

Cvičení 10:

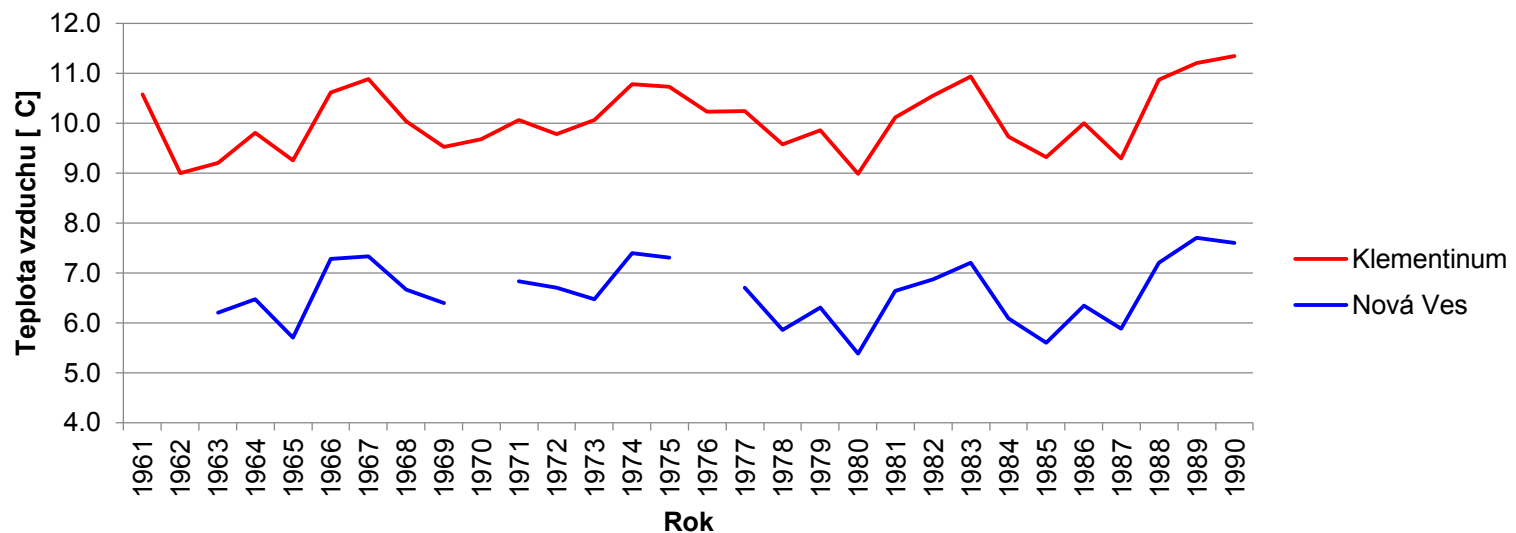
Konstrukce regresního modelu

Statistické metody a zpracování dat 1 (podzim 2016)

Klára Čížková

Řešený příklad – 1. úkol

- Máme soubor průměrných ročních teplot vzduchu ze stanice Nová Ves za období 1961–1990, kde chybějí 4 roky
- S využitím údajů z vaší stanice (viz cvičení 3) doplníme chybějící hodnoty



POZOR! Vaše výsledky se budou lišit!

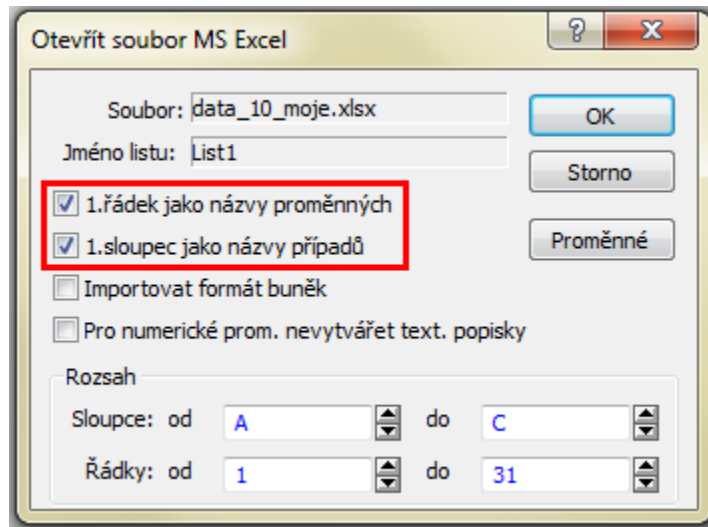
Jako referenční řadu jsem použila řadu průměrných ročních teplot vzduchu ze stanice Praha-Klementinum (1961–1990)

Příprava export dat

1. Otevřeme...
2. De..._Ves
Už hotovo – stačí vybrat případy označené vašim jménem a zkopírovat proměnné ROK, KLEM a NVES (i se záhlavím).
... průměrné roční teploty
... stanice
... pojmenujeme sloupce, uložíme a
... avíme

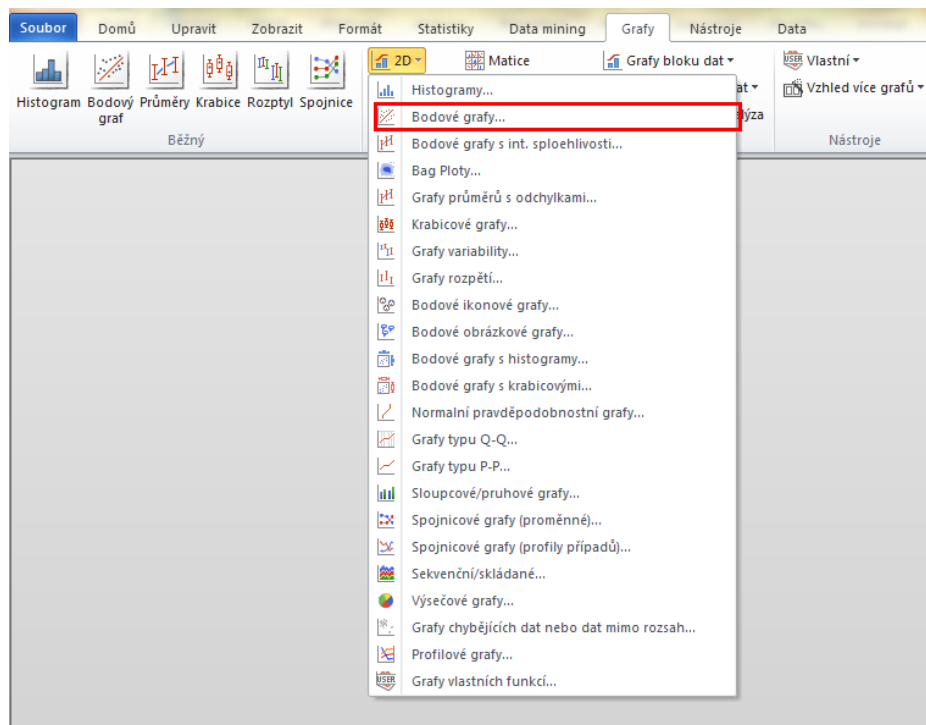
4. Importujeme do STATISTICY: Soubor – Otevřít; vybraný list; 1. řádek jako názvy proměnných, 1.sloupec jako názvy případů; OK

	A	B	C
1	ROK	KLEM	NVES
2	1961	10.6	
3	1962	9.0	
4	1963	9.2	6.2
5	1964	9.8	6.5
6	1965	9.3	5.7
7	1966	10.6	7.3
8	1967	10.9	7.3
9	1968	10.0	6.7
10	1969	9.5	6.4
11	1970	9.7	
12	1971	10.1	6.8
13	1972	9.8	6.7
14	1973	10.1	6.5
15	1974	10.8	7.4
16	1975	10.7	7.3
17	1976	10.2	
18	1977	10.2	6.7
19	1978	9.6	5.9
20	1979	9.9	6.3
21	1980	9.0	5.4
22	1981	10.1	6.6
23	1982	10.6	6.9
24	1983	10.9	7.2
25	1984	9.7	6.1
26	1985	9.3	5.6
27	1986	10.0	6.3
28	1987	9.3	5.9
29	1988	10.9	7.2
30	1989	11.2	7.7
31	1990	11.3	7.6

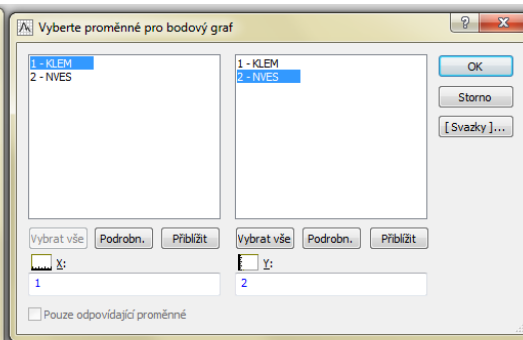
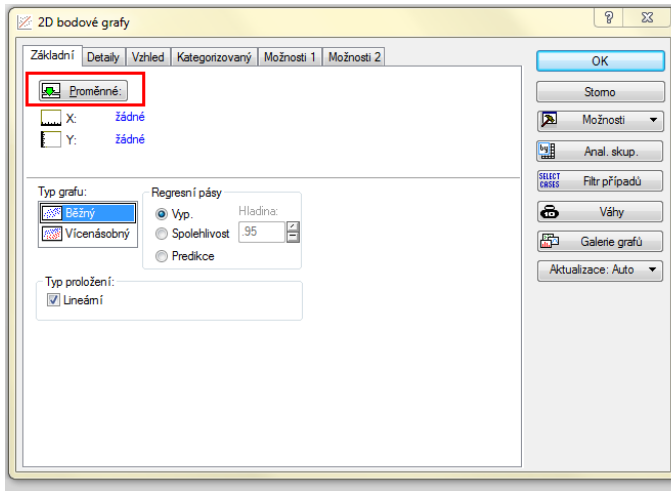


Ověření korelačního vztahu

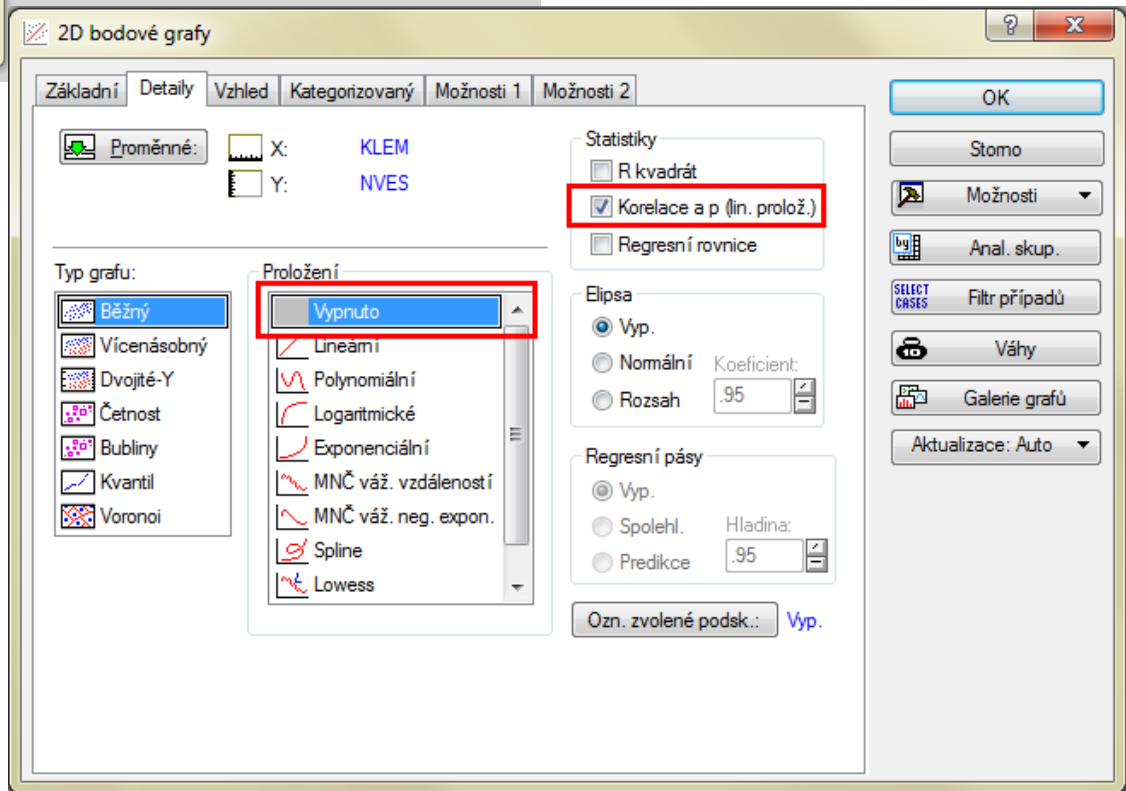
- Pomocí výpočtu **korelačního koeficientu** a vykreslením **bodového grafu** ověříme existenci a těsnost korelačního vztahu mezi průměrnou roční teplotou vzduchu v Nové Vsi a na vaší stanici
- **Před provedením tohoto kroku by bylo vhodné posoudit normalitu rozdělení!**
- Grafy – 2D – Bodové grafy



Ověření korelačního vztahu

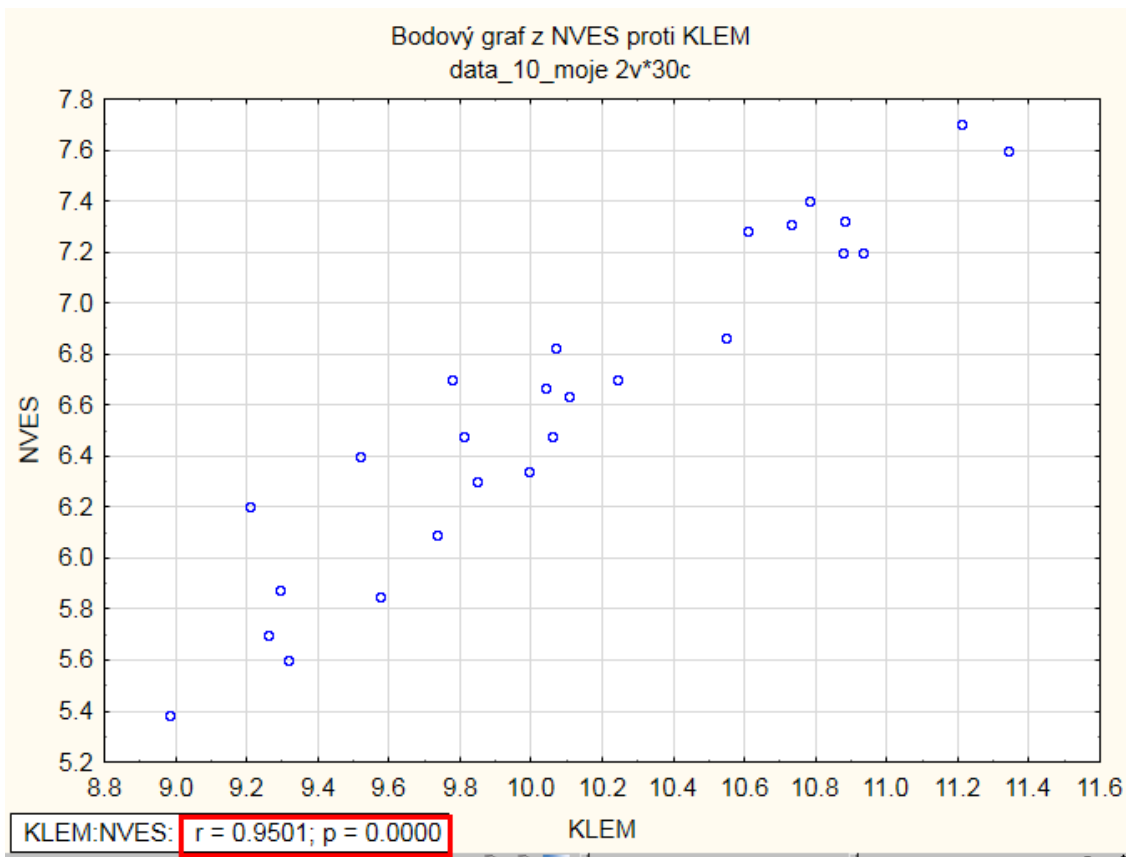


- Nastavení bodového grafu:
 - Vybereme **proměnné** (X = vaše stanice, Y = Nová Ves), OK
 - Na záložce **Detaily** vypneme lineární proložení a zadáme výpočet korelačního koeficientu a p-hodnoty, OK



Ověření korelačního vztahu

- Interpretace bodového grafu a p-hodnoty



$r_{xy} = 0$ → neexistuje lineární vztah

$r_{xy} > 0$ → kladná korelace

$r_{xy} < 0$ → záporná korelace

Nulová hypotéza: mezi průměrnou roční teplotou vzduchu v Nové Vsi a na vaší stanici neexistuje statisticky významný lineární vztah.

Hladina významnosti = 0,05

$p > 0,05$ → H_0 platí

$p < 0,05$ → H_0 zamítáme

Ukázka interpretace: Průměrné roční teploty vzduchu z Nové Vsi a Klementina jsou **velmi silně/středně silně/slabě pozitivně/negativně** korelovány ($r_{xy} = ?,??$). P-hodnota činí $?,??$, takže na hladině významnosti 0,05 **zamítáme/přijímáme** nulovou hypotézu, že mezi těmito soubory neexistuje lineární vztah.

Vytvoření regresního modelu

- Obecný tvar rovnice lineární regrese:

$$y' = a + bx$$

(y' je nejpravděpodobnější teoretická hodnota y odpovídající danému x)

- Koeficient **a** představuje hodnotu y' při $x = 0$
- Koeficient **b** je směrnicí regresní přímky ($b > 0 \rightarrow$ pozitivní regrese; $b < 0 \rightarrow$ negativní regrese)
- Výpočet pomocí **metody nejmenších čtverců** (součet druhých mocnin vzdálenosti bodů korelačního pole od regresní přímky je minimální)



Vytvoření regresního modelu

Vypočítáme parametry regresní přímky a koeficient determinace r^2

- Koeficient determinace r^2 je druhou mocninou korelačního koeficientu, nabývá hodnot od 0 do 1 a udává, jakou část variability závisle proměnné lze vysvětlit použitým regresním modelem

1. Statistika – Vícenásobná regrese – Proměnné: závislá = Nová Ves, nezávislá = vaše stanice – OK

The screenshot displays the Minitab software interface. The top menu bar includes 'Soubor', 'Domů', 'Upravit', 'Zobrazit', 'Formát', 'Statistika', 'Data mining', 'Grafy', 'Nástroje', and 'Data'. The 'Statistika' menu is open, showing options like 'Základní statistiky', ' Vícenásobná regrese', 'ANOVA', 'Neparametrické statistiky', 'Prokládání rozdělání', and 'Rozdělení a simulace'. The ' Vícenásobná regrese' option is highlighted with a red box.

Below the menu, the ' Vícenásobná lineární regrese: data_10_moje' dialog box is open. The 'Základní nastavení' tab is active, and the 'Proměnné' button is highlighted with a red box. The 'Závislé:' field contains 'žádné' and the 'Nezávislé:' field contains 'žádné'. The 'Vážené momenty' checkbox is unchecked, and the 'SV =' field is set to 'W-1'. The 'ChD vynechána' section has 'Celé případy' selected.

To the right, the 'Vyberte závislé a nezávislé proměnné' dialog box is open. It shows two lists: 'Závislá prom. (n. seznam pro dávku):' containing '2' and 'Seznam nezáv. proměnných:' containing '1'. The 'Pouze odpovídající proměnné' checkbox is checked.

Vytvoření regresního modelu

2. Výsledná tabulka udá r_{xy} („vícenásobné R“), **koeficient determinace („R2“)** a adjustovaný koeficient determinace („upravené R2“), který používáme, pokud chceme porovnat více regresních modelů použitých pro vysvětlení stejné proměnné.

Výsledky - vícenásobná regrese: data_10_moje

Výsledky- vícerozm. regrese

Záv.prom. : NVES	vícenás. R = .95014810	F = 222.8664
	R2= .90278142	sv = 1,24
Poč. případů: 26	upravené R2= .89873064	p = .000000
	Směrodatná chyba odhadu : .203534643	
Abs. člen: -2.611134148	Sm. chyba: .6195834	t(24) = -4.214 p = .0003

KLEM b* = .950

(významná b* jsou zvýrazněna červeně)

Alfa pro zvýraznění efektů: .05

Základní výsledky | Detailní výsledky | Residua/předpoklady/předpovědi

Výpočet: Výsledky regrese

OK
Storno
Možnosti
Anal. Skup.

Když vynásobíme koeficient determinace stem, dostaneme procentuální podíl variability vysvětlené daným modelem.

Čím vyšší je hodnota koeficientu determinace, tím je model vhodnější.

V tomto případě vysvětluje použitý regresní model přibližně 90,3 % variability řady průměrných ročních teplot vzduchu v Nové Vsi.

Vytvoření regresního modelu

3. Ve výsledcích najdeme rovněž **směrodatnou chybu odhadu**, která vyjadřuje směrodatnou odchylku reziduálních hodnot. Slouží k posouzení vhodnosti modelu – **čím nižší směrodatná chyba odhadu, tím vhodnější model.**

Výsledky - vícenásobná regrese: data_10_moje

Výsledky- vícerozm. regrese

Záv.prom. : NVES vícenás. R = .95014810 F = 222.8664
R2 = .90278142 sv = 1,24
Poč. případů: 26 upravené R2 = .89873064 p = .000000
Směrodatná chyba odhadu : .203534643
Abs. člen: -2.611134148 Sm. chyba: .6195834 t(24) = -4.214 p = .0003

KLEM b* = .950

(významná b* jsou zvýrazněna červeně)

Alfa pro zvýraznění efektů: .05

Základní výsledky | Detailní výsledky | Residua/předpoklady/předpovědi

Výpočet: Výsledky regrese

OK
Storno
Možnosti
Anal. Skup.

V tomto případě činí směrodatná chyba odhadu 0,20.

TIP: porovnejte koeficient determinace a směrodatnou chybu odhadu, které jste získali vy, s hodnotami z této prezentace (Klementinum). Je lepší vysvětlující proměnnou pro teplotu vzduchu v Nové Vsi vaše řada, nebo řada z Klementina?

4. Parametry regresní rovnice získáme kliknutím na „**Výpočet: Výsledky regrese**“.

Vytvoření regresního modelu

Parametry regresní rovnice:

Výsledky regrese se závislou proměnnou : NVES (data_10_moje)
R= .95014810 R2= .90278142 Upravené R2= .89873064
F(1,24)=222.87 p<.00000 Směrod. chyba odhadu : .20353

	b*	Sm.chyba z b*	b	Sm.chyba z b	t(24)	p-hodn.
N=26						
Abs.člen			-2.61113	0.619583	-4.21434	0.000306
KLEM	0.950148	0.063646	0.91340	0.061184	14.92871	0.000000

$$y' = a + bx$$

Parametr **a** (průsečík regresní přímky s osou Y) je označen **modře**

Parametr **b** (směrnice regresní přímky) je označen **červeně**

Regresní rovnice má v tomto případě následující tvar:

$$T(\text{Nová Ves}) = -2,61113 + 0,91340 T(\text{Klementinum})$$

V záhlaví tabulky opět najdeme korelační koeficient (R), koeficient determinace (R2) i p-hodnotu určující statistickou významnost vztahu.

Ověření vhodnosti modelu

Testujeme, zda se variabilita vysvětlená modelem statisticky významně liší od reziduální (nevysvětlené) variability.

1. Přepneme zpět na výsledky vícenásobné lineární regrese a zvolíme záložku **Detailní výsledky**
2. Zvolíme **ANOVA – celková vhodnost modelu**

Výsledky - vícenásobná regrese: data_10_moje

Výsledky- vícerozm. regrese

Záv.prom. : NVES vícenás. R = .95014810 F = 222.8664
R2 = .90278142 sv = 1,24

Poč. případů: 26 upravené R2 = .89873064 p = .000000
Směrodatná chyba odhadu : .203534643

Abs. člen: -2.611134148 Sm. chyba: .6195834 t(24) = -4.214 p = .0003

KLEM b* = .950

(významná b* jsou zvýrazněna červeně)

Alfa pro zvýraznění efektů: .05

Základní výsledky **Detailní výsledky** Residua/předpoklady/předpovědi

Výpočet: výsledky regrese Parciální korelace

ANOVA (Celk. vhodnost modelu) Redundance

Kovariance koeficientů Výsledky krokové regrese

Současná sweep matice ANOVA upravená pro průměry

OK

Storno

Možnosti

Anal. Skup.

Ověření vhodnosti modelu

3. Interpretujeme výslednou tabulku.

variabilita vysvětlená regresním modelem

zbytková variabilita (nevysvětlená regresním modelem)

celková variabilita

testové kritérium

p-hodnota

Analýza rozptylu (data_10_moje)					
Efekt	Součet čtverců	sv	Průměr čtverců	F	p-hodn.
Regres.	9.23254	1	9.232541	222.8664	0.000000
Rezid.	0.99423	24	0.041426		
Celk.	10.22677				

Nulová hypotéza: mezi reziduální variabilitou a variabilitou vysvětlenou regresním modelem není statisticky významný rozdíl.

Hladina významnosti: 0,05

$p > 0,05$ → nulová hypotéza platí, **model není vhodný**

$p < 0,05$ → zamítáme nulovou hypotézu, existuje statisticky významný rozdíl, **model je vhodný**

Ukázka interpretace: Mezi variabilitou vysvětlenou modelem a reziduální variabilitou **je/není** na hladině významnosti 0,05 statisticky významný rozdíl, použití zvoleného regresního modelu tedy **je/není** vhodné.

Analýza reziduí

Ověří, zda je model regresní přímky vhodný. Rezidua by měla náhodně kolísat kolem nuly (neměla by například růst nebo klesat)

1. Přepneme zpět na výsledky vícenásobné lineární regrese a zvolíme záložku **Residua/ Předpoklady/ Předpovědi**
2. Zvolíme **Reziduální analýza**

Výsledky - vícenásobná regrese: data_10_moje

Výsledky- vícerozm. regrese

Záv.prom. : NVES vícenás. R = .95014810 F = 222.8664
R2 = .90278142 sv = 1,24

Poč. případů: 26 upravené R2 = .89873064 p = .000000
Směrodatná chyba odhadu : .203534643

Abs. člen: -2.611134148 Sm. chyba: .6195834 t(24) = -4.214 p = .0003

KLEM b* = .950

(významná b* jsou zvýrazněna červeně)

Alfa pro zvýraznění efektů: .05

Základní výsledky | Detailní výsledky | **Residua/předpoklady/předpovědi**

Reziduální analýza

Popisné statistiky

Generátor kódu

Předpovědi

Předpověď závislé proměnné

Výpočet interv. spolehlivosti Alfa: .05

Výpočet interv. předpovědi

OK

Storno

Možnosti

Anal.Skup.

Analýza reziduí

3. V okně reziduální analýzy přepneme na záložku **Bodové grafy** a vybereme **Předpovědi vs. rezidua**

Reziduální analýza: data_10_moje

Záv.prom. : NVES vícenás. R = .95014810 F = 222.8664
R2: .90278142 sv = 1,24
Poč. případů: 26 upravené R2: .89873064 p = .000000
Směrodatná chyba odhadu : .203534643
Abs. člen: -2.611134148 Sm. chyba: .6195834 t(24) = -4.214 p < .0003

Pravděp. grafy Odlehle hodnoty Uložit

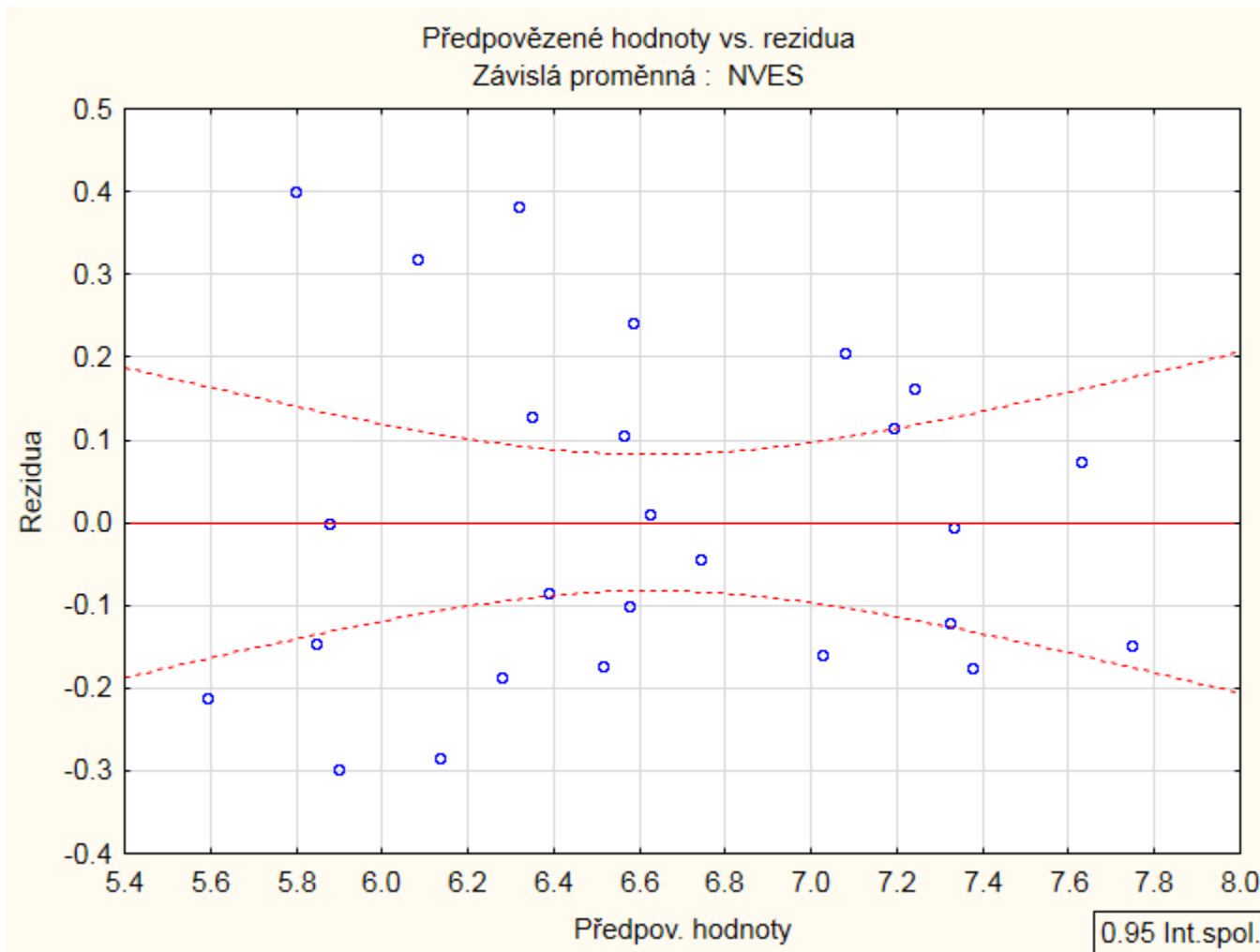
Základ Detaily Rezidua Předpovědi **Bodové grafy**

Předpovědi vs. rezidua Pozorované vs. rezidua²
Předpovědi vs. rezidua² Rezidua vs. odstr. rezidua
Předpovědi vs. pozorované Korelace 2 proměnných
Pozorované vs. rezidua Parciální rezidua

Výpočet
Storno
Možnosti
Anal.Skup.

Analýza reziduí

4. Interpretujeme graf (Kolísají rezidua okolo 0? Pokud ano, vybrali jsme vhodný model).



Graf regresního modelu

Vykreslí regresní přímku i s pásy spolehlivosti, které udávají, kde se bude s 95% pravděpodobností regresní přímka nacházet

1. Přepneme zpět na výsledky reziduální analýzy, zůstaneme na kartě **Bodové grafy**
2. Zvolíme **Korelace 2 proměnných**

Reziduální analýza: data_10_moje

Záv.prom. : NVES vícenás. R = .95014810 F = 222.8664
R2: .90278142 sv = 1,24
Poč. případů: 26 upravené R2: .89873064 p = .000000
Směrodatná chyba odhadu : .203534643
Abs. člen: -2.611134148 Sm. chyba: .6195834 t(24) = -4.214 p < .0003

Pravděp. grafy		Odehlé hodnoty		Uložit	
Základ	Detaily	Rezidua	Předpovědi	Bodové grafy	
<input type="checkbox"/> Předpovědi vs. rezidua	<input type="checkbox"/> Pozorované vs. rezidua ²				
<input type="checkbox"/> Předpovědi vs. rezidua ²	<input type="checkbox"/> Rezidua vs. odstr. rezidua				
<input type="checkbox"/> Předpovědi vs. pozorované	<input checked="" type="checkbox"/> Korelace 2 proměnných				
<input type="checkbox"/> Pozorované vs. rezidua	<input type="checkbox"/> Parciální rezidua				

Výpočet
Stomo
Možnosti
Anal.Skup.

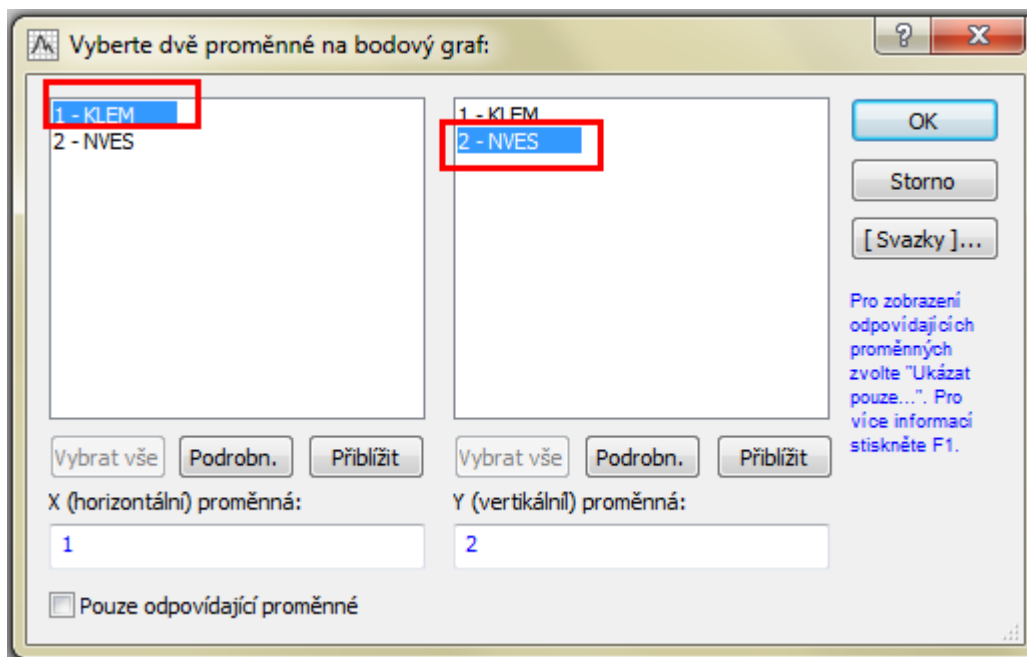
Graf regresního modelu

3. Zvolíme proměnné

X = nezávislá proměnná (vaše stanice)

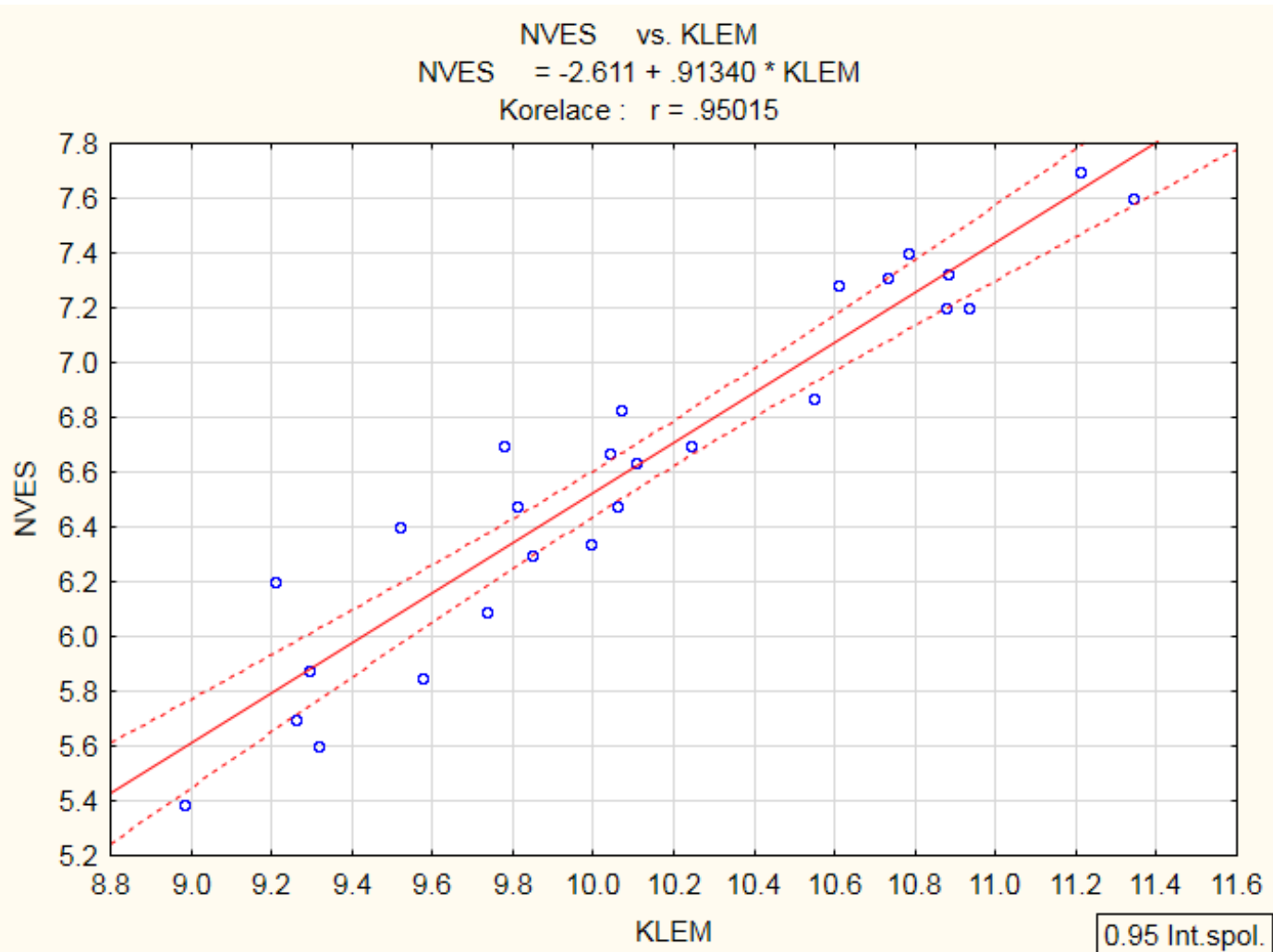
Y = závislá proměnná (Nová Ves)

Potvrdíme OK



Graf regresního modelu

4. Interpretujeme graf (zejména si všímáme směru regresní přímky – kladná, nebo záporná závislost proměnných – a šířky pásů spolehlivosti)



Odhad chybějících měření

Pomocí regresní rovnice a hodnot z vaší stanice je možné vypočítat chybějící data z Nové Vsi ručně, ale lze to udělat i ve STATISTICE:

1. Vrátime se zpět k **výsledkům vícenásobné regrese** (přepneme na okno reziduální analýzy, zadáme Storno)
2. Na kartě **Rezidua/předpoklady/předpovědi** vybereme **Předpověď závislé proměnné**

The screenshot shows the 'Výsledky - vícenásobná regrese: data_10_moje' window. The main area displays regression statistics:

Záv.prom. : NVES	vícenás. R = .95014810	F = 222.8664
Poč. případů: 26	R2 = .90278142	sv = 1,24
Abs.člen: -2.611134148	upravené R2 = .89873064	p = .000000
	Směrodatná chyba odhadu : .203534643	
	Sm. chyba: .6195834	t(24) = -4.214 p = .0003

Below the statistics, it indicates 'KLEM b+ = .950' and '(významná b+ jsou zvýrazněna červeně)'. The bottom panel shows the 'Alfa pro zvýraznění efektů: .05' and three tabs: 'Základní výsledky', 'Detailní výsledky', and 'Rezidua/předpoklady/předpovědi'. The 'Rezidua/předpoklady/předpovědi' tab is active, showing options for 'Předpovědi' with 'Předpověď závislé proměnné' selected. Other options include 'Výpočet interv. spolehlivosti' and 'Výpočet interv. předpovědi'.

Odhad chybějících měření

3. Opíšeme teplotu vzduchu z vaší stanice pro první rok s chybějící teplotou vzduchu z Nové Vsi (1961) a potvrdíme OK

The screenshot shows a data entry application window titled "Data: data_10_moje* (2s krát 30ř)". The main window contains a table with columns "ROK", "KLEM", and "NVES". The data is as follows:

ROK	KLEM	NVES
1961	10.575	
1962	9	
1963	9.20833333	6.2
1975	10.7333333	7.30833333
1976	10.2333333	
1977	10.2416667	6.7
1978	9.575	5.85
1979	9.85	6.3
1980	8.98333333	5.38333333
1981	10.1083333	6.63333333
1982	10.55	6.86666667
1983	10.9333333	7.2
1984	9.73333333	6.09166667
1985	9.31666667	5.6

Overlaid on the table is a dialog box titled "Zadejte hodnoty nezávislých prom." (Enter values of independent variables). The dialog box has a text input field for "KLEM" with the value "10.575". It also has a "Společná hod." (Common value) field with the value "0". The dialog box has "OK", "Storno", and "Použit" buttons.

Odhad chybějících měření

4. Z výsledné tabulky si zapíšeme teplotu vzduchu včetně intervalů spolehlivosti

(data_10_moje)

Předpovězené hodnoty (data_10_moje) proměnné: NVES			
Proměnná	b-váha	Hodnota	b-váha * Hodnot
KLEM	0.913405	10.57500	9.65926
Abs. člen			-2.61113
Předpověď			7.04812
-95.0%LS			6.94662
+95.0%LS			7.14962

odhad teploty
vzduchu v Nové
Vsi

intervaly spolehlivosti (s 95%
pravděpodobností se v nich
bude nacházet reálná hodnota)

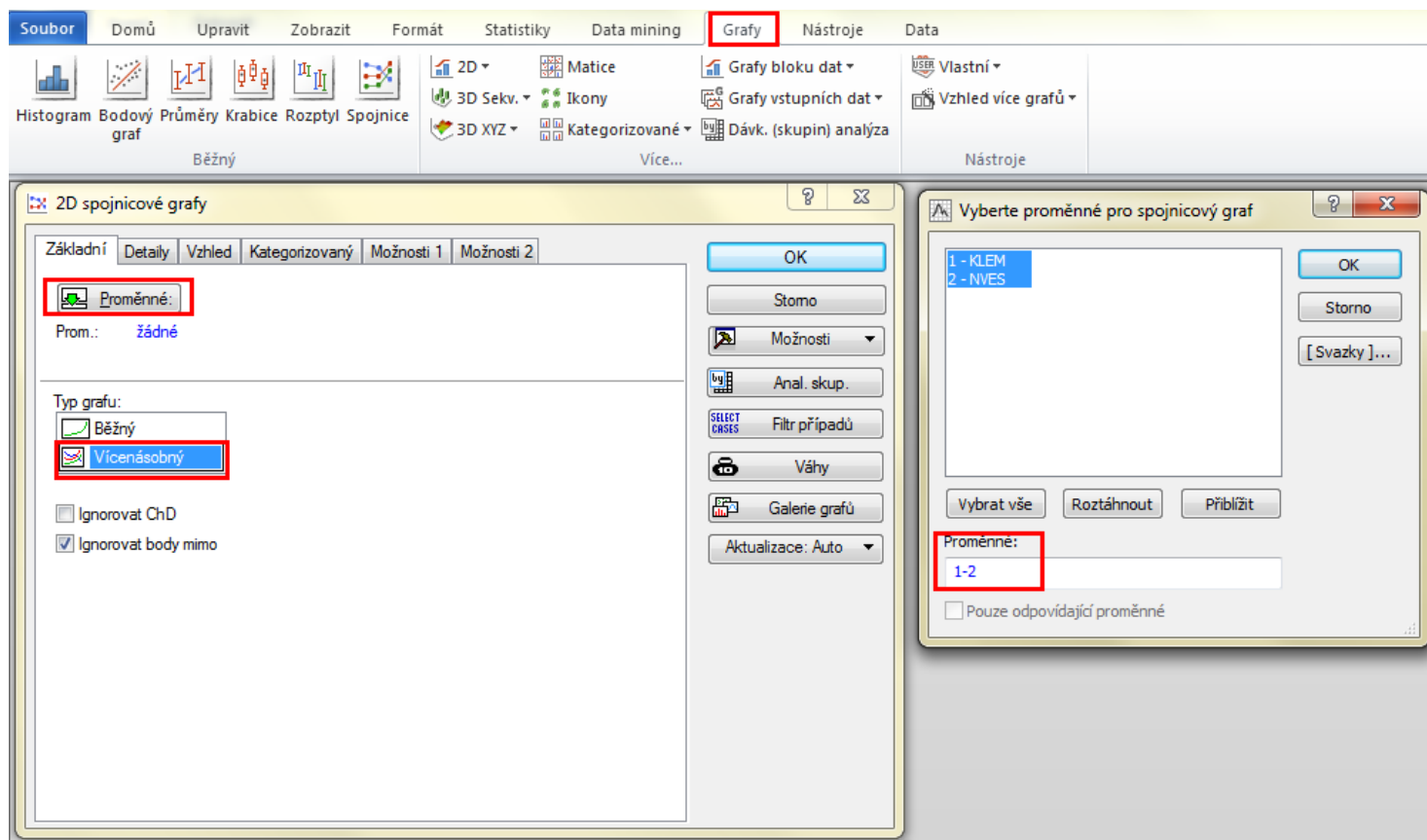
5. Postup opakujeme pro všechny ostatní chybějící hodnoty

6. Do výsledné tabulky zapíšeme vždy rok, teplotu vzduchu z vaší stanice, bodový odhad teploty vzduchu v Nové Vsi a intervaly spolehlivosti

Tip: chcete-li, můžete u prvního odhadu porovnat šířku vašich intervalů spolehlivosti s Klementinem.

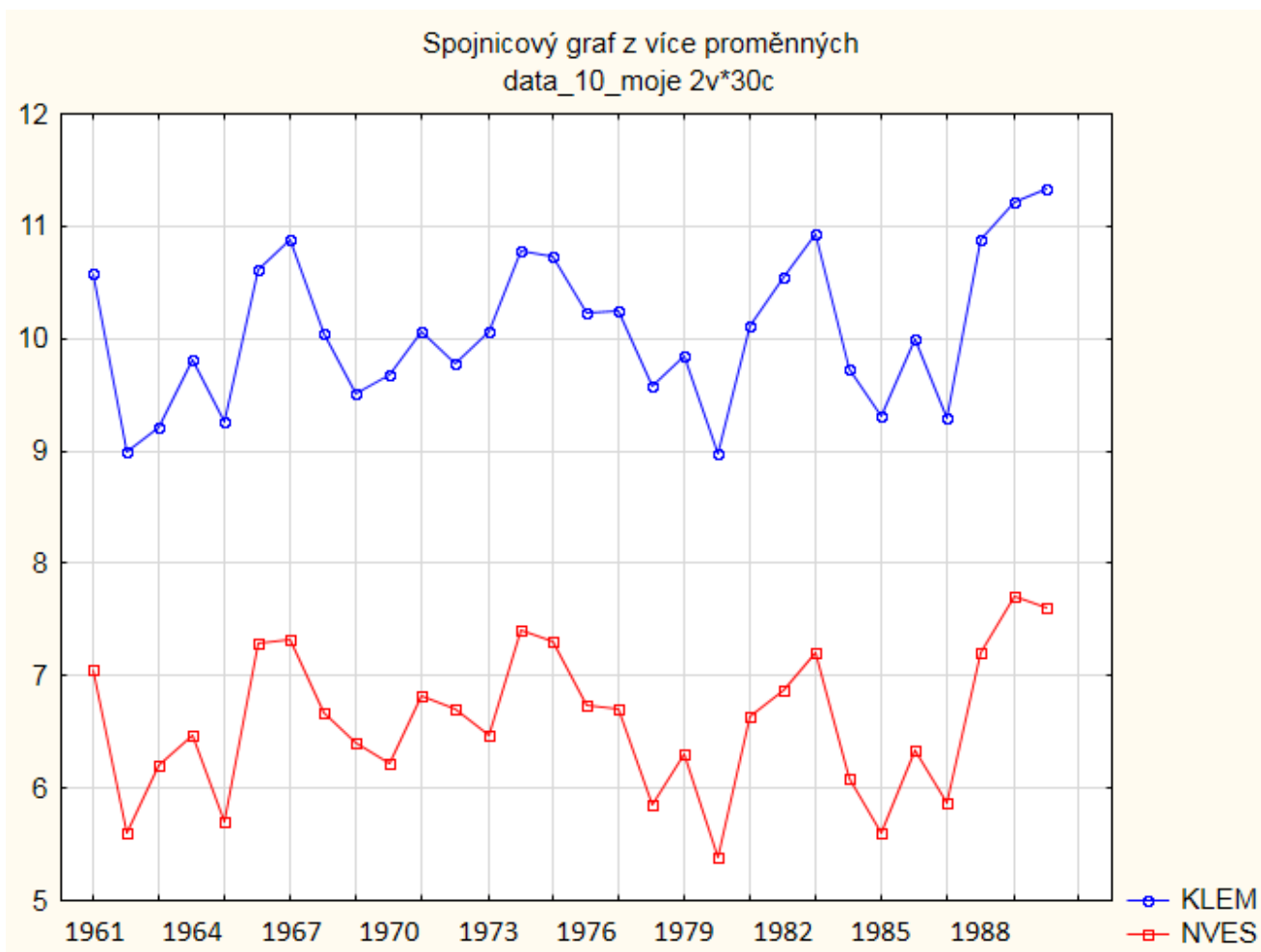
Spojnicový graf

1. Do datového souboru ve STATISTICE doplníme na základě výpočtů chybějící hodnoty teploty vzduchu z Nové Vsi.
 - Na záložce Grafy vybereme Spojnice, následně vybereme vícenásobný typ grafu a obě proměnné a potvrdíme (OK – OK)



Spojnicový graf

2. Graf můžeme upravit (barva spojnic, popis os, legenda...) a popíšeme



Do protokolu – 1. úkol

1. Zadání
2. Tabulka vstupních dat (průměrné roční teploty vzduchu na vaší stanici a v Nové Vsi)
3. Bodový graf s výsledky korelace (r_{xy} , p-hodnota) + interpretace
4. Tabulka s výsledky regresního modelu (vícenásobná regrese) včetně koeficientu determinace a směrodatné chyby odhadu, můžete porovnat vaše výsledky s výsledky z Klementina
5. Regresní rovnice (tabulka z výsledků vícenásobné regrese + vypsání rovnic)
6. Ověření vhodnosti modelu (tabulka ANOVA + interpretace p-hodnoty)
7. Analýza reziduí – graf + jeho popis
8. Graf regresního modelu + jeho popis
9. Tabulka s doplněním chybějících hodnot
10. Spojnicový graf s teplotou vzduchu v Nové Vsi a na vaší stanici + popis
11. Závěr

2. úkol

NEPOVINNÝ

Pokud ho uděláte a budete mít v zápočtovém testu známku na hraně, prof. Dobrovolný se přikloní k té lepší :)

Povinný úkol (bez plus u prof. Dobrovolného) pro všechny, kdo mají **3 absence** a nechtějí opakovat předmět

V souboru **Uzemni_srovnani.xls** najdete údaje porovnávající kraje ČR:

1. Zvolíme ukazatel, který chceme sledovat (**proměnná y**)
2. Ve stejné tabulce najdeme jednu **vysvětlující proměnnou x** , u níž lze **logicky předpokládat, že na ní bude sledovaná proměnná závislá**
3. Sestavíme **lineární regresní model**, který zhodnotíme

2. úkol – regresní model

Postupujeme podobně jako v případě 1. úkolu:

1. Posoudíme **normalitu rozdělení** (v tomto případě nemusíte dělat, ale před výpočtem regresního modelu by se normalita ověřovat měla)
2. Zhodnotíme **korelační vztah** vybraných proměnných pomocí **korelačního pole** (bodový graf), **koeficientu korelace r** a jeho **statistické významnosti** (p-hodnota)
3. Graf **charakterizujeme**, můžeme se i rozhodnout **vyločit některé případy** (kraje), ale vždy musíme rozhodnutí logicky zdůvodnit
4. Vytvoříme **regresní model**, zhodnotíme **koeficient determinace** a **směrodatnou chybu odhadu**, uvedeme i **regresní rovnici**
5. Ověříme **vhodnost modelu** pomocí analýzy ANOVA
6. Provedeme **analýzu reziduí** (graf Předpovědi vs. Rezidua)
7. Vykreslíme **graf regresního modelu** i s **pásky spolehlivosti**
8. V **závěru** zhodnotíme vztah vybraných proměnných

Do protokolu – 2. úkol

1. Zadání
2. Tabulka: zvolené proměnné a vstupní hodnoty + zdůvodnění výběru vysvětlující proměnné
3. Bodový graf s výsledky korelace (r_{xy} , p-hodnota) + interpretace
4. Vyloučené případy a jejich logické zdůvodnění; případně konstatování, že není nutné vyloučit žádný z případů
5. Tabulka s výsledky regresního modelu (vícenásobná regrese) včetně koeficientu determinace a směrodatné chyby odhadu
6. Regresní rovnice (tabulka z výsledků vícenásobné regrese + vypsání rovnic)
7. Ověření vhodnosti modelu (tabulka ANOVA + interpretace p-hodnoty)
8. Analýza reziduí – graf + jeho popis
9. Graf regresního modelu + jeho popis
10. Závěr

Zdroje

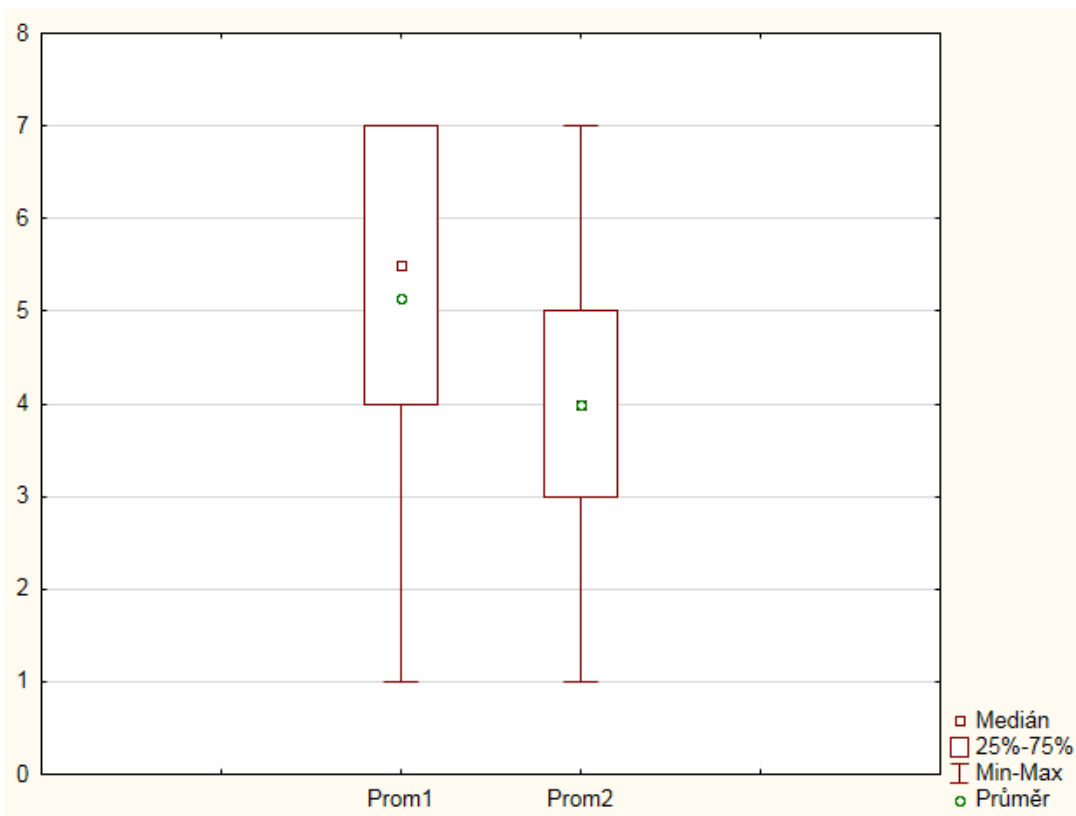
- BUDÍKOVÁ, Marie. Jednoduchá lineární regrese I (přednáška). Brno: Masarykova univerzita, 2016.
- BUDÍKOVÁ, Marie. Jednoduchá lineární regrese II (přednáška). Brno: Masarykova univerzita, 2016.
- DOBROVOLNÝ, Petr. Z1069 Statistické metody a zpracování dat: VII. Regresní počet (přednáška) Brno: Masarykova univerzita, 2016.

Pár slov ke cvičení 8

Jak lze krabicový graf využít k hodnocení normality rozdělení?

Pro ideální normální rozdělení

by byl krabicový graf symetrický podle vodorovné osy procházející středem krabice. Zároveň platí, že se medián rovná průměru a leží rovněž ve středu krabice.



Ukázka krabicového grafu s mediánem a průměrem pro soubor, který nemá normální rozdělení (Prom1) a pro soubor s normálním rozdělením (Prom2).