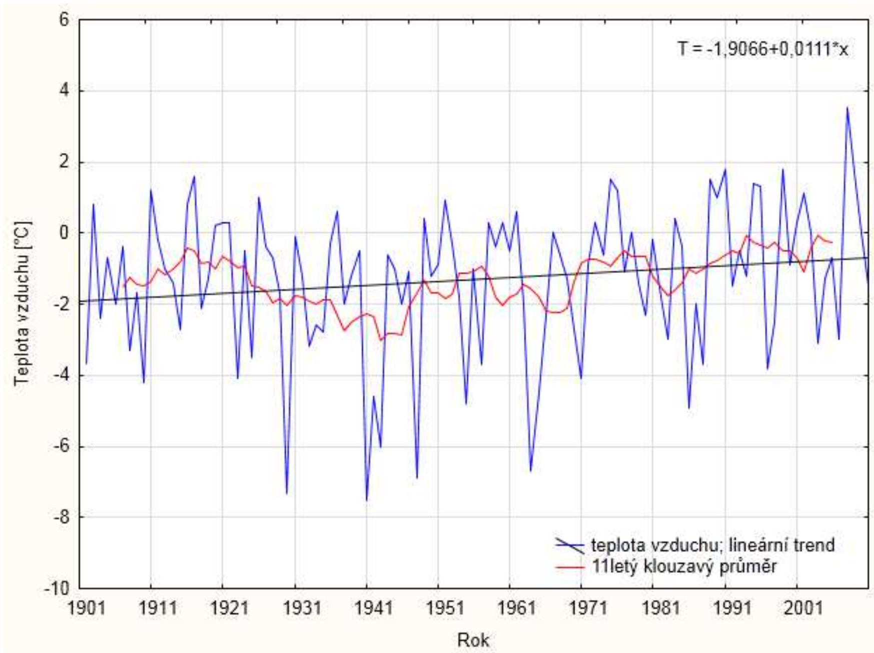
Statistické metody II, cvičení č. 5

Analýza trendu (odhad lineárního trendu a klouzavé průměry)



Brno, 11.4.2016
Klára Ambrožová

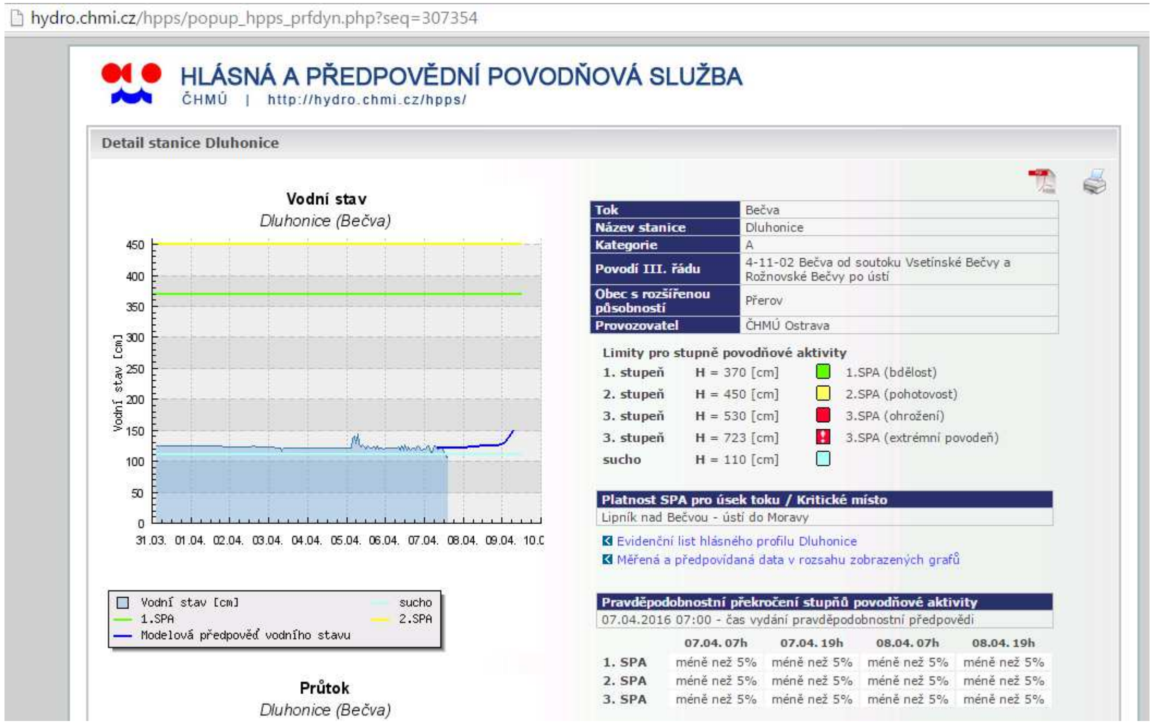
Motivační příklad

Úvod

- Rodina Mráčkova by si ráda postavila rodinný domek.
- Velmi se jim zalíbily Dluhonice, část města Přerov.
- Tímto územím však protéká řeka Bečva a rodina se obává výskytu povodní.
- Řeší tedy otázku: Jak často se v Dluhonicích vyskytují povodně v současné době a co lze očekávat do budoucna?

Data

- Dosažení 2. či vyššího SPA na stanici Dluhonice podle Hydrologických ročenek ČHMÚ (http://voda.chmi.cz/roc/index.html)
- Průměrný počet dosažení v období 1999–2015: 0,82



	1 Rok	2 t	3 yt
1	1999	1	0
2	2000	2	1
3	2001	3	0
4	2002	4	2
5	2003	5	0
6	2004	6	0
7	2005	7	1
8	2006	8	0
9	2007	9	1
10	2008	10	0
11	2009	11	1
12	2010	12	2
13	2011	13	1
14	2012	14	2
15	2013	15	0
16	2014	16	1
17	2015	17	2

Pozn. Data byla z didaktických důvodů upravena, nejde o reálné hodnoty

Mění se počet povodní?

- Nalézt funkci závislosti počtu dosažení 2. SPA na čase

  yt … počet dosažení

  t … čas v letech

$$yt = f\,(t)$$

- Možnosti:
  - Model konstanty

    $yt = b_0$ (~ průměr)

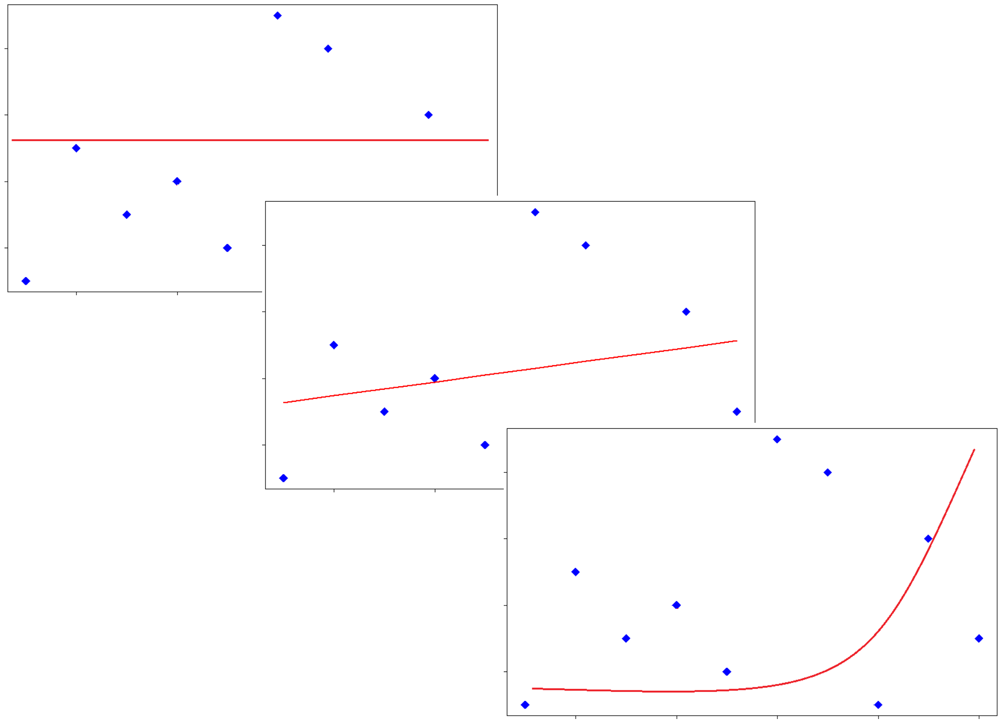
  - Lineární model

    $yt = b_0 + b_1 * t$

  - Exponenciální model

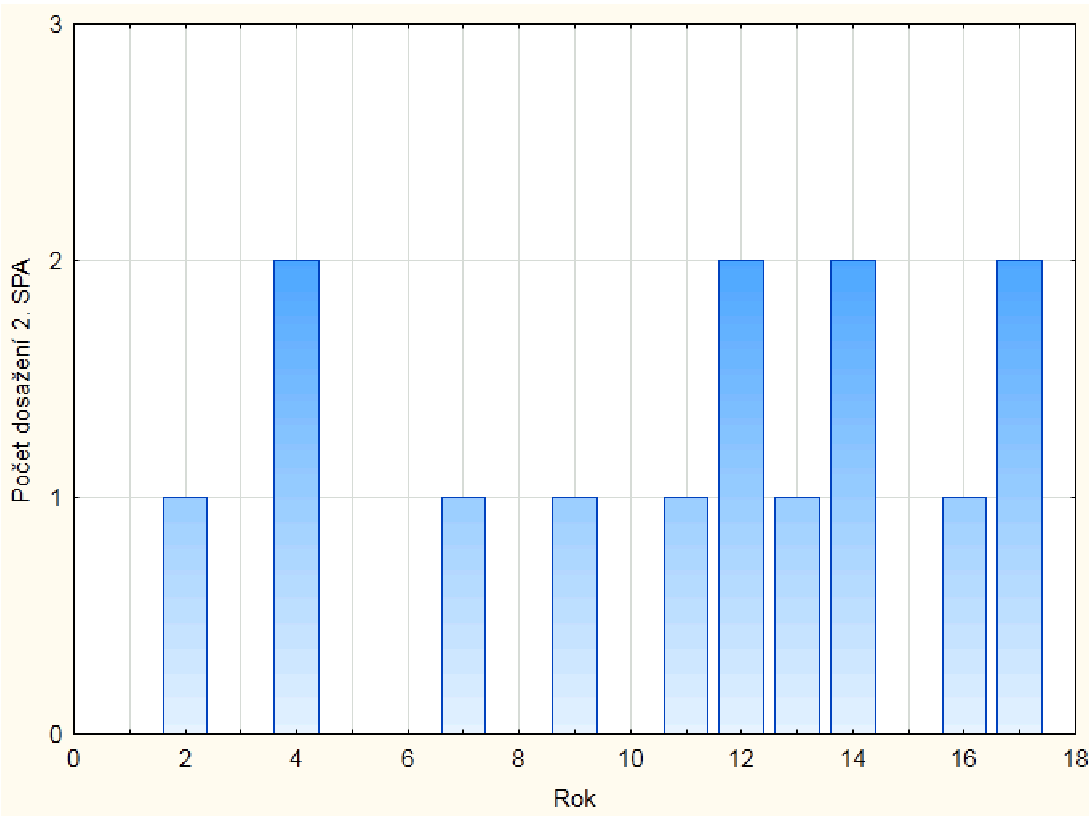
    $yt = b_0 * b_1^{t}$

  …



1) Grafické znázornění

- Subjektivně pohledem zhodnotit, který trend by byl vhodný

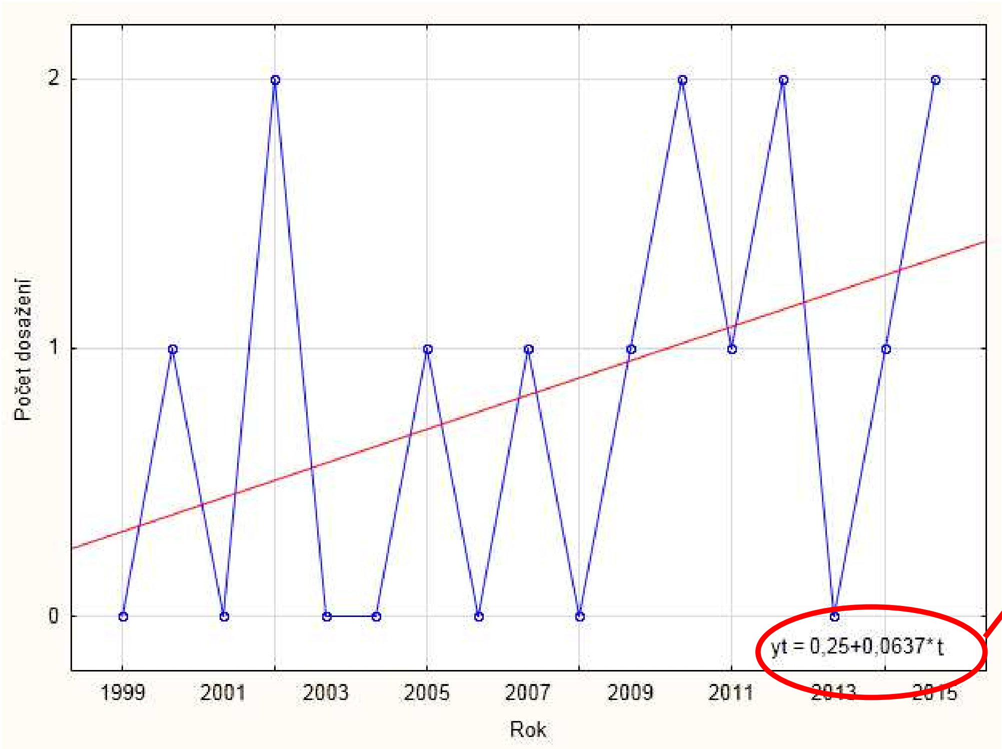


- Trend je obtížné odhadnout → nejlépe začít s nejjednoduším

Pozn. Tento krok je při vypracování cvičení vynechán.

2) Proložení lineárním trendem

- Nalezení vhodné lineární funkce metodou nejmenších čtverců



$$b_1 = \frac{\sum_{t=1}^{n} t y_t - \bar{t}\sum_{t=1}^{n} y_t}{\sum_{t=1}^{n} t^2 - n\bar{t}^2} \qquad b_0 = \bar{y} - b_1\bar{t}$$

yt = 0,25 + 0,0637*x

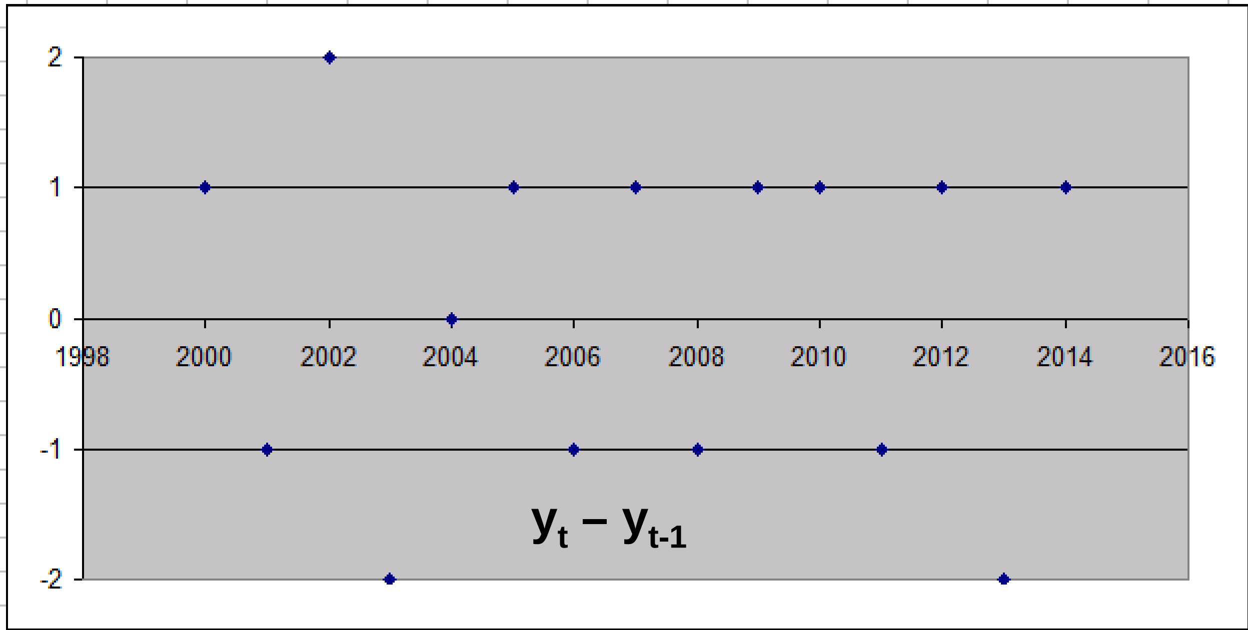
- **Interpretace:** Získali jsme rovnici lineárního trendu, kde koeficient b_1 je kladný (0,0637), v Dluhonicích tedy dochází k nárůstu počtu dosažení a překročení 2. SPA

3) Je lineární model vhodný?

1. Informativní test: 1. diference přibližně konstantní

	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	
y_{t-1}		0	1	0	2	0	0	1	0	1	0	1	2	1	2	0	1	2
y_t	0	1	0	2	0	0	1	0	1	0	1	2	1	2	0	1	2	
$y_t - y_{t-1}$		1	-1	2	-2	0	1	-1	1	-1	1	1	-1	1	-2	1	1	

Rozdíl mezi hodnotou v daném a předchozím roce



Interpretace: 1. diference nejsou konstantní, ale není patrný ani zřetelný nárůst či pokles → nelze vyloučit vhodnost modelu

Pozn. Tento krok je při vypracování cvičení vynechán.

3) Je lineární model vhodný?

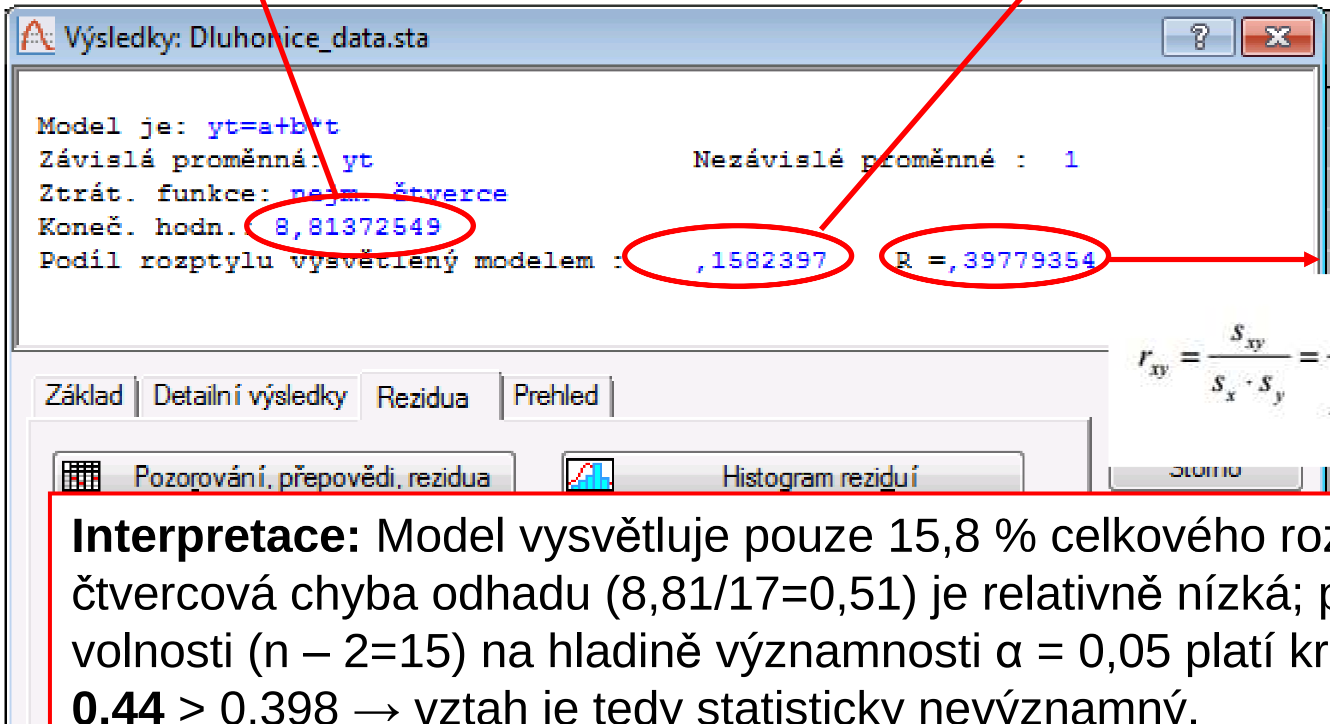
2. Hodnocení pomocí objektivních kritérií

Vydělíme-li počtem hodnot (n), vznikne **střední čtvercová chyba odhadu**

$$M.S.E. = \frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{n}$$

Koeficient determinace (r_{xy}^2; uvádí, jak velkou část rozptylu vysvětluje model)

Koeficient korelace – těsnost lineárního vztahu (ve vzorci x odpovídá t a y odpovídá yt)



$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\frac{1}{n-1}\sum(x_i - \bar{x})\cdot(y_i - \bar{y})}{\sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2 \cdot \frac{1}{n-1}\sum(y_i - \bar{y})^2}}$$

Interpretace: Model vysvětluje pouze 15,8 % celkového rozptylu; střední čtvercová chyba odhadu (8,81/17=0,51) je relativně nízká; pro daný počet stupňů volnosti (n – 2=15) na hladině významnosti α = 0,05 platí kritická hodnota $\mathbf{r_{krit} = 0{,}44}$ > 0,398 → vztah je tedy statisticky nevýznamný.

3) Je lineární model vhodný?

3. **Analýza rozptylu** – aproximuje lineární model data lépe než model konstanty (~ průměr)?

H_0: Model konstanty je dostačující

Efekt	Model je: yt=a+b*t (Dluhonice_data.sta) Záv.prom.:yt 1 Součet čtverců	2 SV	3 Průměrný čtverec	4 F-hodnota	5 p-hodnota
Regrese	13,18627	2,00000	6,593137	11,22080	0,001048
Rezidua	8,81373	15,00000	0,587582		
Celkem	22,00000	17,00000			
Opravený součet	10,47059	16,00000			
Regrese vs. Opravený součet	13,18627	2,00000	6,593137	10,07491	0,001473

Testová statistika F:

$$F = \frac{S_R / p}{S_E / (n - p - 1)}$$

Interpretace: Hodnota testovacího kritéria je 11,22, čemuž odpovídá p-hodnota 0,001 → zamítáme hypotézu, že dostačující je model konstanty.

3) Je lineární model vhodný?

4. **Testování významnosti regresních parametrů** – jsou v rovnici potřeba všechny členy (neboli nestačil by model yt=a nebo yt=b*t)?

Model je: yt=a+b*t (Dluhonice_data.sta)
Záv.prom.:yt
Hladina spolehlivosti:95.0% (alfa =0.050)

	Odhad	Standard chyba	t-hodn. sv = 15	p-hodn.	Dol. sp. Mez	Hor. sp. Mez
a	0,250000	0,388865	0,642897	0,529997	-0,578846	1,078846
b	0,063725	0,037949	1,679226	0,113811	-0,017162	0,144613

Odhad absolutního členu

Odhad směrnice

Testovací kritérium t-testu pro absolutní člen (a) a směrnici (b) vč. odpovídající p-hodnoty

Interpretace: Výsledná rovnice by měla tvar yt = 0,25 + 0,064*t. V daném případě však není ani jeden z regresních parametrů statisticky významný (p-hodnoty jsou větší než 0,05), takže by bylo třeba zvolit jiný model.

ZÁVĚR: Počet dosažených a překročených 2. SPA se s časem mění, nikoliv však lineárně – lineární model proto v tomto případě není vhodný.

Analýza reziduí

- Následně se provádí ještě analýza reziduí
- Rezidua by měla mít následující vlastnosti:
  - normálně rozložená
    (karta Rezidua – Normál.pravd.graf reziduí – body by měly ležet na přímce)
  - mít konstantní rozptyl
    (karta Rezidua – Rezidua vs. Předpovědi – rozmístění kolem nuly by mělo být náhodné)
  - mít nulovou střední hodnotu
    (lze určit pomocí t-testu pro samostatný vzorek – testujeme hypotézu, že se rezidua stat. Významně neliší od nuly)
  - být nezávislá
    (provádí se přes modul Vícenásobná regrese pomocí Durbin-Watsonovy statistiky, která by měla být přibližně v intervalu od 1,4 do 2,6)

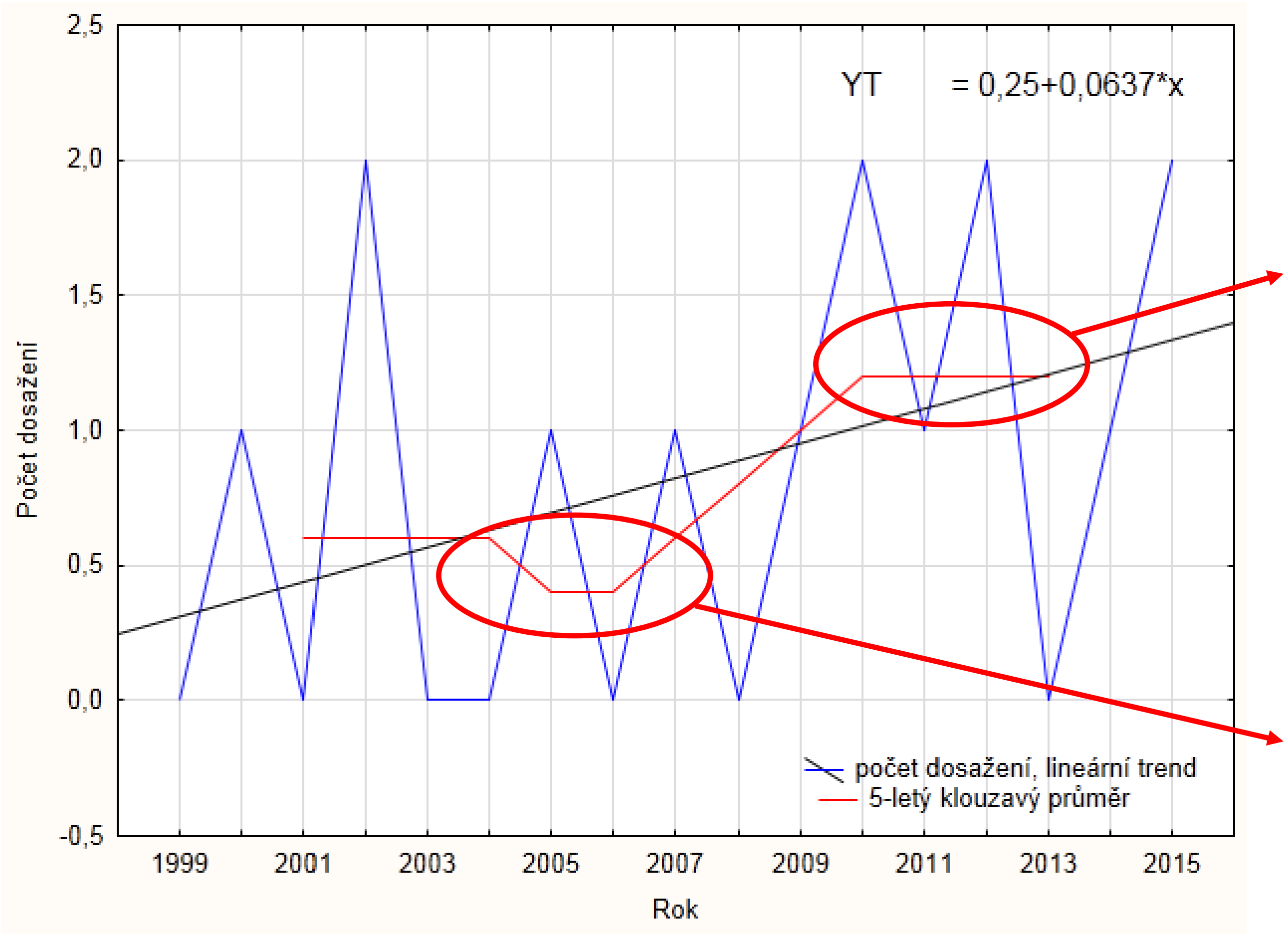
ZÁVĚR: Pokud rezidua splňují výše uvedená kritéria, je model vhodný.

Motivační příklad – pokračování

- Lineární trend nevhodný pro aproximaci počtu dosažení a překročení 2. SPA na stanici Dluhonice
  x počet není v čase konstantní!

- Zjištění období nárůstu a poklesu počtu hodnot = metoda klouzavých průměrů
  - prosté $\frac{1}{5}(y_{t-2} + y_{t-1} + y_t + y_{t+1} + y_{t+2})$
  - vážené $\frac{1}{8}(y_{t-2} + 2y_{t-1} + 2y_t + 2y_{t+1} + y_{t+2})$

Metoda klouzavých průměrů



Od r. 2006 docházelo k nárůstu počtu dosažení a 2010–2013 bylo období nejvyšších dosažených hodnot za dobu sledování

V období 2005–2006 docházelo k poklesu počtu dosažení

ZÁVĚR: Přestože lineární trend nebyl vhodným modelem pro danou problematiku, z analýzy klouzavých průměrů lze zjistit, že v poslední době (od r. 2006) se počet dosažení 2. SPA v Dluhonicích zvyšuje.

Pozn. Ve cvičení se používá 11-letý klouzavý průměr.

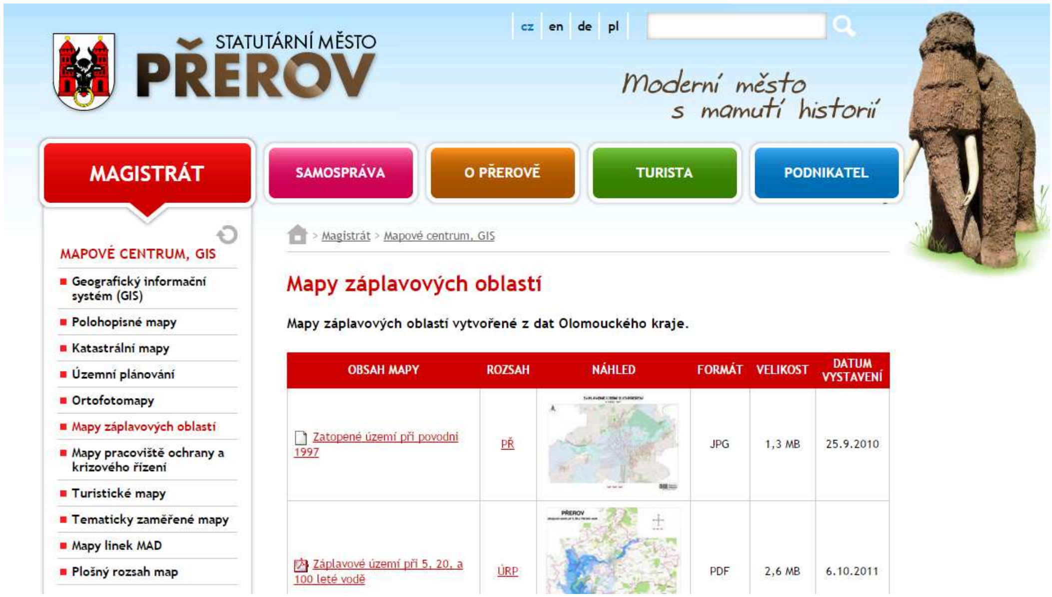
Zdroje:

- BUDÍKOVÁ, Marie. Jednoduchá lineární regrese (přednáška). Brno: Masarykova univerzita,8.4. 2016.

- DOBROVOLNÝ, Petr. Z2069 Statistické metody a zpracování dat II: Analýza časových řad (přednáška) Brno: Masarykova univerzita,8.4. 2016.

Poznámka:

- Pro problém z motivačního příkladu by bylo vhodnější se podívat na záplavová oblastí města Přerov a hodnotit riziko pro konkrétní lokalitu (http://www.prerov.eu/cs/magistrat/mapove-centrum-gis/mapy-zaplavovych-oblasti.html)



Poznámka ke cvičení

- Cvičení bude obsahovat:

  1) Graf časové řady s lineárním trendem vč. rovnice,

  2) Tabulku s koeficienty rovnice (a, b), F-hodnotu analýzy rozptylu, střední čtvercovou chybou odhadu, koeficienty korelace a determinace – vč. p-hodnot tam, kde jsou k dispozici,

  3) NP-graf pro rezidua a obrázek Rezidua vs. Předpovědi,

  4) Graf časové řady s lineárním trendem a 11-letým klouzavým průměrem

- Závěr bude obsahovat:
  - Hodnocení vhodnosti proložení časové řady lineárním trendem
  - Interpretace lineárního trendu – růst či pokles hodnot v studovaném období?
  - Hodnocení klouzavých průměrů – období výskytu minimálních resp. maximálních teplot a také období největšího růstu (resp. poklesu) teplot ve studovaném období

- Všechny kroky v programu Statistica nutné k vypracování jsou uvedeny v popisu cvičení