Jednoduchá lineární regrese

Motivace: Cíl regresní analýzy - popsat závislost hodnot veličiny Y na hodnotách veličiny X.
Nutnost vyřešení dvou problémů:

a) jaký typ funkce se použije k popisu dané závislosti;
b) jak se stanoví konkrétní parametry daného typu funkce?

ad a) Při určení typu funkce je třeba provést teoretický rozbor zkoumané závislosti. Teoretická analýza může upozornit například na to, že
s růstem hodnot veličiny X budou mít hodnoty veličiny Y tendenci monotónně růst či klesat,
tato tendence má charakter zrychlujícího se či zpomalujícího se růstu či poklesu,
jde o závislost, kdy s růstem hodnot veličiny X dochází zpočátku k růstu hodnot veličiny Y, který je po dosažení určitého maxima vystřídán poklesem,
apod.
Můžeme např. zkoumat závislost ceny ojetého auta (veličina Y) na jeho stáří (veličina X). Je zřejmé, že s rostoucím stářím bude klesat cena, ale není jasné, zda lineárně, kvadraticky či dokonce exponenciálně.
Vždy se snažíme o to aby regresní model byl jednoduchý, tj. aby neobsahoval příliš mnoho parametrů. Připadá-li v úvahu více funkcí, posuzujeme jejich vhodnost pomocí různých kritérií – viz dále.
Často však nemáme dostatek informací k provedení teoretického rozboru. Pak se snažíme odhadnout typ funkce pomocí dvourozměrného tečkového diagramu.
Zde se omezíme na funkce, které závisejí lineárně na parametrech $\beta_0, \beta_1, \ldots, \beta_p$.
Zvláštní pozornost budeme věnovat polynomiální funkci 1. stupně $y = \beta_0 + \beta_1 x$.

ad b) Odhady $b_0, b_1, \ldots, b_p$ neznámých parametrů $\beta_0, \beta_1, \ldots, \beta_p$ získáme na základě dvourozměrného datového souboru $\begin{pmatrix} x_1 & y_1 \\ \ldots & \ldots \\ x_n & y_n \end{pmatrix}$ metodou nejmenších čtverců, tj. z podmínky, aby součet čtverců odchylek zjištěných a odhadnutých hodnot byl minimální.

Specifikace klasického modelu lineární regrese

$Y = m(x;\beta_0,\beta_1,\ldots,\beta_p)+\varepsilon$, kde

$m(x;\beta_0,\beta_1,\ldots,\beta_p)$ - teoretická regresní funkce, která lineárně závisí na neznámých regresních parametrech $\beta_0,\beta_1,\ldots,\beta_p$ a známých funkcích $f_1(x),\ldots,f_p(x)$, které již neobsahují neznámé parametry, tj. $m(x;\beta_0,\beta_1,\ldots,\beta_p)=\sum_{j=0}^{p}\beta_j f_j(x)$, přičemž $f_0(x)\equiv 1$.

Jde o deterministickou složku modelu.

Složka ε - náhodná složka modelu. Je to náhodná odchylka od deterministické závislosti Y na X. Popisuje závislost vysvětlované proměnné na neznámých nebo nepozorovaných proměnných a popisuje i vliv náhody. Nelze ji funkčně vyjádřit.

Veličina Y - závisle proměnná (též vysvětlovaná) veličina.

Veličina X - nezávisle proměnná (též vysvětlující) veličina.

Pořídíme n dvojic pozorování $(x_1,y_1),\ldots,(x_n,y_n)$, tj. dvourozměrný datový soubor $\begin{pmatrix} x_1 & y_1 \\ \ldots & \ldots \\ x_n & y_n \end{pmatrix}$.

Pro i = 1, ..., n platí: $y_i = m(x_i;\beta_0,\beta_1,\ldots,\beta_p)+\varepsilon_i$.

O náhodných odchylkách $\varepsilon_1,\ldots,\varepsilon_n$ předpokládáme, že

a) $E(\varepsilon_i)=0$ (odchylky nejsou systematické)
b) $D(\varepsilon_i)=\sigma^2>0$ (všechna pozorování jsou prováděna s touž přesností)
c) $C(\varepsilon_i,\varepsilon_j)=0$ pro $i\neq j$ (mezi náhodnými odchylkami neexistuje žádný lineární vztah)
d) $\varepsilon_i \sim N(0,\sigma^2)$.

V tomto případě hovoříme o klasickém modelu lineární regrese.

Označení

$b_0, b_1, \ldots, b_p$ - odhady regresních parametrů $\beta_0, \beta_1, \ldots, \beta_p$ (nejčastěji je získáme metodou nejmenších čtverců, tj. z podmínky, že výraz

$\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p}\beta_j f_j(x_i)\right)^2$ nabývá svého minima pro $\beta_j = b_j$, j = 0, 1, …, p)

$\hat{m}(x; b_0, \ldots, b_p)$ - empirická regresní funkce

$\hat{y}_i = \hat{m}(x_i; b_0, \ldots, b_p) = \sum_{j=0}^{p} b_j f_j(x_i)$ - regresní odhad i-té hodnoty veličiny Y (i-tá predikovaná hodnota veličiny Y)

$e_i = y_i - \hat{y}_i$ - i-té reziduum

$S_E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ - reziduální součet čtverců

$s^2 = \frac{S_E}{n-p-1}$ - odhad rozptylu σ^2

$S_R = \sum_{i=1}^{n}(\hat{y}_i - m_2)^2$ - regresní součet čtverců $\left(m_2 = \frac{1}{n}\sum_{i=1}^{n} y_i\right)$

$S_T = \sum_{i=1}^{n}(y_i - m_2)^2$ - celkový součet čtverců $(S_T = S_R + S_E)$

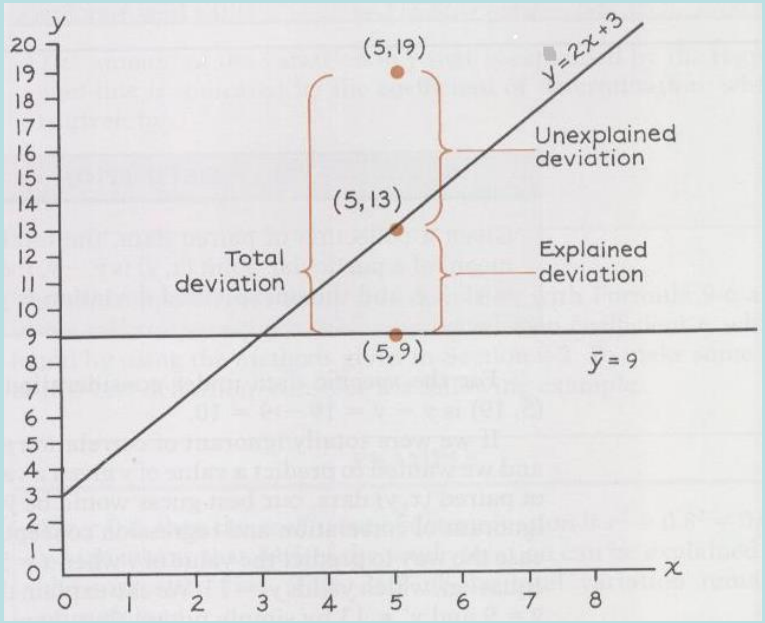
Význam jednotlivých typů součtů čtverců

Předpokládejme, že máme dvourozměrný datový soubor, v němž průměr hodnot závisle proměnné veličiny Y je 9 a závislost veličiny Y na veličině X je popsána regresní přímkou y = 2x + 3. Dvourozměrný tečkový diagram obsahuje bod o souřadnicích (5, 19), který pochází z datového souboru. Na regresní přímce leží bod o souřadnicích (5, 13).

Odchylka zjištěné hodnoty 19 od průměru 9 je v obrázku označena „Total deviation“ a po umocnění je to jedna ze složek celkového součtu čtverců S_T, tj. složka $y_i - m_2$.

Odchylka zjištěné hodnoty 19 od hodnoty 13 na regresní přímce je v obrázku označena „Unexplained deviation“ a po umocnění je to jedna ze složek reziduálního součtu čtverců S_E, tj. složka $y_i - \hat{y}_i$.

Odchylka hodnoty 13 na regresní přímce od průměru 9 je v obrázku označena „Explained deviation“ a po umocnění je to jedna ze složek regresního součtu čtverců S_R, tj. složka $\hat{y}_i - m_2$.



Maticový zápis klasického modelu lineární regrese

$\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$, kde

$\mathbf{y} = (y_1, \ldots, y_n)'$ - vektor pozorování závisle proměnné veličiny Y,

$$\mathbf{X} = \begin{pmatrix} 1 & f_1(x_1) & \ldots & f_p(x_1) \\ \ldots & \ldots & \ldots & \ldots \\ 1 & f_1(x_n) & \ldots & f_p(x_n) \end{pmatrix}$$ - regresní matice

(předpokládáme, že h(**X**) = p+1 < n)

$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ - vektor regresních parametrů,

$\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$ - vektor náhodných odchylek.

Podmínky (a) až (d) lze zkráceně zapsat ve tvaru $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I})$.

Maticově zapsaná metoda nejmenších čtverců vede na rovnice

$\mathbf{X'X\beta} = \mathbf{X'y}$ - systém normálních rovnic

$\mathbf{b} = (\mathbf{X'X})^{-1} \mathbf{X'y}$ – odhad vektoru $\boldsymbol{\beta}$ získaný metodou nejmenších čtverců

$\hat{\mathbf{y}} = \mathbf{Xb}$ – vektor regresních odhadů (vektor predikce)

$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ - vektor reziduí

Vlastnosti odhadu **b**:

- odhad **b** je lineární, neboť je vytvořen lineární kombinací pozorování $y_1, \ldots, y_n$ s maticí vah $(\mathbf{X'X})^{-1}\mathbf{X'}$;
- odhad **b** je nestranný, neboť $E(\mathbf{b}) = \boldsymbol{\beta}$;
- odhad **b** má varianční matici $\text{var}\,\mathbf{b} = \sigma^2(\mathbf{X'X})^{-1}$;
- odhad $\mathbf{b} \sim N_{p+1}(\boldsymbol{\beta}, \sigma^2 (\mathbf{X'X})^{-1})$ vzhledem k platnosti podmínky (d);
- pro odhad **b** platí Gaussova - Markovova věta: Odhad $\mathbf{b} = (\mathbf{X'X})^{-1} \mathbf{X'y}$ je nejlepší nestranný lineární odhad vektoru $\boldsymbol{\beta}$.

Příklad

U šesti obchodníků byla zjišťována poptávka po určitém druhu zboží loni (veličina X - v kusech) a letos (veličina Y - v kusech).

číslo obchodníka	1	2	3	4	5	6
poptávka loni (X)	20	60	70	100	150	260
poptávka letos (Y)	50	60	60	120	230	320

Předpokládejte, že závislost letošní poptávky na loňské lze vystihnout regresní přímkou. Sestavte regresní matici, vypočtěte odhady regresních parametrů a napište rovnici regresní přímky. Interpretujte parametry regresní přímky.

Řešení:

Sestavíme regresní matici.

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \text{ tedy } \mathbf{X} = \begin{pmatrix} 1 & 20 \\ 1 & 60 \\ 1 & 70 \\ 1 & 100 \\ 1 & 150 \\ 1 & 260 \end{pmatrix}.$$

Podle vzorce $\mathbf{b} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$ získáme odhady regresních parametrů.

Nejprve vypočítáme matici $\mathbf{X}'\mathbf{X} = \begin{pmatrix} 6 & 660 \\ 660 & 109000 \end{pmatrix}$ a k ní inverzní matici $(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 0{,}499084 & -0{,}003022 \\ -0{,}003022 & 0{,}000027 \end{pmatrix}$.

Dále získáme součin $\mathbf{X}'\mathbf{y} = \begin{pmatrix} 840 \\ 138500 \end{pmatrix}$ a nakonec vektor odhadů regresních parametrů: $\mathbf{b} = \begin{pmatrix} 0{,}499084 & -0{,}003022 \\ -0{,}003022 & 0{,}000027 \end{pmatrix} \cdot \begin{pmatrix} 840 \\ 138500 \end{pmatrix} = \begin{pmatrix} 0{,}6868 \\ 1{,}2665 \end{pmatrix}$.

Regresní přímka má tedy rovnici

$$y = 0{,}6868 + 1{,}2665\, x.$$

Znamená to, že při nulové loňské poptávce by letošní poptávka činila 0,6868 kusů a při zvýšení loňské poptávky o 10 kusů by se letošní poptávka zvedla o 12,665 kusů.

Výpočet pomocí systému STATISTICA

Vytvoříme nový datový soubor se dvěma proměnnými X a Y a 6 případy:

Statistiky – Vícerozměrná regrese – Závisle proměnná Y, nezávisle proměnná X - OK – OK – Výpočet: Výsledky regrese.

N=6	Výsledky regrese se závislou proměnnou : Y (Tabulka1) R= ,97197702 R2= ,94473932 Upravené R2= ,93092415 F(1,4)=68,384 p<,00117 Směrod. chyba odhadu : 29,219					
	Beta	Sm.chyba beta	B	Sm.chyba B	t(4)	Úroveň p
Abs.člen			0,68681	20,6423	0,03327	0,97505
X	0,97197	0,11753	1,26648	0,1531	8,26947	0,00116

Ve výstupní tabulce najdeme koeficient b_0 ve sloupci B na řádku označeném Abs. člen, koeficient b_1 ve sloupci B na řádku označeném X.

Rovnice regresní přímky:

$y = 0{,}686813 + 1{,}266484\ x$.

Znamená to, že při nulové loňské poptávce by letošní poptávka činila 0,6868 kusů a při zvýšení loňské poptávky o 10 kusů by se letošní poptávka zvedla o 12,665 kusů.

Testování významnosti modelu jako celku (celkový F-test)

Na hladině významnosti α testujeme

H_0: $(\beta_1,\ldots,\beta_p)' = (0,\ldots,0)'$ proti H_1: $(\beta_1,\ldots,\beta_p)' \neq (0,\ldots,0)'$.

(Nulová hypotéza říká, že dostačující je model konstanty.)

Testová statistika: $F = \dfrac{S_R/p}{S_E/(n-p-1)}$ má rozložení F(p, n-p-1), pokud H_0 platí.

Kritický obor: $W = \langle F_{1-\alpha}(p, n-p-1), \infty)$.

$F \in W \Rightarrow H_0$ zamítáme na hladině významnosti α.

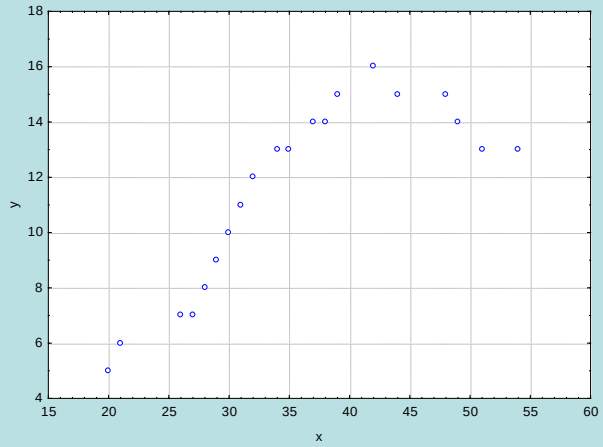
Výsledky F-testu zapisujeme do tabulky analýzy rozptylu:

zdroj variability	součet čtverců	stupně volnosti	podíl	statistika F
model	S_R	p	S_R/p	$\dfrac{S_R/p}{S_E/(n-p-1)}$
reziduální	S_E	n-p-1	$S_E/(n-p-1)$	-
celkový	S_T	n-1	-	-

Příklad:
Majitelé prodejny počítačových her nechali své prodavače absolvovat kurz prodejních dovedností. Poté zjišťovali po dobu 20 dnů, kolik osob navštíví během otevírací doby prodejnu (proměnná X) a jaká je v tento den tržba (proměnná Y, udává se v tisících Kč a je zaokrouhlená).

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
x_i	20	21	2	27	28	29	30	31	32	34	35	37	38	39	42	44	48	49	51	54
y_i	5	6	7	7	8	9	10	11	12	13	13	14	14	15	16	15	15	14	13	13

Dvourozměrný tečkový diagram



Z grafu závislosti Y na X vyplývá, že s rostoucím počtem zákazníků se tržby zvyšují, avšak při denním počtu zákazníků asi 42 dosahují svého maxima a pak už zase klesají (vyšší počet zákazníků obsluha prodejny nezvládá a zákazníci odcházejí, aniž by nakoupili). Zdá se tedy, že vhodným modelem závislosti tržeb na počtu zákazníků bude regresní parabola
$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$.
Odhadněte parametry regresního modelu a proveďte celkový F-test.

Řešení:

Vytvoříme nový datový soubor se třemi proměnnými X, Xkv, Y a o 20 případech. Do proměnných X a Y napíšeme zjištěné hodnoty a do Dlouhého jména proměnné Xkv napíšeme = X^2.

Získání odhadů b_0, b_1, b_2:

Statistiky – Vícerozměrná regrese – Závisle proměnná Y, nezávisle proměnné X, Xkv - OK – OK – Výpočet: Výsledky regrese.

	Výsledky regrese se závislou proměnnou : y (prodejna_software. R= ,95519276 R2= ,91239322 Upravené R2= ,90208653 F(2,17)=88,524 p<,00000 Směrod. chyba odhadu : 1,0623					
N=20	b*	Sm.chyba z b*	b	Sm.chyba z b	t(17)	p-hodn.
Abs.člen			-20,772	3,37325	-6,1579	0,00001
x	4,5264	0,54822	1,565	0,18955	8,2565	0,00000
xkv	-3,7383	0,54822	-0,017	0,00253	-6,8191	0,00000

Regresní parabola má tedy tvar: $y = -20{,}7723 + 1{,}5651x - 0{,}0173x^2$.

Výsledky celkového F-testu jsou uvedeny v záhlaví výstupní tabulky. Testová statistika F nabývá hodnoty 88,524, odpovídající p-hodnota je blízká 0, tedy na hladině významnosti 0,05 zamítáme hypotézu, že dostačující je model konstanty.

Podrobnější výsledky získáme v tabulce analýzy rozptylu:

Aktivujeme Výsledky–vícenásobná regrese – Detailní výsledky – ANOVA

	Analýza rozptylu (prodejna_software.sta)				
Efekt	Součet čtverců	sv	Průměr čtverců	F	p-hodn.
Regres.	199,814	2	99,9070	88,5244	0,00000
Rezid.	19,185	17	1,1285		
Celk.	219,000				

Testování významnosti regresních parametrů (dílčí t-testy)

Na hladině významnosti α pro j = 0,1, ..., p testujeme hypotézu
H_0: $\beta_j = 0$ proti H_1: $\beta j \neq 0$.

Testová statistika: $T_j = \frac{b_j}{s_{b_j}}$ má rozložení t(n-p-1), pokud H_0 platí.

Kritický obor: $W = \left(-\infty, -t_{1-\alpha/2}(n-p-1)\right\rangle \cup \left\langle t_{1-\alpha/2}(n-p-1), \infty\right)$.

$T_j \in W \Rightarrow H_0$ zamítáme na hladině významnosti α.

Příklad:

V předešlém příkladě, kde byla modelována závislost tržby na počtu zákazníků regresní parabolou, proveďte dílčí t-testy o nevýznamnosti jednotlivých regresních parametrů

Řešení:

Stačí interpretovat výstupní tabulku vícenásobné regrese:

Výsledky regrese se závislou proměnnou : y (prodejna_software.
R= ,95519276 R2= ,91239322 Upravené R2= ,90208653
F(2,17)=88,524 p<,00000 Směrod. chyba odhadu : 1,0623

N=20	b*	Sm.chyba z b*	b	Sm.chyba z b	t(17)	p-hodn.
Abs.člen			-20,772	3,37325	-6,1579	0,00001
x	4,5264	0,54822	1,565	0,18955	8,2565	0,00000
xkv	-3,7383	0,54822	-0,017	0,00253	-6,8191	0,00000

Sloupec označený t(17) obsahuje realizace testových statistik a sloupec p-hodn. pak odpovídající p-hodnoty. Ve všech třech případech jsou p-hodnoty menší než 0,05, tedy na hladině významnosti 0,05 zamítáme hypotézy o nevýznamnosti regresních parametrů β_0, β_1, β_2.

Kritéria pro posouzení vhodnosti zvolené regresní funkce

a) Index determinace

$ID^2 = \frac{S_R}{S_T} = 1 - \frac{S_E}{S_T}$ - index determinace $(0 \le ID^2 \le 1)$

- udává, jakou část variability závisle proměnné veličiny Y lze vysvětlit zvolenou regresní funkcí (často se udává v %);
- je zároveň mírou těsnosti závislosti proměnné Y na proměnné X;
- je to obecná míra, nezávislá na typu regresní funkce (lze použít i pro měření nelineární závislosti);
- je to míra, která nebere v úvahu počet parametrů regresní funkce. U regresních funkcí s více parametry vychází tedy obvykle vyšší než u regresních funkcí s méně parametry;
- tato míra není symetrická.

Za vhodnější se považuje ta regresní funkce, pro niž je index determinace vyšší. V případě, že porovnáváme několik modelů s rozdílným počtem parametrů, používáme adjustovaný index determinace:

$ID_{adj}^2 = ID^2 - \frac{(1 - ID^2)p}{n - p - 1}$ - adjustovaný index determinace

V příkladu s prodejem software najdeme index determinace ve výstupní tabulce regrese:

Výsledky regrese se závislou proměnnou : y (prodejna_software.
R= ,95519276 R2= ,91239322 Upravené R2= ,90208653
F(2,17)=88,524 p<,00000 Směrod. chyba odhadu : 1,0623

N=20	b*	Sm.chyba z b*	b	Sm.chyba z b	t(17)	p-hodn.
Abs.člen			-20,772	3,37325	-6,1579	0,00001
x	4,5264	0,54822	1,565	0,18955	8,2565	0,00000
xkv	-3,7383	0,54822	-0,017	0,00253	-6,8191	0,00000

Index determinace je zde označen jako R2, nabývá hodnoty 0,9124 a říká nám, že 91,24% variability tržeb je vysvětleno regresní parabolou. Adjustovaný index determinace je označen Upravené R2.

b) Testové kritérium F

Za vhodnější je považována ta regresní funkce, u níž je hodnota testové statistiky $F = \frac{S_R / p}{S_E / (n - p - 1)}$ pro test významnosti modelu jako celku vyšší.

Ve výstupní tabulce regrese je testová statistika F uvedena v záhlaví:

Výsledky regrese se závislou proměnnou : y (prodejna_software.
R= ,95519276 R2= ,91239322 Upravené R2= ,90208653
F(2,17)=88,524 p<,00000 Směrod. chyba odhadu : 1,0623

N=20	b*	Sm.chyba z b*	b	Sm.chyba z b	t(17)	p-hodn.
Abs.člen			-20,772	3,37325	-6,1579	0,00001
x	4,5264	0,54822	1,565	0,18955	8,2565	0,00000
xkv	-3,7383	0,54822	-0,017	0,00253	-6,8191	0,00000

V našem příkladě je označena F(2,17) a nabývá hodnoty 88,524.

c) Reziduální součet čtverců a reziduální rozptyl

Reziduální součet čtverců: $S_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

Za vhodnější považujeme funkci, která má reziduální součet čtverců nižší. Reziduální součet čtverců lze použít pouze tehdy, když srovnáváme funkce se stejným počtem parametrů.

Reziduální rozptyl: $s^2 = \frac{S_E}{n - p - 1}$

Za vhodnější považujeme tu funkci, která má reziduální rozptyl nižší. Reziduální rozptyl můžeme použít vždy, bez ohledu na to, kolik parametrů mají srovnávané regresní funkce.

Obě charakteristiky najdeme v tabulce ANOVA:

	Analýza rozptylu (prodejna_software.sta)				
Efekt	Součet čtverců	sv	Průměr čtverců	F	p-hodn.
Regres.	199,814	2	99,9070	88,5244	0,00000
Rezid.	19,185	17	1,1285		
Celk.	219,000				

Reziduální součet čtverců je 19,1859 a reziduální rozptyl je 1,12858.

d) Střední absolutní procentuální chyba predikce (MAPE)

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

Za vhodnější považujeme tu funkci, která má MAPE nižší.

Systém STATISTICA MAPE neposkytuje, tuto chybu musíme vypočítat.

Statistiky – Vícerozměrná regrese – Závisle proměnná Y, nezávisle proměnné x, xkv - OK – OK – zvolíme Rezidua/předpoklady/předpovědi – Reziduální analýza – Uložit – Uložit rezidua & předpovědi – vybereme proměnnou y - OK.
K vzniklému datovému souboru přidáme jednu novou proměnnou, nazveme ji chyba a do jejího Dlouhého jména napíšeme =100*abs((v1-v2)/v1)
Pomocí Statistiky – Základní statistiky/tabulky – Popisné statistiky zjistíme průměr proměnné chyba. V našem případě je MAPE 9,31%.

e) Analýza reziduí

Rezidua považujeme za odhady náhodných odchylek a klademe na ně stejné požadavky jako na náhodné odchylky, tj.
mají být nezávislá,
mají být normálně rozložená,
mají mít nulovou střední hodnotu,
mají mít konstantní rozptyl (tj. jsou homoskedastická).

Nezávislost reziduí (autokorelaci) posuzujeme např. pomocí Durbinovy – Watsonovy statistiky, která by se měla nacházet v intervalu $\langle 1{,}4;2{,}6 \rangle$ (to je ovšem pouze orientační vodítko, korektní postup spočívá v porovnání této statistiky s tabelovanou kritickou hodnotou).

Normalitu reziduí ověřujeme pomocí testů normality (např. Lilieforsovou variantou Kolmogorovova – Smirnovova testu nebo Shapirovým – Wilksovým testem) či graficky pomocí N-P plotu.

Testování nulovosti střední hodnoty reziduí provádíme pomocí jednovýběrového t-testu.

Homoskedasticitu reziduí posuzujeme pomocí grafu závislosti reziduí na predikovaných hodnotách. V tomto grafu by rezidua měla být rovnoměrně rozptýlena.

Příklad: Proveďte analýzu reziduí pro příklad s modelováním závislosti tržby na počtu zákazníků.

Posouzení nezávislosti reziduí pomocí Durbinovy – Watsonovy statistiky:

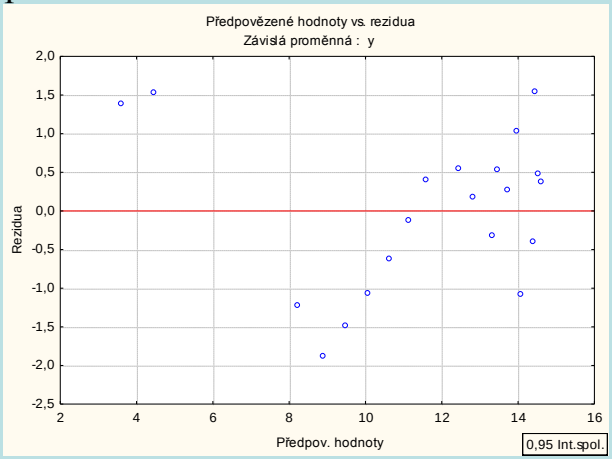
Statistiky – Vícenásobná regrese – proměnná Závislá: y, nezávislá x, xkv – OK – na záložce Residua/předpoklady/předpovědi vybereme Reziduální analýza - Detaily – Durbin-Watsonova statistika:

	Durbin-Watson.d	Sériové korelace
Odhad	0,70250	0,59924

Hodnota této statistiky je nízká, svědčí o tom, že rezidua jsou kladně korelovaná.

Posouzení homoskedasticity reziduí

Reziduální analýza – Bodové grafy – Předpovědi vs. rezidua



Je vidět, že rezidua nejsou kolem 0 rozmístěna náhodně. Model s regresní parabolou tedy není úplně vhodný.

Testování nulovosti střední hodnoty reziduí:

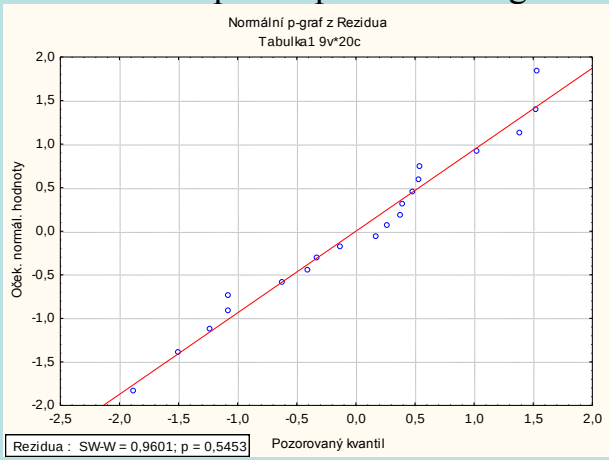
Pro proměnnou Rezidua z tabulky uložené pomocí Reziduální analýzy provedeme jednovýběrový t-test: Statistiky - Základní statistiky/tabulky – t-test, samost. vzorek – OK – proměnné Rezidua – OK.

Proměnná	Průměr	Sm.odch.	N	Sm.chyba	Referenční konstanta	t	SV	p
Rezidua	-0,00000	1,00488	20	0,22469	0,00	-0,00000	19	1,00000

Na hladině významnosti 0,05 nezamítáme hypotézu, že střední hodnota reziduí je 0.

Posouzení normality reziduí:

Na záložce Pravděpodobnostní grafy zvolíme Normální pravděpodobnostní graf reziduí:



Rezidua se řadí kolem ideální přímky, lze tedy soudit, že se řídí normálním rozložením.

Závěr: V neprospěch regresní paraboly hovoří hodnota Durbinovy – Watsonovy statistiky a graf závislosti reziduí na predikovaných hodnotách.

Popis časových řad

Pojem časové řady: Časovou řadou rozumíme řadu hodnot $y_{t_1},\ldots,y_{t_n}$ určitého ukazatele uspořádanou podle přirozené časové posloupnosti $t_1 < \ldots < t_n$. Jsou-li časové intervaly (t_1, t_2), ..., (t_{n-1}, t_n) stejně dlouhé (ekvidistantní), zjednodušeně zapisujeme časovou řadu jako y_1, ..., y_n. Přitom ukazatel je veličina, která charakterizuje nějaký jev v určitém prostoru a určitém čase (okamžiku či intervalu).

Druhy časových řad

a) Časová řada okamžiková: příslušný ukazatel udává, kolik jevů existuje v daném časovém okamžiku (např. počet obyvatelstva k určitému dnu).

b) Časová řada intervalová: příslušný ukazatel udává, kolik jevů vzniklo či zaniklo v určitém časovém intervalu (např. počet sňatků během roku). Nejsou-li jednotlivé časové intervaly ekvidistantní, musíme provést očištění časové řady od důsledků kalendářních variací.

Příklad: Máme k dispozici údaje o tržbě obchodní organizace (v tis. Kč) v jednotlivých měsících roku 1995: 2400, 2134, 2407, 2445, 2894, 3354, 3515, 3515, 3225, 3063, 2694, 2600. Vypočtěte očištěné údaje.

Řešení: Průměrná délka měsíce je 365/12 dne. Očištěná hodnota

pro leden $y_1^{(o)} = 2400 \cdot \dfrac{365}{12 \cdot 31} = 2354{,}84$,

pro únor $y_2^{(o)} = 2134 \cdot \dfrac{365}{12 \cdot 28} = 2318{,}18$.

Pro ostatní měsíce analogicky dostaneme

2361,71; 2478,96; 2839,54; 3400,58, 3448,86; 3448,86; 3269,79; 3005,36; 2731,42; 2551,08.

Výpočet pomocí systému STATISTICA:

Vytvoříme nový datový soubor o třech proměnných: trzba, dm (délky jednotlivých měsíců) a ot (očištěná tržba) a 12 případech. Do proměnné trzba zapíšeme zjištěné hodnoty. Do proměnné dm vložíme délky jednotlivých měsíců, tj. 31, 28, 30, …, 31. Do Dlouhého jména proměnné ot napíšeme =trzba*365/(12*dm).

	1 trzba	2 dm	3 ot
1	2400	31	2354,83
2	2134	28	2318,18
3	2407	31	2361,70
4	2445	30	2478,95
5	2894	31	2839,54
6	3354	30	3400,58
7	3515	31	3448,85
8	3515	31	3448,85
9	3225	30	3269,79
10	3063	31	3005,36
11	2694	30	2731,41
12	2600	31	2551,07

Grafické znázornění okamžikové časové řady

Použijeme spojnicový diagram. Na vodorovnou osu vynášíme časové okamžiky $t_1, ..., t_n$, na svislou osu odpovídající hodnoty $y_1, ..., y_n$. Dvojice bodů (t_i, y_i), $i = 1, ..., n$ spojíme úsečkami.

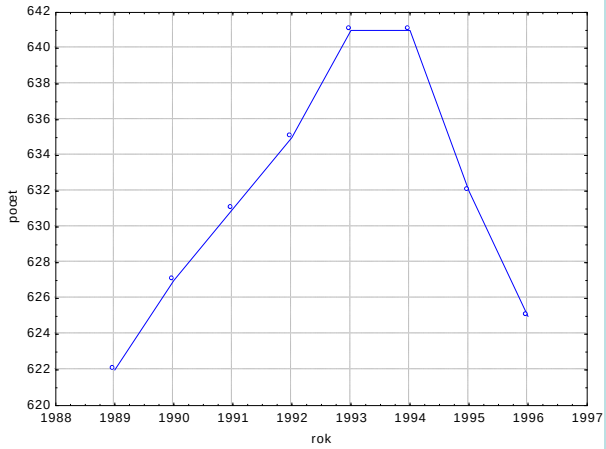
Příklad: Časová řada obsahuje údaje o počtu zaměstnanců určité akciové společnosti v letech 1989 – 1996 vždy k 31.12.

1989	1990	1991	1992	1993	1994	1995	1996
622	627	631	635	641	641	632	625

Znázorněte tuto časovou řadu graficky.

Řešení pomocí systému STATISTICA:

Vytvoříme datový soubor o dvou proměnných nazvaných rok a pocet a 8 případech.
Grafy – Bodové grafy – odškrtneme Lineární proložení – Proměnné X – rok, Y – počet – OK – OK. 2x klikneme na pozadí grafu – vybereme Graf: obecné – zaškrtneme Spojnice – OK.



Grafické znázornění intervalové časové řady

Použijeme sloupkový diagram. Je to soustava obdélníků, kde šířka obdélníku je rovna délce intervalu a výška odpovídá hodnotě ukazatele v daném intervalu. Ke znázornění intervalové časové řady lze použít i spojnicový diagram, přičemž na vodorovnou osu vynášíme středy příslušných intervalů.

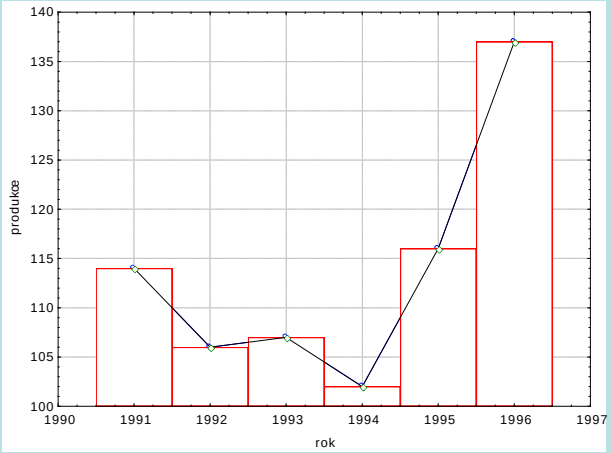
Příklad: Máme k dispozici údaje o produkci určitého podniku (v tisících výrobků) v letech 1991-1996.

1991	1992	1993	1994	1995	1996
114	106	107	102	116	137

Znázorněte tuto časovou řadu graficky.

Řešení pomocí systému STATISTICA:

Vytvoříme datový soubor o dvou proměnných nazvaných rok a produkce a 6 případech.
Grafy – Bodové grafy – odškrtneme Lineární proložení – Proměnné X – rok, Y – produkce – OK – OK. 2x klikneme na pozadí grafu – vybereme Graf: obecné – zaškrtneme Spojnice – Přidat nový graf – typ Sloupcový graf – OK. Do sloupců označených jako Nový1, Nový2 okopírujeme hodnoty proměnných rok a produkce. Ve Všech možnostech: Sloupce upravíme šířku sloupce na 1.



Průměr okamžikové časové řady

Nejprve vypočteme průměry pro jednotlivé dílčí intervaly (t_1, t_2), (t_2, t_3), ..., (t_{n-1}, t_n): $\frac{y_1 + y_2}{2}, \frac{y_2 + y_3}{2}, \ldots, \frac{y_{n-1} + y_n}{2}$. Jsou-li všechny tyto intervaly stejně dlouhé, vypočteme prostý chronologický průměr okamžikové časové řady:

$$\bar{y} = \frac{1}{n-1} \sum_{i=2}^{n} \frac{y_{i-1} + y_i}{2} = \frac{1}{n-1} \left(\frac{y_1}{2} + \sum_{i=2}^{n-1} y_i + \frac{y_n}{2} \right).$$

Nemají-li intervaly stejnou délku, vypočteme $d_i = t_i - t_{i-1}$, $i = 2, \ldots, n$ a použijeme vážený chronologický průměr okamžikové časové řady:

$$\bar{y} = \frac{1}{\sum_{i=2}^{n} d_i} \sum_{i=2}^{n} \frac{y_{i-1} + y_i}{2} \cdot d_i .$$

Příklad: Časová řada vyjadřuje počet obyvatelstva ČSSR (v tisících) v letech 1965 až 1974 vždy ke dni 31.12.

Rok	1965	1966	1967	1968	1969	1970	1971	1972	1973	1974
počet	14194	14271	14333	14387	14443	14345	14419	14576	14631	14738

Charakterizujte tuto časovou řadu chronologickým průměrem.

Řešení: $\bar{y} = \frac{1}{9} \left(\frac{14194}{2} + 14271 + \ldots + 14631 + \frac{14738}{2} \right) = 14430.$

Průměr intervalové časové řady

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \,.$$

Příklad: Vypočtěte průměrnou hodnotu roční časové řady HDP ČR (v miliardách Kč) v letech 1994 až 2000.

1994	1995	1996	1997	1998	1999	2000
1303,6	1381,1	1447,7	1432,8	1401,3	1390,6	1433,8

Řešení: $\bar{y} = \frac{1}{7}(1303{,}6 + \ldots + 1433{,}8) = 1398{,}7$.

Dynamické charakteristiky časových řad

Absolutní přírůstky

1. diference: $\Delta y_i = y_i - y_{i-1}, i = 2,\ldots,n$

2. diference: $\Delta^{(2)} y_i = \Delta y_i - \Delta y_{i-1} = y_i - 2y_{i-1} + y_{i-2}, i = 3,\ldots,n$

atd.

(Diferencování má velký význam při odhadu trendu časové řady regresními metodami.)

Průměrný absolutní přírůstek: $\overline{\Delta} = \frac{\sum_{i=2}^{n} \Delta y_i}{n-1} = \frac{y_n - y_1}{n-1}$

Relativní přírůstek

$\delta_i = \frac{\Delta y_i}{y_{i-1}}, i = 2,\ldots,n$

(Relativní přírůstek po vynásobení 100 udává, o kolik procent se změnila hodnota v čase t_i oproti času t_{i-1}.)

Koeficient růstu (tempo růstu)

$k_i = \frac{y_i}{y_{i-1}}, i = 2,\ldots,n$

(Koeficient růstu po vynásobení 100 udává, na kolik procent hodnoty v čase t_{i-1} vzrostla či poklesla hodnota v čase t_i.)

Průměrný koeficient růstu

$\overline{k} = \sqrt[n-1]{k_2 \cdot k_3 \cdot \ldots \cdot k_n} = \sqrt[n-1]{\frac{y_n}{y_1}}$

Průměrný relativní přírůstek

$\overline{\delta} = \overline{k} - 1$

Příklad: Pro časovou řadu HDP ČR v letech 1994 až 2000 (v miliardách Kč) vypočtěte základní charakteristiky dynamiky a graficky znázorněte 1. diference a koeficienty růstu.

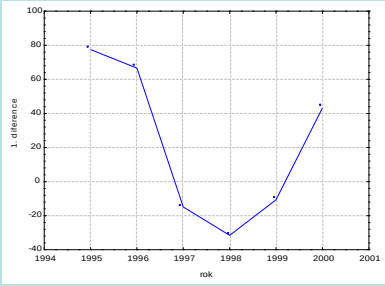
Řešení:

rok	HDP	Δy_i	k_i	δ_i
1994	1303,6	x	x	x
1995	1381,1	77,5	1,059	0,059
1996	1447,7	66,6	1,048	0,048
1997	1432,8	-14,7	0,990	-0,010
1998	1401,3	-31,5	0,978	-0,022
1999	1390,6	-10,7	0,992	-0,008
2000	1433,8	43,2	1,031	0,031

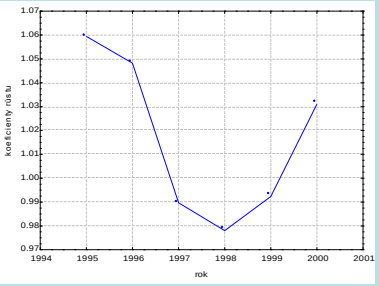
Průměrný absolutní přírůstek: $\overline{\Delta} = \frac{1433{,}8 - 1303{,}6}{6} = 21{,}7$, tzn., že v období 1994 – 2000 rostl HDP průměrně o 21,7 miliard Kč ročně.

Průměrný koeficient růstu: $\overline{k} = \sqrt[6]{\frac{1433{,}8}{1303{,}6}} = 1{,}016$, tzn., že v období 1994 – 2000 rostl HDP průměrně o 1,6% ročně.

Graf 1. diferencí:

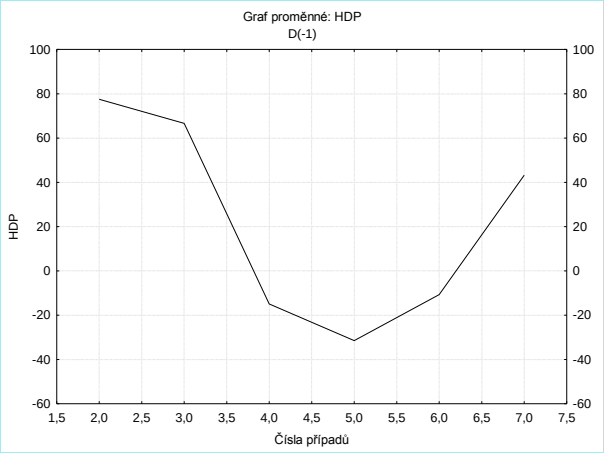


Graf koeficientů růstu:



Výpočet pomocí systému STATISTICA

Statistiky – Pokročilé lineární/nelineární modely – Časové řady/predikce – Proměnné HDP – OK – OK (transformace, autokorelace, kříž. korelace, grafy) – Diferencování - OK (transformovat vybrané řady) – vykreslí se graf.



Vrátíme se do Transformace proměnných – Uložit proměnné. Otevře se nové datové okno, kde v proměnné HDP_1 jsou uloženy 1. diference.

	HDP	HDP_1
1	1303,600	
2	1381,100	77,500
3	1447,700	66,600
4	1432,800	-14,900
5	1401,300	-31,500
6	1390,600	-10,700
7	1433,800	43,200

Výpočet relativních přírůstků: $\delta_i = \frac{\Delta y_i}{y_{i-1}}$ pro i = 2,...,n

Vrátíme se do Transformace proměnných – označíme proměnnou, kterou chceme transformovat (HDP) – vybereme Posun – OK, (Transformovat vybrané řady) – vykreslí se graf.
Vrátíme se do Transformace proměnných – Uložit proměnné. Tato transformovaná veličina se uloží do tabulky pod názvem HDP_1 (proměnná s 1. diferencemi se přejmenuje na HDP_2). Přidáme novou proměnnou RP a do jejího Dlouhého jména napíšeme vzorec =HDP_2/HDP_1.

Výpočet koeficientů růstu: $k_i = \frac{y_i}{y_{i-1}}$ pro i = 2,...,n

Do tabulky přidáme proměnnou KR a do jejího Dlouhého jména napíšeme vzorec =HDP/HDP_1. Získáme tabulku

	1 HDP	2 HDP_2	3 HDP_1	4 RP	5 KR
1	1303,60				
2	1381,10	77,50	1303,60	0,05945	1,05945
3	1447,70	66,60	1381,10	0,04822	1,04822
4	1432,80	-14,90	1447,70	-0,0102	0,98970
5	1401,30	-31,50	1432,80	-0,0219	0,97801
6	1390,60	-10,70	1401,30	-0,0076	0,99236
7	1433,80	43,20	1390,60	0,03106	1,03106
8			1433,80		

Pomocí Grafy - 2D Grafy – Spojnicové grafy (Proměnné) vykreslíme průběh relativních přírůstků a koeficientů růstu.

Průměrný absolutní přírůstek a průměrný koeficient růstu vypočteme na kalkulačce pomocí vzorců
$\bar{\Delta} = \frac{1433,8 - 1303,6}{6} = 21,7$ a $\bar{k} = \sqrt[6]{\frac{1433,8}{1303,6}} = 1,016$.

Aditivní model časové řady

Předpokládejme, že pro časovou řadu $y_1, ..., y_n$ platí model

$y_t = f(t) + \varepsilon_t$, $t = 1, ..., n$, kde

$f(t)$ je neznámá trendová funkce (trend), kterou považujeme za systematickou (deterministickou) složku časové řady (popisuje hlavní tendenci dlouhodobého vývoje časové řady),

ε_t je náhodná složka časové řady zahrnující odchylky od trendu. Náhodná složka splňuje předpoklady

$E(\varepsilon_t) = 0$,

$D(\varepsilon_t) = \sigma^2$,

$C(\varepsilon_t, \varepsilon_{t+h}) = 0$,

$\varepsilon_t \sim N(0, \sigma^2)$ (říkáme, že ε_t je bílý šum).

Odhad trendu časové řady pomocí klouzavých průměrů

Podstata klouzavých průměrů

Předpokládáme, že časová řada se řídí aditivním modelem

$y_t = f(t) + \varepsilon_t$, t = 1, ..., n.

Odhad trendu v bodě t získáme určitým zprůměrováním původních pozorování z jistého okolí uvažovaného časového okamžiku t. Můžeme si představit, že podél dané časové řady klouže okénko, v jehož rámci se průměruje. Nechť toto okénko zahrnuje d členů nalevo od bodu t a d členů napravo od bodu t. Hovoříme pak o vyhlazovacím okénku šířky h = 2d + 1. Prvních a posledních d hodnot trendu neodhadujeme, protože pro $t \in \{1,\ldots,d\} \cup \{n-d+1,\ldots,n\}$ není vyhlazovací okénko symetrické. Odhad trendu ve středu vyhlazovacího okénka je dán vztahem:

$$\hat{f}(t) = \frac{1}{2d+1}\left(y_{t-d} + y_{t-d+1} + \ldots + y_{t+d}\right) = \frac{1}{2d+1}\sum_{k=0}^{2d} y_{t-d+k}$$, t = d+1, ..., n-d.

Šířka vyhlazovacího okénka

Velmi důležitou otázkou je stanovení šířky vyhlazovacího okénka. Je-li okénko příliš široké, bude se odhad trendu blížit přímce (říkáme, že je přehlazen) a zároveň se ztratí velký počet členů na začátku a na konci časové řady. Je-li naopak okénko úzké, bude se odhad trendu blížit původním hodnotám (říkáme, že odhad je podhlazen). Nejčastěji se volí šířka okénka h = 3, 5, 7, pro čtvrtletní hodnoty pak 4.

Příklad: Časová řada 215, 219, 222, 235, 202, 207, 187, 204, 174, 172, 201, 272 udává roční objemy vývozu piva (v miliónech litrů) z Československa v letech 1980 až 1991.

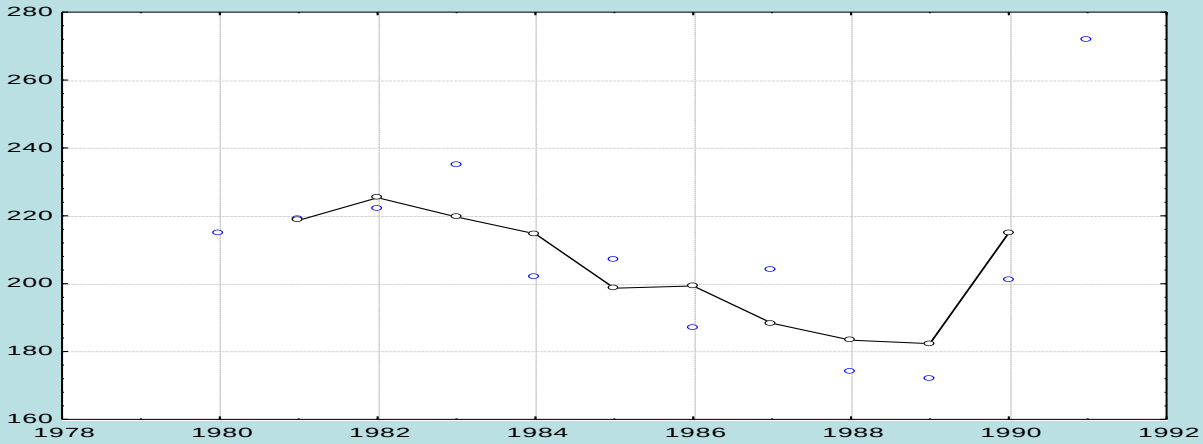
a) Odhadněte trend této časové řady pomocí klouzavých průměrů s vyhlazovacím okénkem šířky 3 a poté 5.
b) Graficky znázorněte průběh časové řady s odhadnutým trendem.

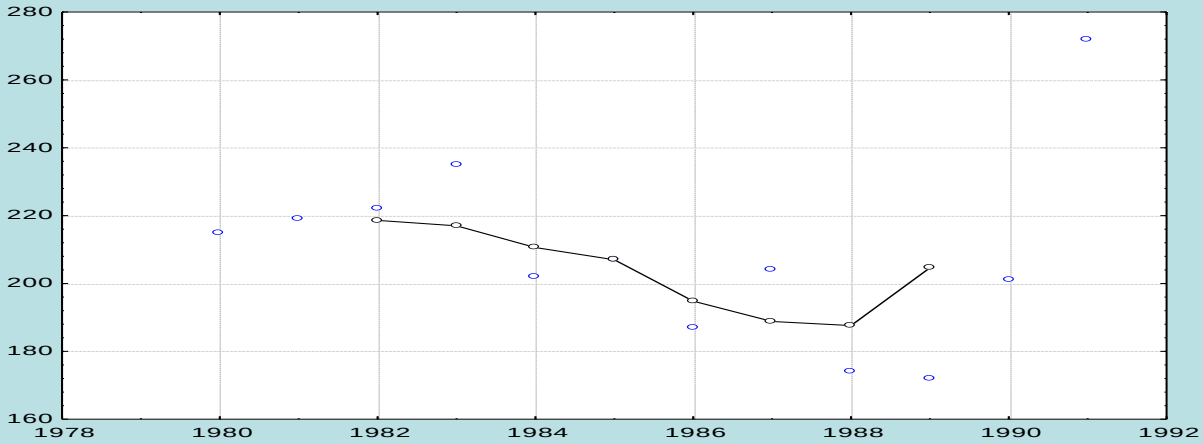
Řešení pomocí systému STATISTICA:

Vytvoříme datový soubor export_piva.sta o dvou proměnných ROK a VYVOZ a dvanácti případech. Statistiky – Pokročilé lineární/nelineární modely – Časové řady/predikce – Proměnné Y – OK– OK (transformace, autokorelace, kříž. korelace, grafy) – Vyhlazování – zaškrtneme N-bod. klouzavý průměr, N = 3 – OK (Transformovat vybrané řady) – vykreslí se graf, vrátíme se do Transformace proměnných – Uložit proměnné. Otevře se nový spreadsheet, kde v proměnné VYVOZ_1 jsou uloženy klouzavé průměry pro N = 3. Totéž uděláme pro případ N = 5. Ve spreadsheetu se proměnná VYVOZ_1 přepíše na VYVOZ_2 a nová proměnná se uloží jako VYVOZ_1. Nově vzniklé proměnné nazveme KP3 a KP5. K datovému souboru přidáme proměnnou ROK, do jejíhož Dlouhého jména napíšeme =1979+v0.

	export_piva.sta			
	1 rok	2 VYVOZ	3 KP3	4 KP5
1	1980	215,000		
2	1981	219,000	218,667	
3	1982	222,000	225,333	218,600
4	1983	235,000	219,667	217,000
5	1984	202,000	214,667	210,600
6	1985	207,000	198,667	207,000
7	1986	187,000	199,333	194,800
8	1987	204,000	188,333	188,800
9	1988	174,000	183,333	187,600
10	1989	172,000	182,333	204,600
11	1990	201,000	215,000	
12	1991	272,000		

Grafické znázornění časové řady s odhadnutým trendem provedeme pomocí vícenásobných bodových grafů.





Porovnání empirického a teoretického rozložení

Motivace: Možnost použití statistických testů je podmíněna nějakými předpoklady o datech. Velmi často je to předpoklad o typu rozložení, z něhož získaná data pocházejí. Mnoho testů je založeno na předpokladu normality. (Testování normality bylo probráno ve 2. kapitole.) Opomíjení předpokladů o typu rozložení může v praxi vést i ke zcela zavádějícím výsledkům, proto je nutné věnovat tomuto problému patřičnou pozornost.

V této kapitole se seznámíme s testem dobré shody, který je (po splnění určitých předpokladů) použitelný k ověření shody empirického rozložení s jakýmkoliv teoretickým rozložením. Tato univerzálnost je ovšem provázena poněkud sníženou silou testu. Proto byly pro některá rozložení vyvinuty speciální testy využívající charakteristických vlastností těchto rozložení. Zde uvedeme tzv. jednoduché testy exponenciálního a Poissonova rozložení.

Testy dobré shody

Popis testu

Testujeme hypotézu, která tvrdí, že náhodný výběr $X_1, ..., X_n$ pochází z rozložení s distribuční funkcí $\Phi(x)$.

Spojitý případ:

- data rozdělíme do r třídicích intervalů $\langle u_j, u_{j+1} \rangle$, j = 1, ..., r
- zjistíme absolutní četnost n_j j-tého třídicího intervalu
- vypočteme pravděpodobnost p_j, že náhodná veličina X s distribuční funkcí $\Phi(x)$ se bude realizovat v j-tém třídicím intervalu. Platí-li nulová hypotéza, pak $p_j = \Phi(u_{j+1}) - \Phi(u_j)$.

Diskrétní případ:

- určíme varianty $x_{[j]}$, j = 1, …, r
- pro variantu $x_{[j]}$ zjistíme absolutní četnost n_j
- vypočteme pravděpodobnost p_j, že náhodná veličina X s distribuční funkcí $\Phi(x)$ se bude realizovat variantou $x_{[j]}$. Platí-li nulová hypotéza, pak $p_j = \Phi(x_{[j]}) - \lim_{x \to x_{[j]}-} \Phi(x) = P(X = x_{[j]})$.

Testová statistika: $K = \sum_{j=1}^{r} \frac{(n_j - np_j)^2}{np_j}$. Platí-li nulová hypotéza, pak $K \approx \chi^2(r-1-p)$, kde p je počet odhadovaných parametrů daného rozložení. (Např. pro normální rozložení p = 2, protože z dat odhadujeme střední hodnotu a rozptyl.) Pokud žádný parametr nemusíme odhadovat, hovoříme o úplně specifikovaném problému. Nulovou hypotézu zamítáme na asymptotické hladině významnosti α, když $K \geq \chi^2_{1-\alpha}(r-1-p)$. Aproximace se považuje za vyhovující, když $np_j \geq 5$, j = 1, ..., r.

Upozornění: Při nesplnění podmínky $np_j \geq 5$, j = 1, ..., r je třeba některé intervaly resp. varianty slučovat, což vede ke ztrátě informace. Ve spojitém případě je hodnota testové statistiky K silně závislá na volbě třídicích intervalů

Příklad: (Testování shody empirického a teoretického rozložení při úplně specifikovaném problému)
Ze souboru rodin s pěti dětmi bylo náhodně vybráno 84 rodin a byl zjišťován počet chlapců:

Počet chlapců	0	1	2	3	4	5
Počet rodin	3	10	22	31	14	4

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že rozložení počtu chlapců se řídí binomickým rozložením Bi(5; 0,5).
Řešení: Počet chlapců v náhodně vybrané rodině s 5 dětmi je náhodná veličina s rozložením Bi(5; 0,5), její pravděpodobnostní funkce je

$$p_j = \binom{5}{j}\frac{1}{32},\ j = 0,1,\ldots,5.$$

Výpočty potřebné pro stanovení testové statistiky K uspořádáme do tabulky.

j	n_j	p_j	np_j
0	3	0,03125	84.0,03125=2,625
1	10	0,15625	84.0,15625=13,125
2	22	0,3125	84.0,3125=26,25
3	31	0,3125	84.0,3125=26,25
4	14	0,15625	84.0,15625=13,125
5	4	0,03125	84.0,03125=2,625

Podmínky dobré aproximace nejsou splněny, sloučíme tedy první dvě varianty a poslední dvě varianty.

j	n_j	p_j	np_j	$\frac{(n_j - np_j)^2}{np_j}$
0 a 1	13	0,1875	84.0,1875=15,75	0,480159
2	22	0,3125	84.0,3125=26,25	0,688095
3	31	0,3125	84.0,3125=26,25	0,859524
4 a 5	18	0,1875	84.0,1875=15,75	0,321429

Vypočteme realizaci testové statistiky: K = 0,48059 + 0,688095 + 0,859524 + 0,321429 = 2,3492, počet tříd r = 4, počet odhadovaných parametrů p = 0, r – p - 1 = 3, kritický obor $W = \langle \chi^2_{1-\alpha}(r-p-1), \infty) = \langle \chi^2_{0,95}(3), \infty) = \langle 7,8147; \infty)$. Protože $K \notin W$, nulovou hypotézu nezamítáme na asymptotické hladině významnosti 0,05.

Výpočet pomocí systému STATISTICA:

Vytvoříme datový soubor se dvěma proměnnými a čtyřmi případy. Proměnná nj obsahuje zjištěné četnosti (po sloučení variant), proměnná npj pak teoretické četnosti.

Statistiky – Neparametrická statistika – Pozorované vs. očekávané $\chi 2$ – OK – Proměnné – Pozorované četnosti nj, očekávané četnosti npj – OK – Výpočet.

	Pozorované vs. očekávané četnosti (Tabul Chi-Kvadr. = 2,349206 sv = 3 p = ,503161			
Případ	pozorov. nj	očekáv. npj	P - O	(P-O)^2 /O
C: 1	13,0000	15,7500	-2,7500	0,48015
C: 2	22,0000	26,2500	-4,2500	0,68809
C: 3	31,0000	26,2500	4,7500	0,85952
C: 4	18,0000	15,7500	2,2500	0,32142
Sčt	84,0000	84,0000	0,0000	2,34920

V záhlaví výstupní tabulky je uvedena hodnota testového kritéria (2,349206), počet stupňů volnosti = 3 a p-hodnota (0,503161). Nulová hypotéza se tedy nezamítá na asymptotické hladině významnosti 0,05.

Příklad: (Testování shody empirického a teoretického rozložení při neúplně specifikovaném problému – diskrétní případ)
V tabulce jsou roztříděny fotbalové zápasy určité soutěže podle počtu vstřelených branek.

Počet branek	0	1	2	3	4 a víc
Počet zápasů	19	30	17	10	8

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že jde o výběr z Poissonova rozložení.

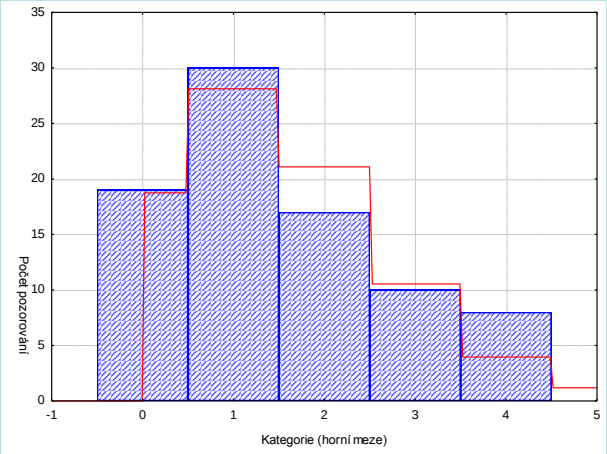
Výpočet pomocí systému STATISTICA:
Vytvoříme datový soubor s dvěma proměnnými a 5 případy. Proměnná POCET obsahuje počet vstřelených branek, proměnná CETNOST pak počet zápasů, v nichž bylo dosaženo zjištěného počtu branek.
Statistiky – Prokládání rozdělení – Diskrétní rozdělení – Poissonovo – OK – Proměnná POCET – klikneme na ikonu se závažím – Proměnná vah CETNOST – Stav Zapnuto – OK – Výpočet.

Proměnná:POCET, Rozdělení:Poissonovo, Lambda = 1,500 (branky.sta)
Chí-kvadrát = 2,07051, sv = 3, p = 0,55790

Kategorie	Pozorované Četnosti	Kumulativ. Pozorované	Procent Pozorované	Kumul. % Pozorované	Očekáv. Četnosti	Kumulativ. Očekáv.	Procent Očekáv.	Kumul. % Očekáv.
<= 0,00000	19	19	22,6190	22,619(	18,7429	18,7429	22,3130	22,313(
1,00000	30	49	35,7142	58,333	28,1144	46,8573	33,4695	55,782
2,00000	17	66	20,2381	78,571	21,0858	67,9431	25,1021	80,884
3,00000	10	76	11,9047	90,476	10,5429	78,4860	12,5510	93,435
< Nekonečno	8	84	9,5238	100,000	5,5139	84,0000	6,5642	100,000

V tomto případě je parametr λ Poissonova rozložení neznámý, je odhadnut pomocí výběrového průměru a odhad činí 1,5. Podmínky dobré aproximace jsou splněny, dokonce všechny teoretické četnosti jsou větší než 5. Dále je v záhlaví výstupní tabulky uvedena hodnota testového kritéria (2,07051), počet stupňů volnosti r – p – 1 = 5 – 1 – 1 = 3 a p-hodnota (0,5578). Nulová hypotéza se tedy nezamítá na asymptotické hladině významnosti 0,05.

Pro vytvoření grafu se vrátíme do Proložení diskrétních rozložení – Základní výsledky – Graf pozorovaného a očekávaného rozdělení.



Příklad: (Testování shody empirického a teoretického rozložení při neúplně specifikovaném problému – spojitý případ)
U 48 studentek VŠE v Praze byla zjišťována výška (v cm): 165 170 170 179 170 168 174 162 167 165 170 173 183 176 165 168 171 178 168 168 169 163 172 184 176 175 176 169 168 170 166 160 167 162 162 166 170 168 155 162 169 166 160 169 165 163 168 163
Pomocí testu dobré shody testujte na hladině významnosti 0,05 hypotézu, že data pocházejí z normálního rozložení. Pomocí histogramu posuďte vizuálně předpoklad normality.

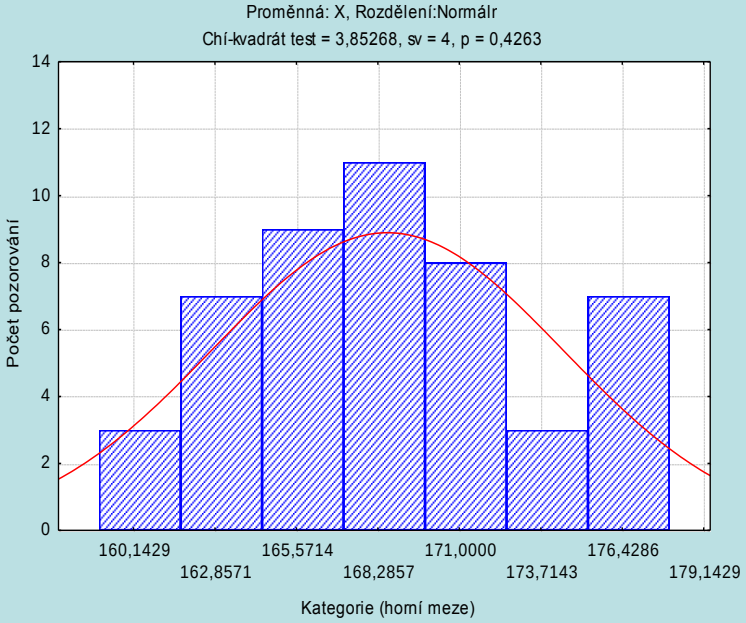
Výpočet pomocí systému STATISTICA:
Statistiky - Prokládání rozdělení – ponecháme implicitní nastavení na normální rozložení – OK – Proměnná X – OK – na záložce Parametry změníme Počet kategorií na 7 (podle Sturgesova pravidla) – Výpočet.

	Proměnná: X, Rozdělení:Normální (vyska.sta) Chí-kvadrát = 1,09280, sv = 1 (uprav.) , p = 0,29585							
Horní hranice	Pozorované Četnosti	Kumulativ. Pozorované	Procent Pozorované	Kumul. % Pozorované	Očekáv. Četnosti	Kumulativ. Očekáv.	Procent Očekáv.	Kumul. % Očekáv.
<= 157,14286	1	1	2,0833	2,0833	1,1970	1,1970	2,4938	2,4939
162,28571	6	7	12,5000	14,5833	5,5148	6,7118	11,4892	13,9831
167,42857	12	19	25,0000	39,5833	13,4622	20,1740	28,0462	42,0293
172,57143	19	38	39,5833	79,1667	15,8914	36,0655	33,1072	75,1366
177,71429	6	44	12,5000	91,6667	9,0770	45,1425	18,9104	94,0470
182,85714	2	46	4,1666	95,8333	2,5036	47,6462	5,2159	99,2629
< Nekonečno	2	48	4,1666	100,000	0,3538	48,0000	0,7370	100,000

Při tomto roztřídění dat do 7 intervalů nejsou splněny podmínky dobré aproximace, ve třech intervalech jsou teoretické četnosti pod 5. Změníme tedy dolní mez na 159 a horní na 178.

Horní hranice	Proměnná: X, Rozdělení:Normální (vyska.sta) Chí-kvadrát = 3,85268, sv = 4, p = 0,42631 Pozorované Četnosti	Kumulativ. Pozorované	Procent Pozorované	Kumul. % Pozorované	Očekáv. Četnosti	Kumulativ. Očekáv.	Procent Očekáv.	Kumul. % Očekáv.
<= 161,71429	3	3	6,2500(	6,250(	5,72299(	5,7230(	11,9229.	11,922(
164,42857	7	10	14,5833:	20,833:	5,67594(	11,3989.	11,8248(	23,747(
167,14286	9	19	18,7500(	39,583:	7,86263:	19,2615'	16,3804(	40,128:
169,85714	11	30	22,9166'	62,500(	8,81245(	28,0740:	18,3592(	58,487(
172,57143	8	38	16,6666'	79,166'	7,99151(	36,0655(	16,6489(	75,136(
175,28571	3	41	6,2500(	85,416'	5,86355(	41,9291(	12,2157(	87,352:
< Nekonečno	7	48	14,5833:	100,000(	6,07089(	48,0000(	12,6477(	100,000(

V tomto případě jsou podmínky dobré aproximace splněny. Testová statistika se realizuje hodnotou 3,85268, p-hodnota je 0,42631, tedy na asymptotické hladině významnosti 0,05 hypotézu o normalitě nezamítáme. Podívejme se ještě na histogram s proloženou Gaussovou křivkou: Na záložce Základní výsledky zvolíme Graf pozorovaného a očekávaného rozdělení.



Jednoduchý test exponenciálního a Poissonova rozložení

Jednoduchý test exponenciálního rozložení

Testujeme hypotézu, která tvrdí, že náhodný výběr $X_1, ..., X_n$ pochází z exponenciálního rozložení. Označme M výběrový průměr a S^2 výběrový rozptyl tohoto náhodného výběru. Víme, že střední hodnota náhodné veličiny $X \sim Ex(\lambda)$ je $E(X) = 1/\lambda$ a rozptyl je $D(X) = 1/\lambda^2$. Test založíme na statistice $K = \frac{(n-1)S^2}{M^2}$, která se v případě platnosti H_0 asymptoticky řídí rozložením $\chi^2(n-1)$. Kritický obor: $W = \langle 0, \chi^2_{\alpha/2}(n-1)\rangle \cup \langle \chi^2_{1-\alpha/2}(n-1), \infty)$. Jestliže $K \in W$, H_0 zamítáme na asymptotické hladině významnosti α.

Příklad

Byla zkoumána doba životnosti 45 součástek (v hodinách). Průměrná životnost byla m = 99,93 a rozptyl s^2 = 7328,91. Na asymptotické hladině významnosti 0,05 testujte hypotézu, že daný náhodný výběr pochází z exponenciálního rozložení.

Řešení:

Testovou statistiku K vypočteme podle vzorce $K = \frac{(n-1)S^2}{M^2}$. Kritický obor má tvar: $W = \langle 0; \chi^2_{\alpha/2}(n-1)\rangle \cup \langle \chi^2_{1-\alpha/2}(n-1); \infty)$.

V našem případě $K = 32{,}2924$, $W = \langle 0; 27{,}575\rangle \cup \langle 64{,}202; \infty)$, H_0 tedy nezamítáme na asymptotické hladině významnosti 0,05.

Jednoduchý test Poissonova rozložení

Testujeme hypotézu, která tvrdí, že náhodný výběr $X_1, ..., X_n$ pochází z Poissonova rozložení. Označme M výběrový průměr a S^2 výběrový rozptyl tohoto náhodného výběru. Víme, že střední hodnota náhodné veličiny $X \sim Po(\lambda)$ je $E(X) = \lambda$ a rozptyl je $D(X) = \lambda$. Test založíme na statistice $K = \frac{(n-1)S^2}{M}$, která se v případě platnosti H_0 asymptoticky řídí rozložením $\chi^2(n-1)$. Kritický obor: $W = \langle 0, \chi^2_{\alpha/2}(n-1)\rangle \cup \langle \chi^2_{1-\alpha/2}(n-1), \infty)$. Jestliže $K \in W$, H_0 zamítáme na asymptotické hladině významnosti α.

Příklad

Studujeme rozložení počtu pacientů, kteří během 75 dnů přijdou na pohotovost. Osmihodinovou pracovní dobu rozdělíme do půlhodinových intervalů a v každém intervalu zjistíme počet příchozích pacientů:

Počet pacientů	0	1	2	3	4	4	6	7	8	9	10
Pozrovaná četnost	79	188	282	275	196	114	45	10	7	3	1

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že daný náhodný výběr pochází z Poissonova rozložení.

Řešení:

Celkový počet pacientů je n = 1200. Realizaci výběrového průměru M získáme jako vážený průměr počtu pacientů (m = 2,8033) a realizaci výběrového rozptylu S^2 získáme jako vážený rozptyl počtu pacientů (s^2 = 2,7086). Testovou statistiku vypočteme podle vzorce $K = \frac{(n-1)S^2}{M}$, tedy K = 1158,5, kritický obor

$$W = \langle 0, \chi^2_{\alpha/2}(n-1)\rangle \cup \langle \chi^2_{1-\alpha/2}(n-1), \infty) = \langle 0, \chi^2_{0,025}(1199)\rangle \cup \langle \chi^2_{0,975}(1199), \infty) =$$

$$= \langle 0; 1104,93\rangle \cup \langle 1296,86; \infty).$$

Protože testová statistika se nerealizuje v kritickém oboru, H_0 nezamítáme na asymptotické hladině významnosti 0,05.