všechny kombinace hodnot vysvětlujících veličin je rozptyl vysvětlované veličiny konstantní
- **nezávislost**: náhodné složky vysvětlovaných veličin jsou nezávislé
- **normalita**: náhodná složka má normální rozdělení
- předpoklady lze ověřovat (regresní diagnostika)
- někdy pomohou transformace

# použití reziduí

- pomocí regrese hledáme model pro závislost nebo predikci (střední hodnoty) příštích pozorování
- celkovou schopnost vysvětlit závisle proměnnou hodnotíme pomocí **koeficientu determinace**

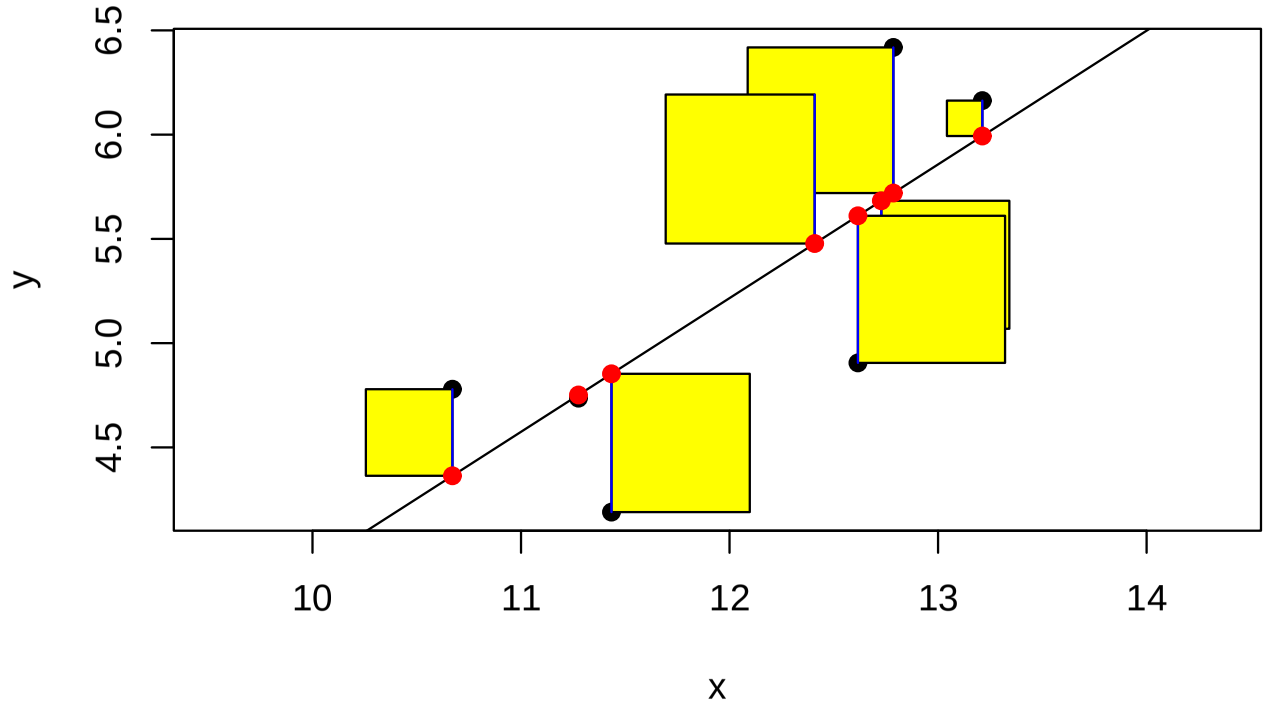
$$R^2 = 1 - \frac{S_e}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

- v čitateli posledního výrazu **rezidua**

$$u_i = Y_i - \hat{Y}_i$$

  (rozdíl **naměřená - vyrovnaná** hodnota vysvětlované proměnné)

- rezidua lze použít k hodnocení (diagnostice) regrese




$Y_i$, $\hat{Y}_i$

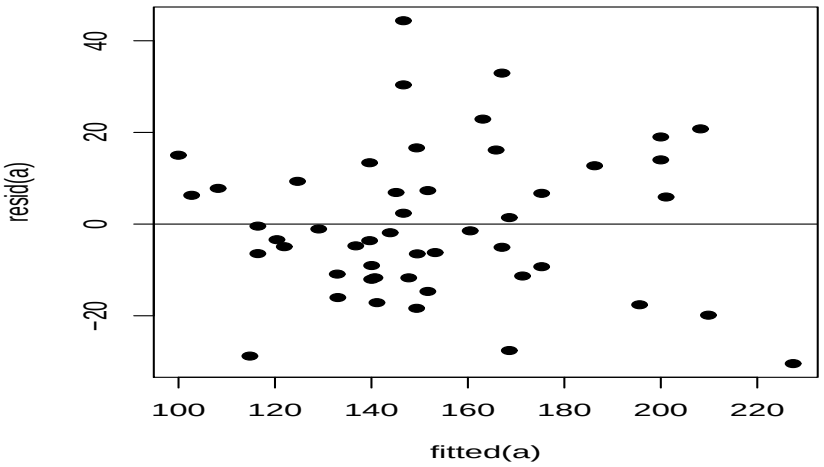
# diagnostika pomocí reziduí

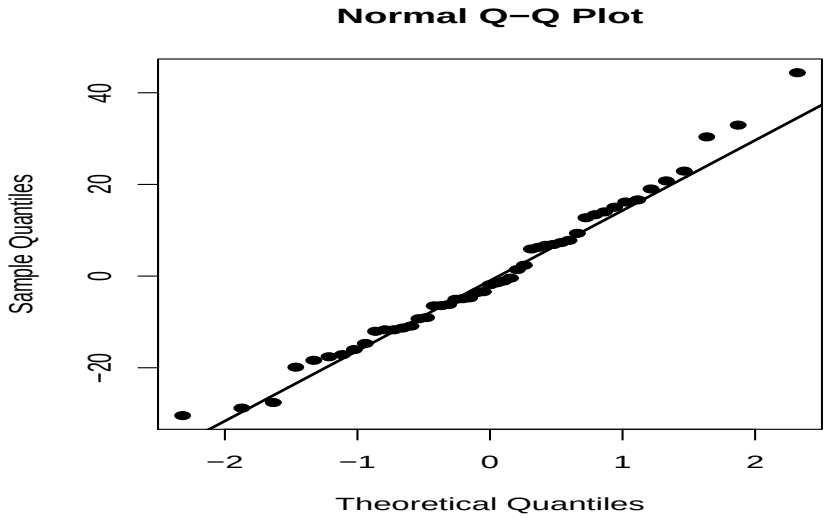
- histogram reziduí nebo normální diagram (k ověření normálního rozdělení)
- grafické znázornění bodů $[\hat{Y}_i, u_i]$ nebo $[x_i, u_i]$ (k ověření konstantního rozptylu či tvaru závislosti)

**ukázky diagnostiky**

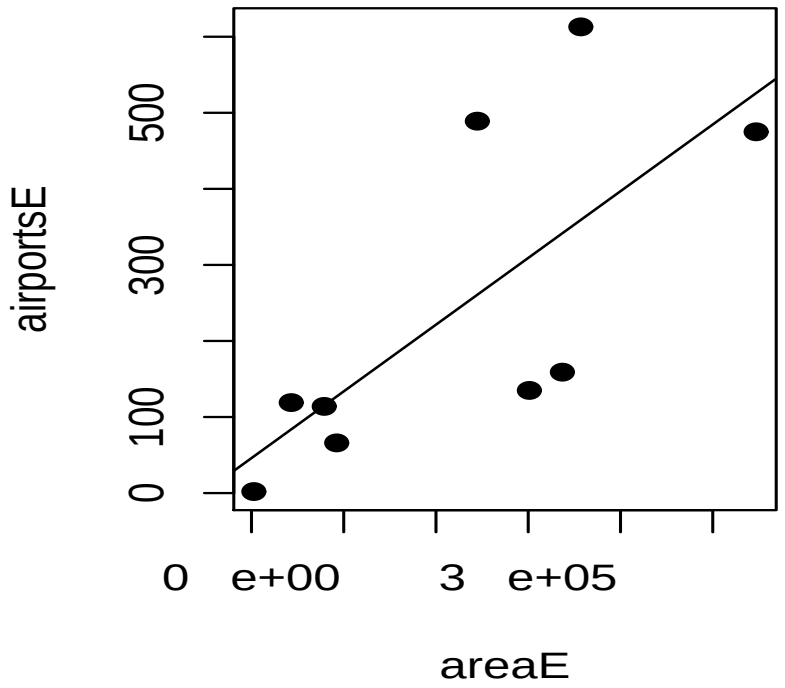
vlevo: rezidua spíše kladná než záporná, možná jsme měli raději vysvětlovat odmocninu z mortality

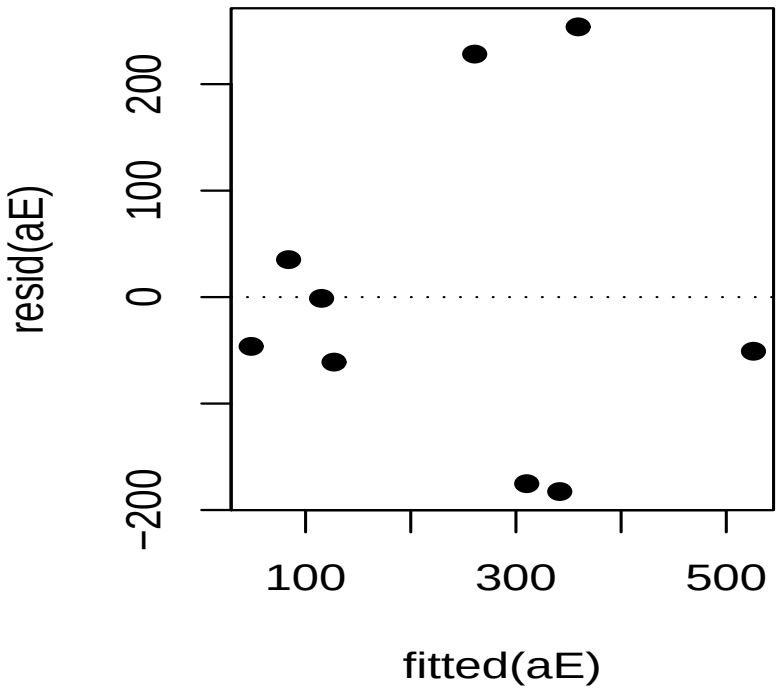
vpravo: normální diagram, ukazuje, že s předpokladem o normálním rozdělením není problém (body těsně kolem přímky)



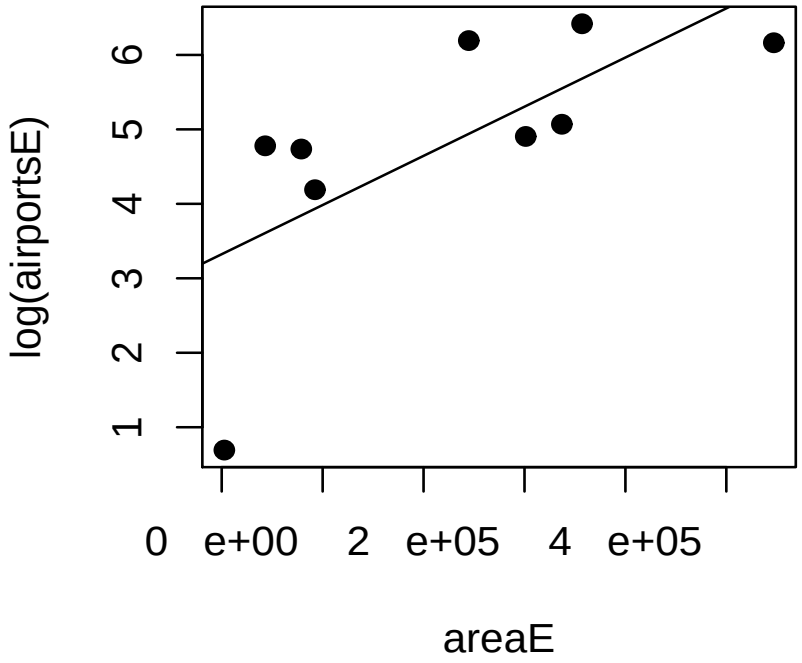



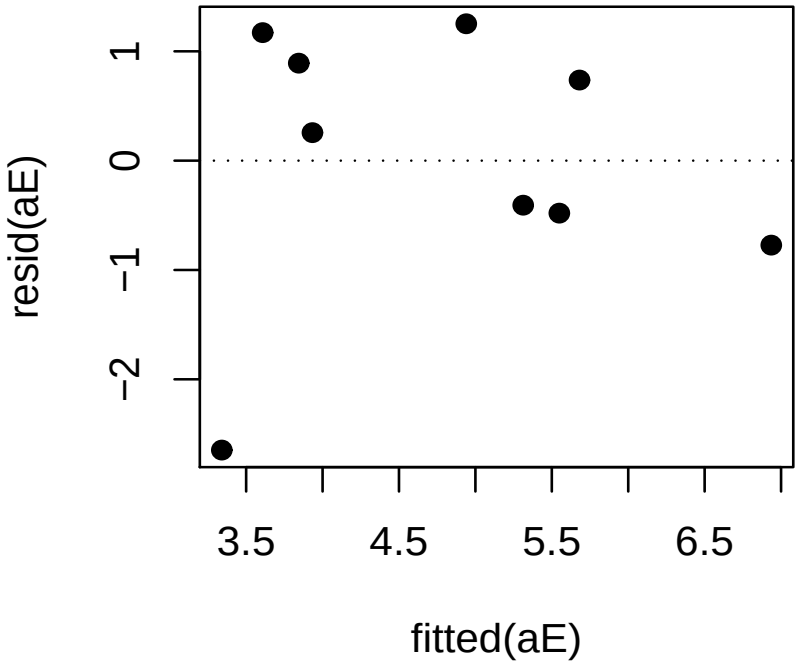
nekonstantní rozptyl (trychtýřovité rozšiřování mraku reziduí)



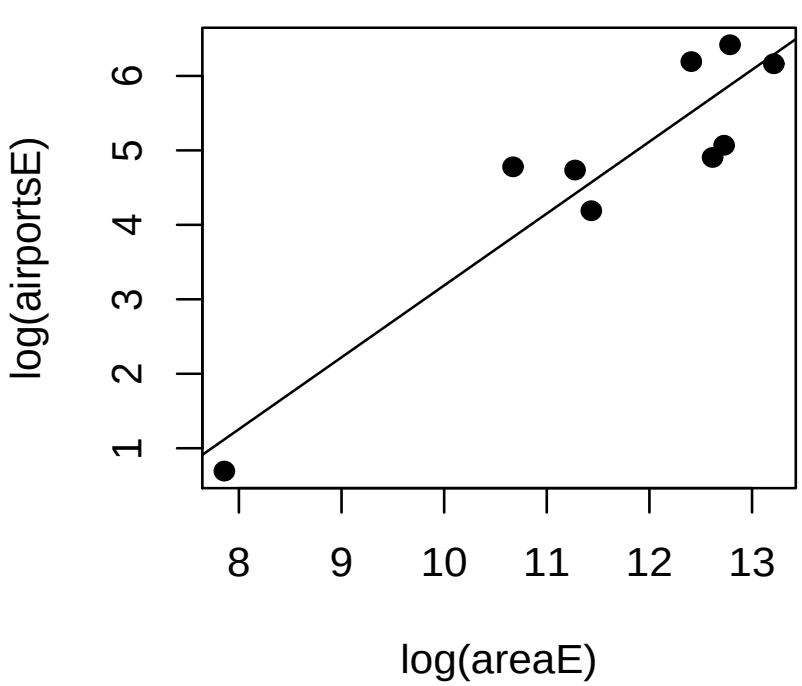



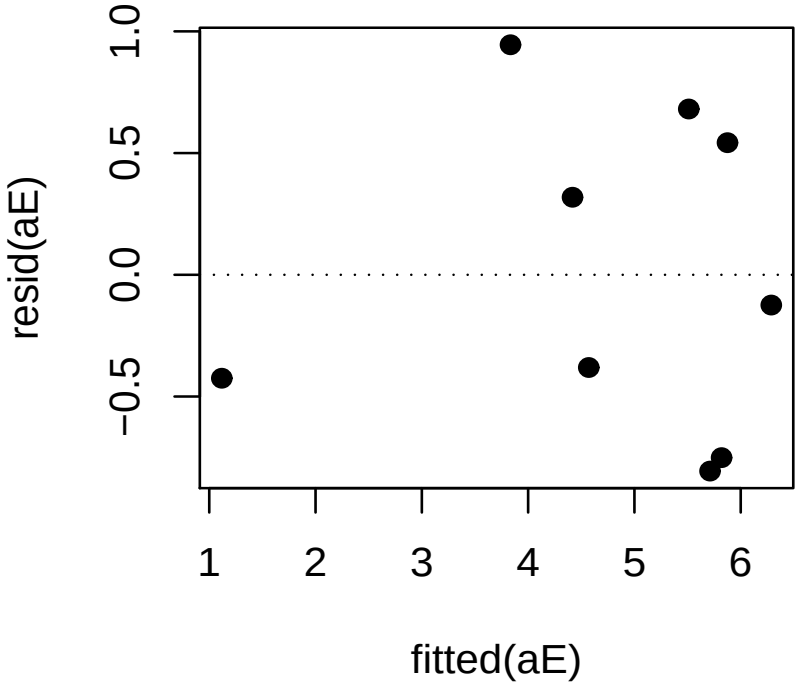
odlehlé pozorování (první pozorování daleko od vodorovné osy)



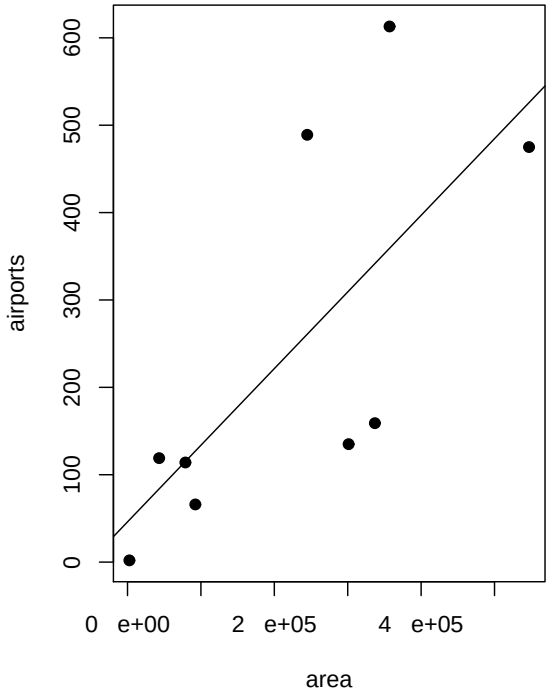



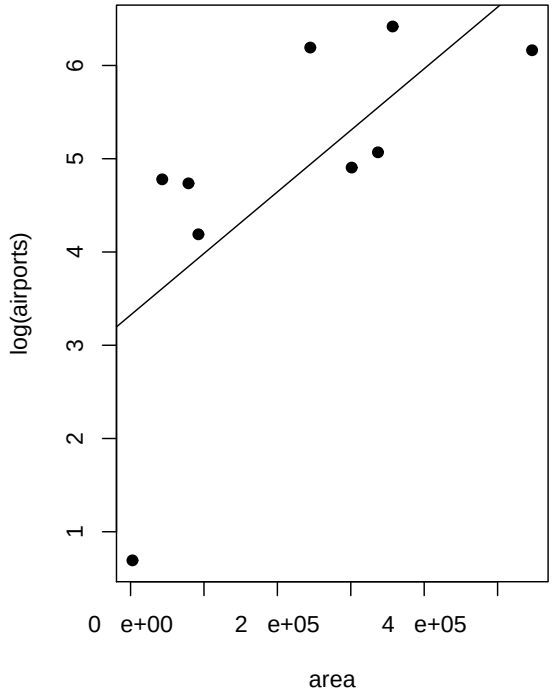
vlivné pozorování (první pozorování daleko ve vodorovném směru)

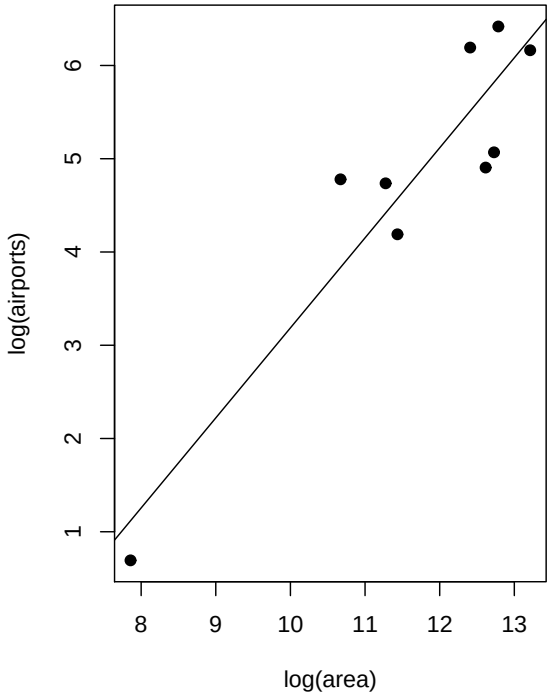


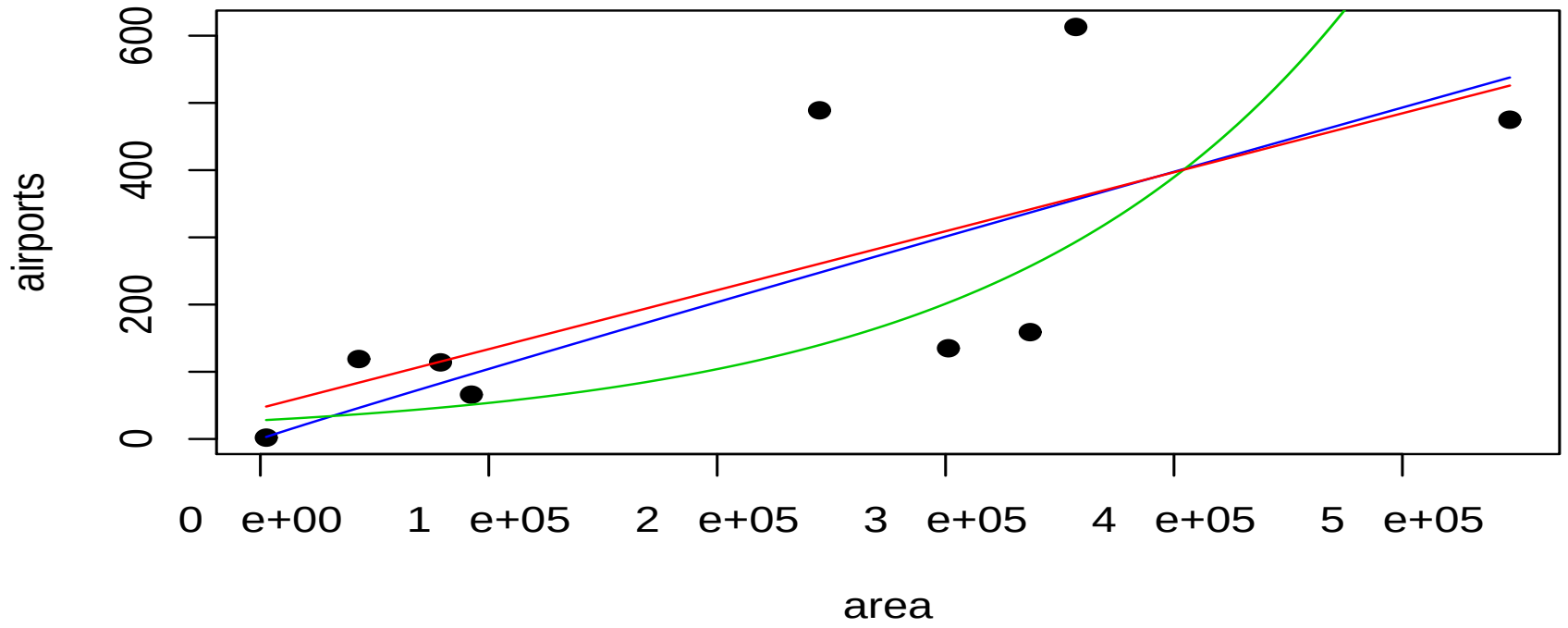



# ukázka transformace: počet letišť na ploše evropského státu



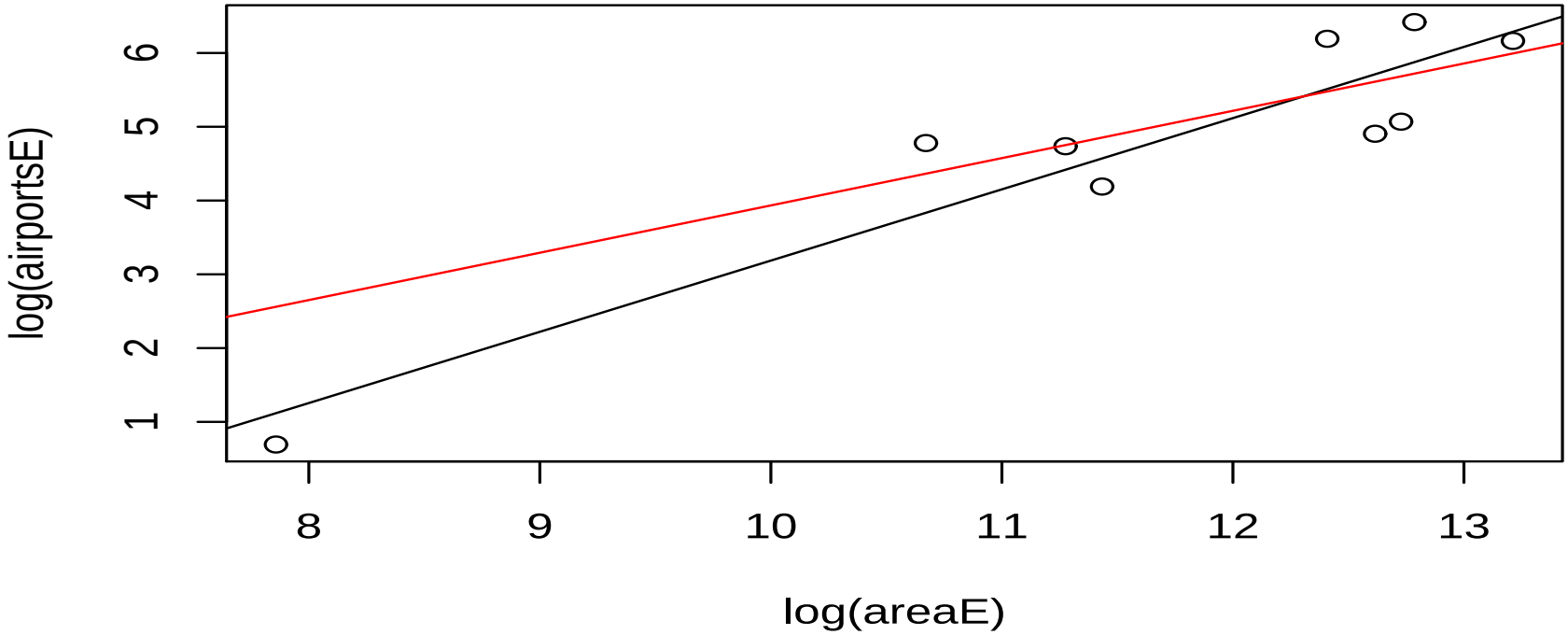





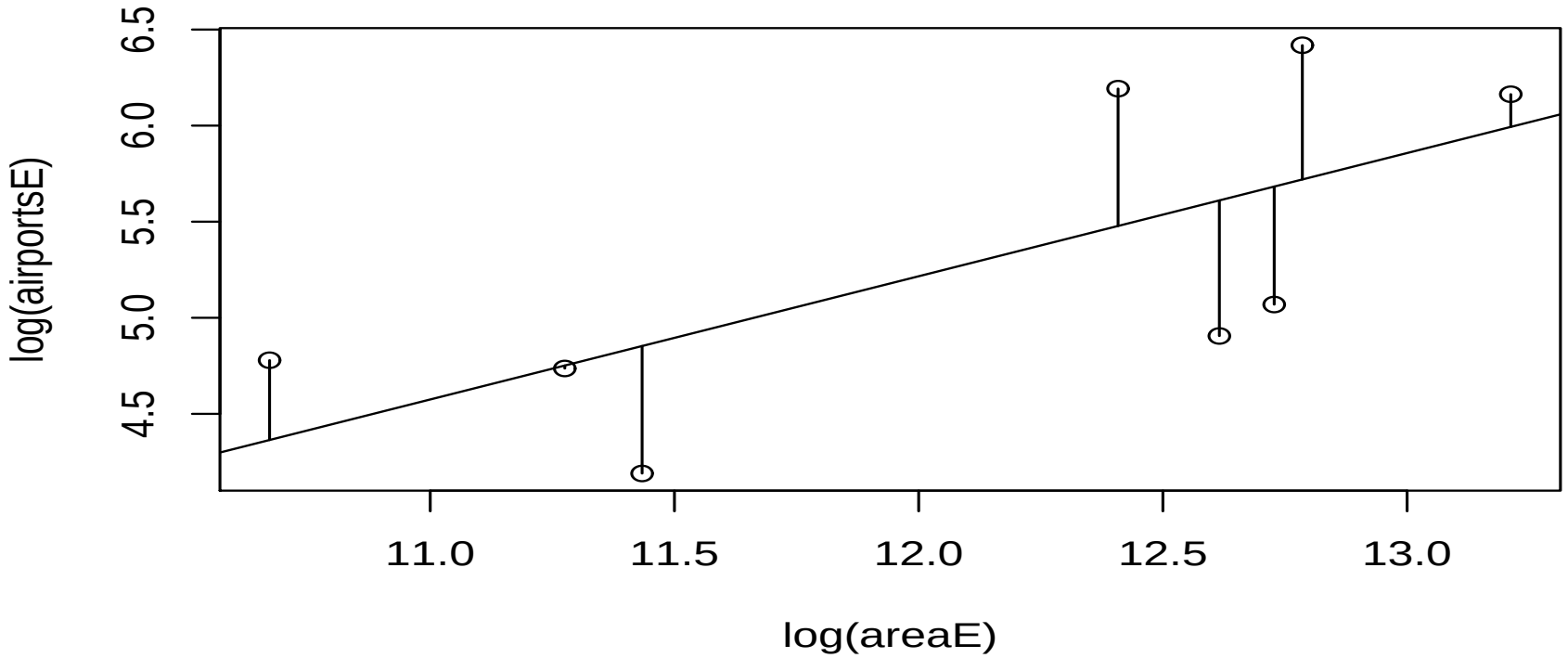

y=46+0,0009 x;log(y)=3,3+0,000007 x;log(y)=-6,5+0,97 log(x)

$R^2 = 51{,}4$ %; $R^2 = 48{,}0$ %; $R^2 = 86{,}1$ %

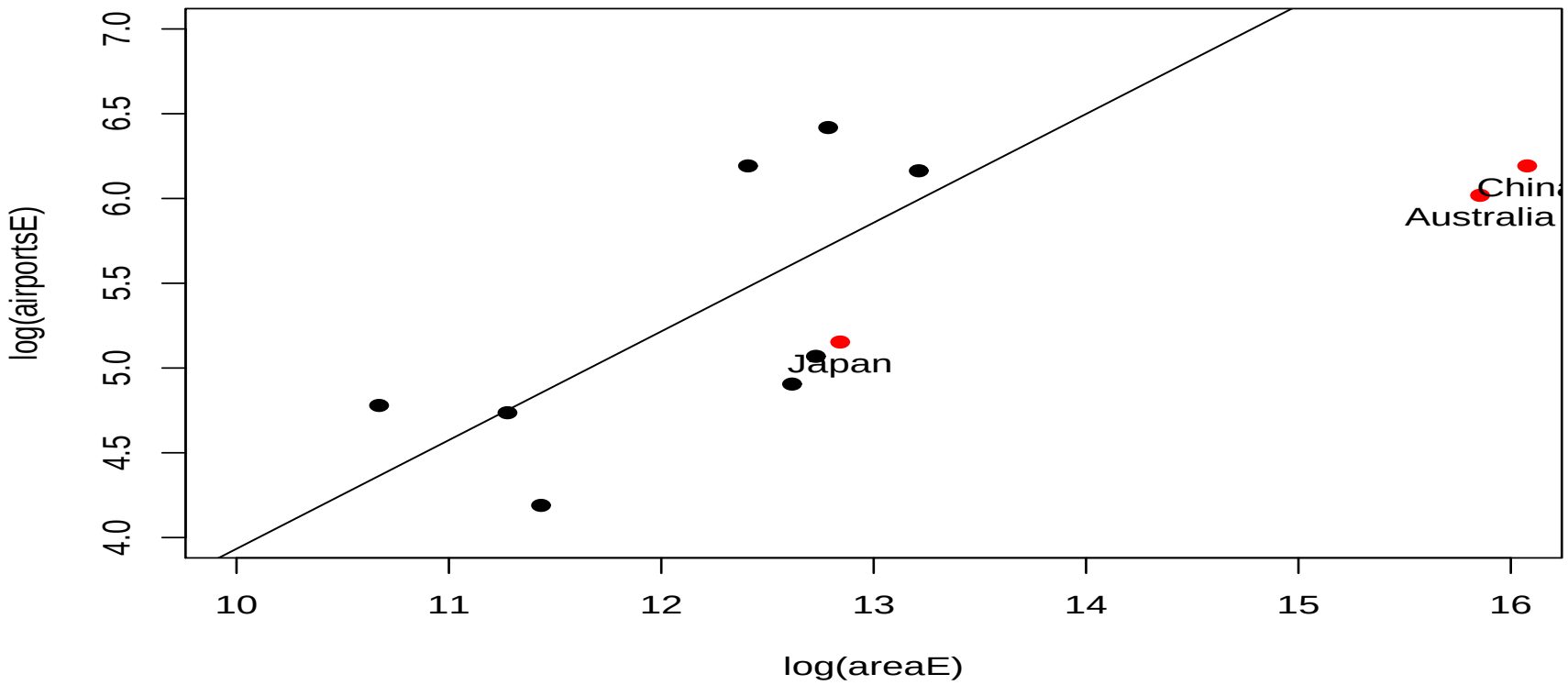
počet letišť: vynechat Lucembursko?



počet letišť: po vynechání Lucemburska



stejná závislost pro Japonsko, Čínu, Austrálii?



# shrnutí

- regrese slouží pro
  - predikci (středních hodnot) budoucích pozorování
  - prokazování závislosti na zvoleném regresoru
  - ověřování modelu o závislosti

- nejsou-li splněny základní předpoklady $\Rightarrow$ pochybné závěry
  - obtížně lze predikovat mimo obor měření
  - je-li malé $R^2$, nespolehlivá předpověď, ale závislost může být prokazatelná
  - vysvětlovaná proměnná může záviset na více nezávisle proměnných, nutná opatrnost (confounding)

# poznámky

- **souvislost regresní přímky a korelačního koeficientu:** testovou statistiku pro $\mathsf{H}_0 : \beta_1 = 0$ lze spočítat také jako

$$T = \frac{r_{xy}}{\sqrt{1 - r_{xy}^2}} \sqrt{n-2}$$

- je-li $|T|$ velké, závislost je prokázána, lze použít (nutno předpokládat normální rozdělení)

- metoda nejmenších čtverců je velmi citlivá na mimořádně umístěná pozorování

- příklad: počet letišť vers. velikost státu (obojí v logaritmech)
  - všech 9 států: $r$ =0,717; $T$ =2,720; $p$ =3,0 %
  - bez Lucemburska: $r$ =0,654; $T$ =2,226; $p$ =7,9 %

# Spearmanův korelační koeficient

- vlastně korelační koeficient pořadí
- citlivě reaguje i na nelineární, ale monotonní závislost
- k prokazování závislosti netřeba normální rozdělení, slabší test
- pro $n \geq 10$ lze předpokládat $r_S\sqrt{n-1} \sim \mathsf{N}(0,1)$
- závislost (proti oboustranné alternativě) prokázána, pokud

$$|r_S\sqrt{n-1}| \geq z(\alpha/2)$$

- závislost (proti jednostranné alternativě) prokázána, pokud

$$r_S\sqrt{n-1} \geq z(\alpha) \text{ resp. } r_S\sqrt{n-1} \leq -z(\alpha)$$

# příklad: počet letišť (přesně $p$ =2,9 %)

| plocha | 78,9 | 43,1 | 337,0 | 547,0 | 357,0 | 301,2 | 92,4 | 244,8 |
|---|---|---|---|---|---|---|---|---|
| letišť | 114 | 119 | 159 | 475 | 613 | 135 | 66 | 489 |
| $R_i$ | 2 | 1 | 6 | 8 | 7 | 5 | 3 | 4 |
| $Q_i$ | 2 | 3 | 5 | 6 | 8 | 4 | 1 | 7 |
| $R_i - Q_i$ | 0 | -2 | 1 | 2 | -1 | 1 | 2 | -3 |
| $(R_i - Q_i)^2$ | 0 | 4 | 1 | 4 | 1 | 1 | 4 | 9 |

$H_0$ : počet letišť **nezávisí** na velikosti státu

$H_1$ : počet letišť **roste** s velikostí státu (jednostranná alternativa)

$$r_S = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (R_i - Q_i)^2 = 1 - \frac{6 \cdot 24}{8(64-1)} = 1 - \frac{144}{504} = 0{,}714$$

$$Z_0 = r_S \sqrt{n-1} = 0{,}714 \cdot \sqrt{7} = 1{,}89$$

$$p = \mathsf{P}(Z \geq Z_0) = 1 - \Phi(Z_0) = 1 - 0{,}971 = 0{,}029$$

na 5% hladině jsme (při jednostranné alternativě) závislost prokázali

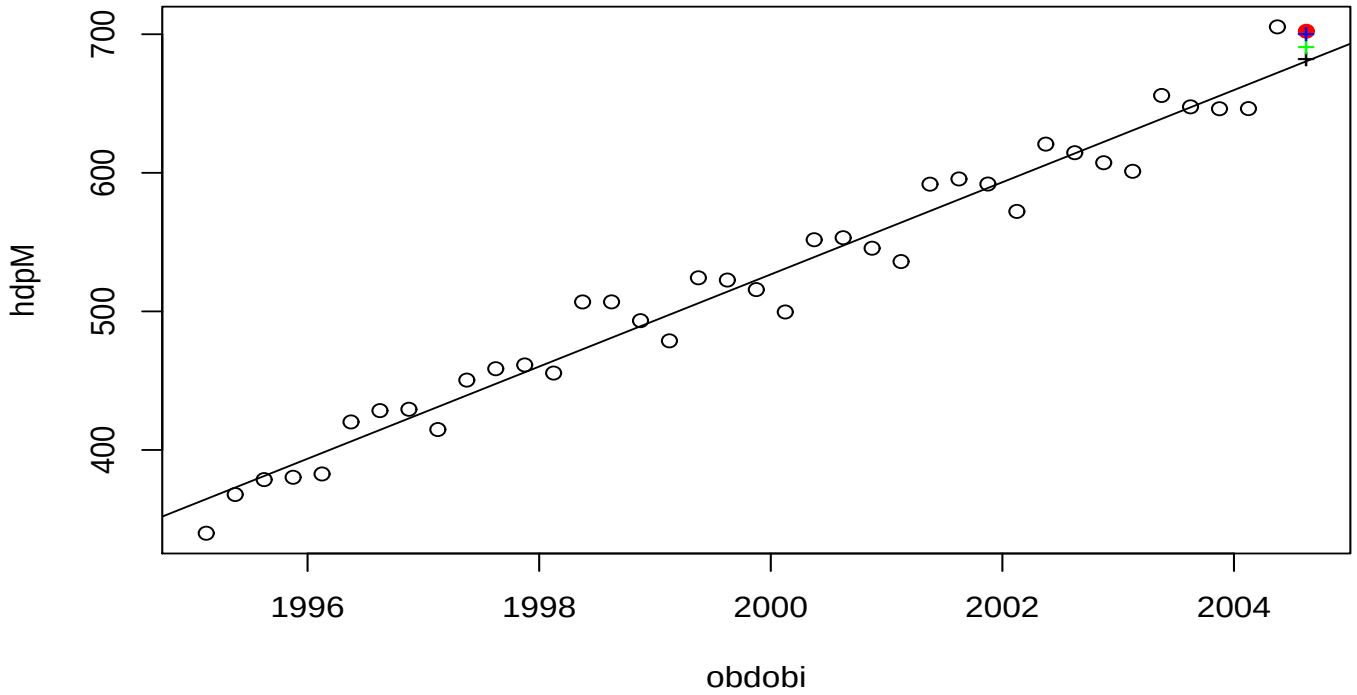
# vyrovnávání

- mechanismus k vyhlazování dat, spíše technický
- cílem je např. nahradit chybějící pozorování (budoucí pozorování – **predikce**, chybějící pozorování - **interpolace**) nebo odstranit nahodilé výkyvy
- často (náš příklad) porušen předpoklad nezávislosti pozorování, proto by obyčejná regrese dala nesprávně přesnost odhadů, testy o parametrech,
- koeficienty nelze snadno statisticky hodnotit, někdy vůbec
- nejen metoda nejmenších čtverců

# časové řady

- spojitý znak měřený v pravidelných časových intervalech
- složeno z několika složek
  - **trend** (dlouhodobý vývoj)
  - **sezónní složka** (periodická složka se známou periodou, např. denní/roční chod teplot, čtvrtletní chod ekonomických veličin)
  - **periodická složka** (s neznámou periodou)
  - **autokorelace** (chybové složky na rozdíl od regrese nejsou mezi sebou nezávislé)
- první dvě složky lze vedle regrese **odhalit** pomocí
  - klouzavých průměrů
  - exponenciálního vyrovnávání
- **prokázat** pomocí regrese

# příklad: HDP v ČR po čtvrtletích (běžné ceny)



# příklad: HDP [mil. Kč]

- predikce bez ohledu na kvartály: 360,4 + 33,3 (rok-1995)
  pro 3. čtvrtletí 2004 tedy předpověď 680,5

- predikce s ohledem na kvartály

$$
\begin{aligned}
est(HDP) &= 339,4 + 33,1(\text{rok} - 1995) \\
est(HDP) &= (335,3 + 38,5) + 33,1(\text{rok} - 1995) \\
est(HDP) &= (335,3 + 30,2) + 33,1(\text{rok} - 1995) \\
est(HDP) &= (335,3 + 18,1) + 33,1(\text{rok} - 1995)
\end{aligned}
$$

  předpověď tedy 335,3 + 30,2 + 33,1 · (2004,75–1995) = 688,6

- predikce s ohledem na autokorelaci (odhad autokorelačního koeficientu $\hat{\rho}$ =0,59): 688,6 + 0,59 · 16,7 = 698,4

- skutečnost: 702,2

# autokorelace, periodicita

- speciální postupy pro zbývající složky (periodická složka s neznámou periodou, autokorelace), nelze mechanicky použít lineární regresi či lineární vyhlazování
- náhodné (chybové) členy v regresním modelu by měly být nezávislé, někdy ale daný člen závisí na předchozím; síla závislosti popsána pomocí autokorelačního koeficientu $\rho$
  - kladné $\rho$: po sobě jdoucí členy podobné (častý případ)
  - záporné $\rho$: po sobě jdoucí členy nepodobné
  - $\rho = 0$: po sobě jdoucí členy nezávislé (v regresi se požaduje)
- autokorelaci a periodicitu lze prokazovat (odhadovat, hodnotit) až po odstranění trendu a sezónního kolísání

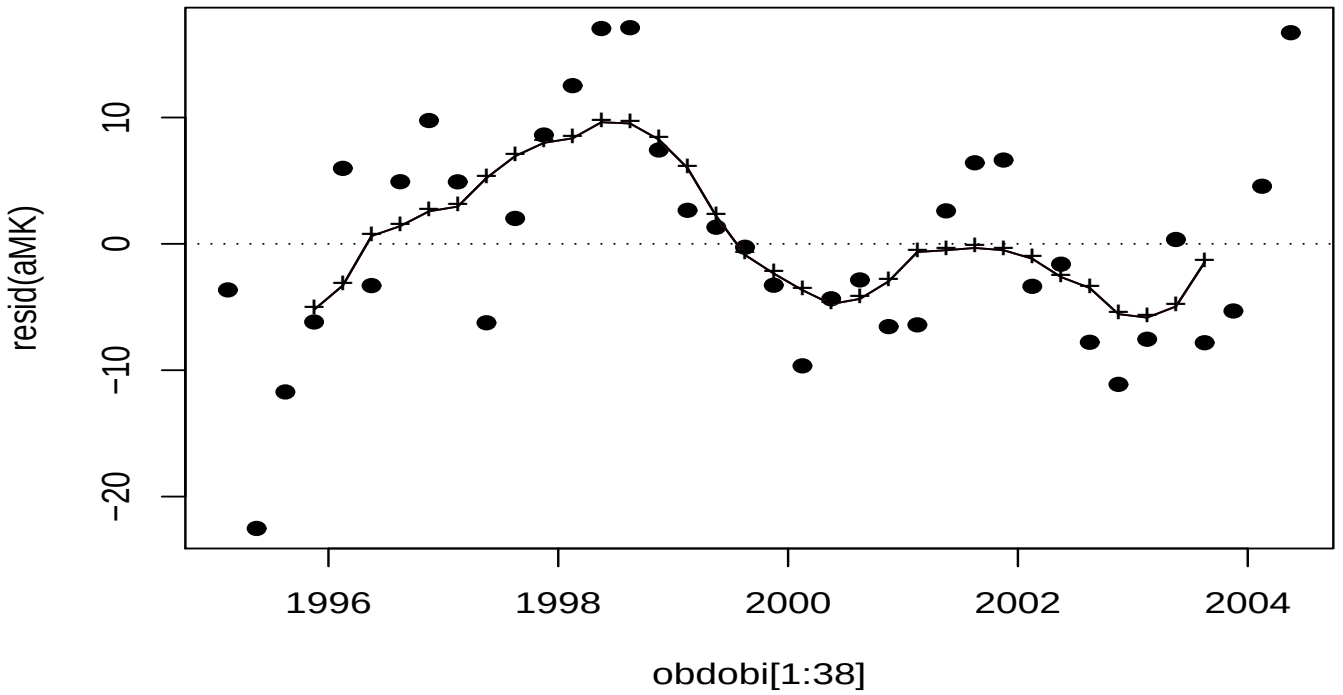
# klouzavé průměry (moving averages)

- pozorování $Y_i$ porovnáváme s průměrem z $m$ sousedních pozorování (včetně $Y_i$ samotného), např. pro $m = 5$

$$\frac{1}{5}(Y_{i-2} + Y_{i-1} + Y_i + Y_{i+1} + Y_{i+2})$$

- snaha vyhladit nahodilé výchylky, zachovat „průměrný“ vývoj
- vhodné zejména k interpolaci, k nalézání dosud přehlížených systematických vlivů
- u HDP vezmeme nejprve v úvahu lineární trend a čtvrtletní periodicitu, spočítáme rezidua (to, co nevysvětlíme lineárním trendem a čtvrtletními sezonními výkyvy)

# příklad: klouzavé průměry reziduí ($m$ =7)



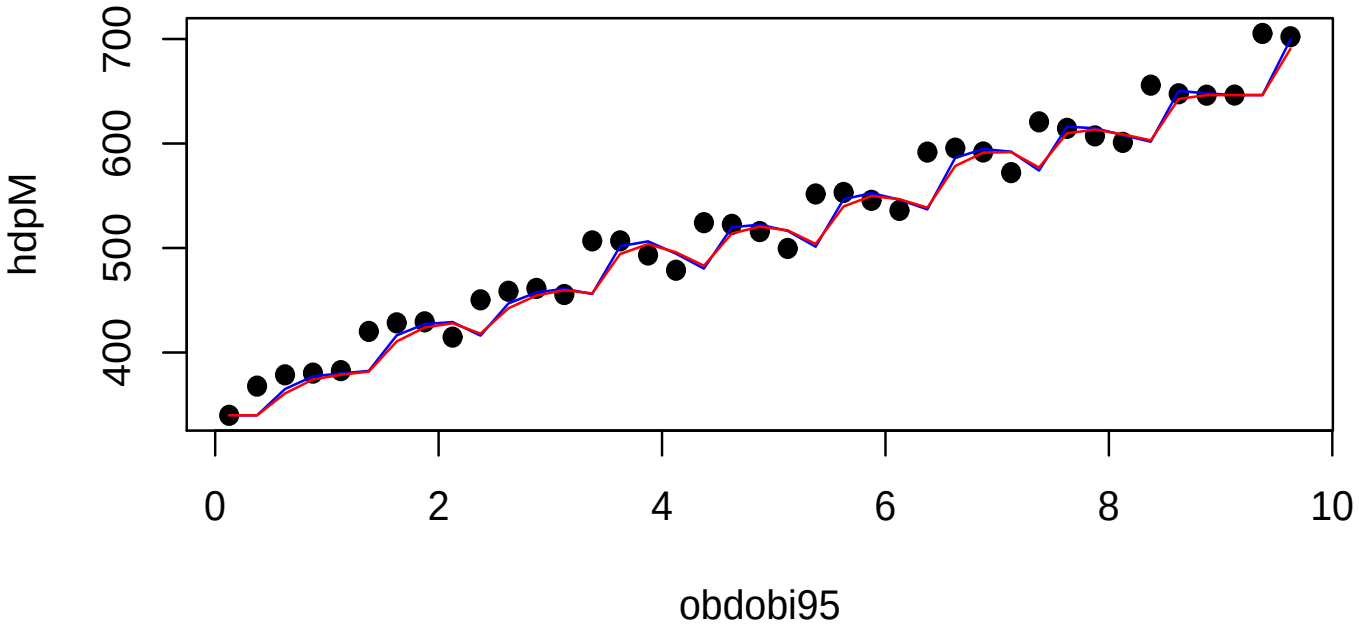
# exponenciální vyrovnávání

- představa: v posledním pozorování se projevuje vliv všech předchozích
- tento vliv je postupně utlumován: největší vliv má předchozí pozorování, nejmenší pozorování v čase nejvzdálenější
- vážený průměr mezi předpovědí předchozího pozorování a skutečným předchozím pozorováním

$$w\hat{Y}_{i-1} + (1-w)Y_{i-1}$$

- $w$ – váha „historie“, čím je větší, tím méně kolísání
- použitelné k předpovědi

příklad: exponenciální vyrovnávání (699,4 pro w=0,1, 690,6 pro w=0,25)



exponenciální vyrovnávání reziduí (704,0 pro w=0,1, 701,6 pro w=0,25)