Téma 12: Regresní přímka

Vzorový 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.
a) Vypočtěte odhady regresních parametrů, napište rovnici regresní přímky a interpretujte její parametry. Do dvourozměrného tečkového diagramu zakreslete regresní přímku s 95% pásem spolehlivosti a 95% predikčním pásem.
b) Najděte odhad rozptylu, proveďte celkový F-test a rovněž dílčí t-testy o významnosti regresních parametrů..
c) Najděte 95% intervaly spolehlivosti pro regresní parametry a zjistěte relativní chyby odhadů regresních parametrů.
d) Vypočtěte index determinace a interpretujte ho. Vypočtěte rovněž střední absolutní procentuální chybu predikce (MAPE) a najděte regresní odhad letošní poptávky při loňské poptávce 110 kusů.

Řešení:

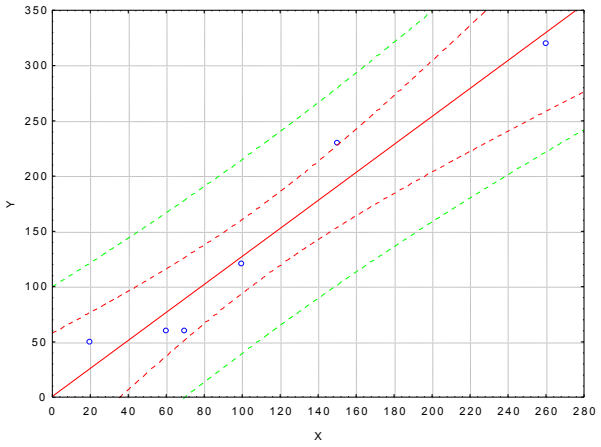
Ad a) Vytvoříme nový datový soubor se dvěma proměnnými X a Y a 6 případy:

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

	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					
N=6	Beta	Sm.chyba beta	B	Sm.chyba B	t(4)	Úroveň p
Abs.člen			0,686813	20,64236	0,033272	0,975052
X	0,971977	0,117538	1,266484	0,15315	8,269474	0,001167

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.

Do dvourozměrného tečkového diagramu nyní nakreslíme regresní přímku s 95% regresními pásy. Grafy – Bodové grafy – Proměnné X, Y – OK – na záložce Detaily zvolíme Regresní pásy, zaškrtneme Spolehl. – OK. Dále potřebujeme přidat predikční pás. 2x klikneme na vytvořený graf a v Možnostech grafu vybereme Regresní pásy – Přidat nový pár pásů – Typ Predikční – ve Vzoru změnime barvu na zelenou – OK.



Vzhled grafu naznačuje, že přímka je vhodným modelem závislosti letošní poptávky na loňské poptávce.

Ad b) Abychom získali odhad rozptylu, vrátíme se do Výsledky – vícenásobná regrese – Detailní výsledky – ANOVA.

	Analýza rozptylu (Tabulka1)				
Efekt	Součet čtverců	sv	Průměr čtverců	F	Úroveň p
Regres.	58384,89	1	58384,89	68,38420	0,001167
Rezid.	3415,11	4	853,78		
Celk.	61800,00				

Odhad rozptylu najdeme na řádku Rezid., ve sloupci Průměr čtverců, tedy $s^2 = 853{,}78$.

Testovou statistiku F-testu a odpovídající p-hodnotu najdeme v záhlaví výstupní tabulky regrese:

	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					
N=6	Beta	Sm.chyba beta	B	Sm.chyba B	t(4)	Úroveň p
Abs.člen			0,686813	20,64236	0,033272	0,975052
X	0,971977	0,117538	1,266484	0,15315	8,269474	0,001167

Zde F = 68,384, p-hodnota < 0,00117, tedy na hladině významnosti 0,05 zamítáme hypotézu o nevýznamnosti modelu jako celku.
Výsledky F-testu jsou rovněž uvedeny v tabulce ANOVA.

Výsledky dílčích t-testů jsou uvedeny ve výstupní tabulce regrese. Testová statistika pro test hypotézy H_0: $\beta_0 = 0$ je 0,033272, p-hodnota je 0,975052. Hypotézu o nevýznamnosti úseku regresní přímky tedy nezamítáme na hladině významnosti 0,05. Testová statistika pro test hypotézy H_0: $\beta_1 = 0$ je 8,269474, p-hodnota je 0,001167. Hypotézu o nevýznamnosti směrnice regresní přímky tedy zamítáme na hladině významnosti 0,05.

Ad c) Ve výstupní tabulce výsledků regrese přidáme za proměnnou Úroveň p tři nové proměnné: dm (pro dolní meze 95% intervalů spolehlivosti pro regresní parametry), hm (pro horní meze 95% intervalů spolehlivosti pro regresní parametry) a chyba (pro relativní chyby odhadů regresních parametrů).
Do Dlouhého jména proměnné dm napíšeme:
=v3-v4*VStudent(0,975;4)
Do Dlouhého jména proměnné hm napíšeme:
=v3+v4*VStudent(0,975;4)
Do Dlouhého jména proměnné chyba napíšeme:
=100*abs(0,5*(hm-dm)/v3)

	Výsledky regrese se závislou proměnnou : Prom2 (Tabulka1) R= ,97197702 R2= ,94473932 Upravené R2= ,93092415 F(1,4)=68,384 p<,00117 Směrod. chyba odhadu : 29,219								
N=6	Beta	Sm.chyba beta	B	Sm.chyba B	t(4)	Úroveň p	dm =v3-v4*V	hm =v3+v4*	chyba =100*abs
Abs.člen			0,686813	20,64236	0,033272	0,975052	-56,6256	57,99918	8344,681
Prom1	0,971977	0,117538	1,266484	0,15315	8,269474	0,001167	0,841266	1,691701	33,57463

Vidíme, že $-56,63 < \beta_0 < 58$ s pravděpodobností aspoň 0,95 a $0,841 < \beta_1 < 1,692$ s pravděpodobností aspoň 0,95.
Relativní chyba odhadu parametru β_0 činí 8344,68% a relativní chyba odhadu parametru β_1 činí 33,57%. V obou případech jsou chyby příliš velké.

Ad d) Index determinace je uveden v záhlaví původní výstupní tabulky pod označením R2:

	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					
N=6	Beta	Sm.chyba beta	B	Sm.chyba B	t(4)	Úroveň p
Abs.člen			0,686813	20,64236	0,033272	0,975052
X	0,971977	0,117538	1,266484	0,15315	8,269474	0,001167

V našem případě $ID^2 = 0,9447$, tedy variabilita letošní poptávky je z 94,5% vysvětlena regresní přímkou.

Abychom vypočetli MAPE, tak ve výsledcích Vícenásobné regrese zvolíme záložku Rezidua / předpoklady / předpovědi – Reziduální analýza – Uložit – Uložit rezidua a předpovědi – Vybrat vše – OK. Ve vzniklé tabulce přidáme proměnnou chyby a do jejího Dlouhého jména napíšeme
=100*abs(v4/v2)
Pak spočteme průměr této proměnné a zjistíme, že MAPE = 25,17%.

Pro výpočet predikované hodnoty zvolíme Rezidua/předpoklady/předpovědi Předpovědi závisle proměnné X: 110 OK. Ve výstupní tabulce je hledaná hodnota označena jako Předpověď.

Proměnná	Předpovězené hodnoty (Tabulka1) proměnné: Y		
	B-váž	Hodnota	B-váž * Hodnot
X	1,266484	110,0000	139,3132
Abs. člen			0,6868
Předpověď			140,0000
-95,0%LS			106,8803
+95,0%LS			173,1197

Při loňské poptávce 110 kusů je predikovaná hodnota letošní poptávky 140 kusů.

Ad e) Při analýze reziduí nejprve posoudíme nezávislost reziduí pomocí Durbinova – Watsonovy statistiky: Na záložce Rezidua/předpoklady/předpovědi zvolíme Reziduální analýza - Pokročilá – Durbinova – Watsonova statistika.

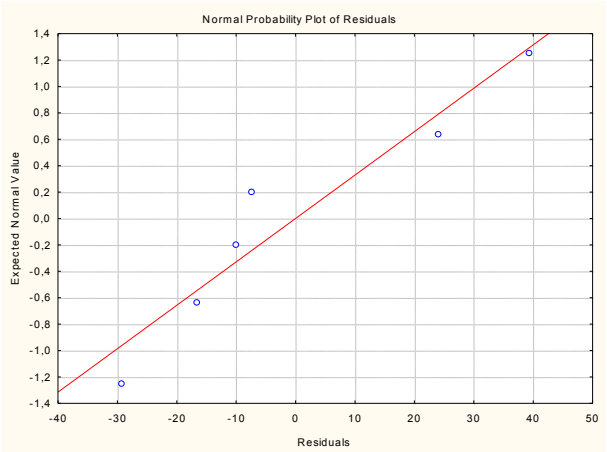
	Durbin-Watsonovo d (poptavka.sta) a sériové korelace reziduí	
	Durbin-Watson.d	Sériové korelace
Odhad	2,022847	-0,113505

Tato statistika je blízká číslu 2, tedy rezidua můžeme považovat za nezávislá.

Normalitu reziduí posoudíme Lilieforsovou variantou K-S testu a S-W testem:

Proměnná	Testy normality (Tabulka6)				
	N	max D	Lilliefors p	W	p
Rezidua	6	0,277184	p < ,15	0,911935	0,449251

Ani jeden z testů nezamítá hypotézu o normalitě reziduí na hladině významnosti 0,05. Graficky posoudíme normalitu N-P plotem:



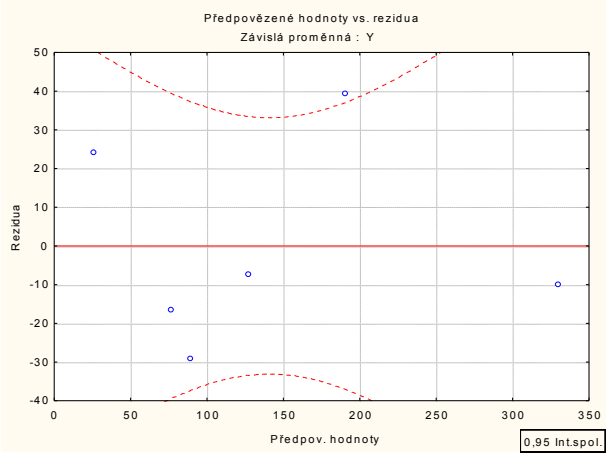
Vidíme, že rezidua se od ideální přímky neodchylují příliš výrazně.

Nulovost střední hodnoty reziduí ověříme jednovýběrovým t-testem:

Proměnná	Test průměrů vůči referenční konstantě (hodnotě) (Tabulka6)							
	Průměr	Sm.odch.	N	Sm.chyba	Referenční konstanta	t	SV	p
Rezidua	-0,000003	26,13469	6	10,66944	0,00	-0,000000	5	1,000000

Vidíme, že p-hodnota je 1, tudíž na hladině významnosti 0,05 nezamítáme hypotézu, že rezidua mají nulovou střední hodnotu.

Homoskedasticitu reziduí posoudíme pomocí grafu závislosti reziduí na predikovaných hodnotách veličiny Y: Na záložce Rezidua/předpoklady/předpovědi zvolíme Reziduální analýza –Bodové grafy – Předpovědi vs. Rezidua



Rezidua nevykazují žádnou závislost na predikovaných hodnotách.

Příklad k samostatnému řešení
Použijte datový soubor s údaji o mobilech. Zkoumejte závislost objemu mobilu na jeho hmotnosti. Proveďte všechny úkoly, které byly popsány ve vzorovém příkladu.