Cvičení 12: Regresní analýza

Úkol 1.: 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

a) Orientačně ověřte předpoklad, že data pocházejí z dvourozměrného normálního rozložení. Vypočtěte výběrový koeficient korelace mezi X a Y, interpretujte jeho hodnotu a na hladině významnosti 0,05 testujte hypotézu, že X a Y jsou nezávislé náhodné veličiny.

b) 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.

c) Najděte odhad rozptylu, vypočtěte index determinace a interpretujte ho.

d) Najděte 95% intervaly spolehlivosti pro regresní parametry a zjistěte relativní chyby odhadů regresních parametrů.

e) Na hladině významnosti 0,05 proveďte celkový F-test.

f) Na hladině významnosti 0,05 proveďte dílčí t-testy.

g) Vypočtěte regresní odhad letošní poptávky při loňské poptávce 110 kusů.

h) Nakreslete dvourozměrný tečkový diagram s proloženou regresní přímkou.

i) Spočtěte střední absolutní procentuální chybu predikce (MAPE)

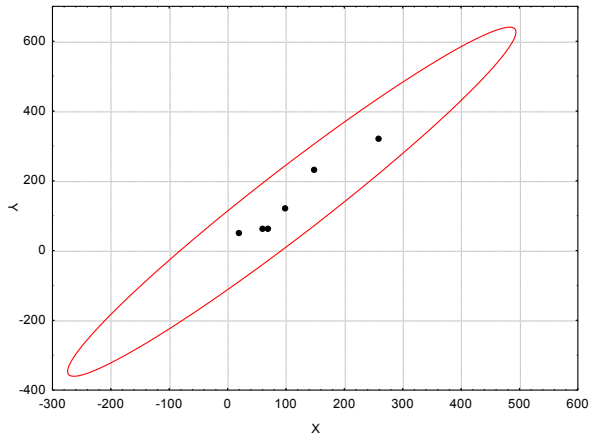
j) Proveďte analýzu reziduí.

Návod:
Načteme nový datový soubor obchodnici.sta se dvěma proměnnými X a Y a 6 případy:

	1 X	2 Y
1	20	50
2	60	60
3	70	60
4	100	120
5	150	230
6	260	320

a) Orientačně ověřte předpoklad, že data pocházejí z dvourozměrného normálního rozložení. Vypočtěte výběrový koeficient korelace mezi X a Y, interpretujte jeho hodnotu a na hladině významnosti 0,05 testujte hypotézu, že X a Y jsou nezávislé náhodné veličiny.

Zobrazíme dvourozměrný tečkový diagram s proloženou elipsou 95% konstantní hustoty pravděpodobnosti, s jehož pomocí posoudíme dvourozměrnou normalitu dat: Grafy – Bodové grafy – vypneme Typ proložení – Proměnné X, Y - OK . Na záložce Detaily vybereme Elipsa Normální – OK. Ve vzniklém dvourozměrném tečkovém diagramu změníme rozsah zobrazených hodnot na vodorovné a svislé ose, abychom viděli celou elipsu



Ze vzhledu diagramu je patrné, že předpoklad dvourozměrné normality je oprávněný a že mezi loňskou a letošní poptávkou existuje vcelku silná přímá lineární závislost.

Testování hypotézy o nezávislosti: Statistika – Základní statistiky /Tabulky - Korelační matice – OK – 2 seznamy proměnných X, Y, OK. Na záložce Možnosti zaškrtneme Zobrazit detailní tabulku výsledků – Souhrn.

Prom. X & prom. Y	Korelace (Tabulka1) Označ. korelace jsou významné na hlad. p < ,05000 (Celé případy vynechány u ChD) Průměr	Sm.Odch.	r(X,Y)	r2	t	p	N	Konst. záv.: Y	Směr. záv: Y	Konst. záv.: X	Směrnic záv.: X
X	110,0000	85,3229									
Y	140,0000	111,1755	0,971977	0,944739	8,269474	0,001167	6	0,686813	1,266484	5,566343	0,745955

Ve výstupní tabulce najdeme hodnotu výběrového korelačního koeficientu R_{12} (r = 0,971977, tzn. že mezi X a Y existuje velmi silná přímá lineární závislost), realizaci testové statistiky t = 8,269474 a p-hodnotu pro test hypotézy o nezávislosti (p = 0,001167, H_0 tedy zamítáme na hladině významnosti 0,05).

b) Předpokládejte, že závislost letošní poptávky na loňské lze vystihnout regresní přímkou. Vypočtěte odhady regresních parametrů a napište rovnici regresní přímky. Interpretujte parametry regresní přímky.

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			0,686813	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.

c) Najděte odhad rozptylu, vypočtěte index determinace a interpretujte ho.

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 $s^2 = 853{,}78$.
Index determinace je uveden v záhlaví původní výstupní tabulky pod označením R2. V našem případě $ID^2 = 0{,}9447$, tedy variabilita letošní poptávky je z 94,5% vysvětlena regresní přímkou.

d) Najděte 95% intervaly spolehlivosti pro regresní parametry.

Ve výstupní tabulce výsledků regrese přidáme za proměnnou Úroveň p dvě nové proměnné dm (pro dolní meze 95% intervalů spolehlivosti pro regresní parametry) a hm (pro horní meze 95% intervalů spolehlivosti pro regresní parametry). Do Dlouhého jména proměnné dm resp. hm napíšeme: =v3-v4*VStudent(0,975;4) resp. =v3+v4*VStudent(0,975;4)

	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	dm =v3-v4*V	hm =v3+v4*
Abs.člen			0,686813	20,64236	0,033272	0,975052	-56,6256	57,99918
X	0,971977	0,117538	1,266484	0,15315	8,269474	0,001167	0,841266	1,691701

Vidíme, že $-56{,}63 < \beta_0 < 58$ s pravděpodobností aspoň 0,95 a $0{,}841 < \beta_1 < 1{,}692$ s pravděpodobností aspoň 0,95.

e) Na hladině významnosti 0,05 proveďte celkový F-test.

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

f) Na hladině významnosti 0,05 proveďte dílčí t-testy a vypočtěte relativní chyby odhadů regresních parametrů.

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.
K upravené výstupní tabulce s mezemi intervalů spolehlivosti přidáme proměnnou chyba. Do jejího Dlouhého jména napíšeme
=100*abs(0,5*(hm-dm)/v3)

N=6	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								
	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

Výsledek pro parametr β_0: Protože p = 0,975 < 0,05, hypotézu o nevýznamnosti regresního parametru β_0 (tj. posunutí regresní přímky) nezamítáme na hladině významnosti 0,05.
Výsledek pro parametr β_1: Protože p = 0,0012 < 0,05, hypotézu o nevýznamnosti regresního parametru β_1 (tj. směrnice regresní přímky) zamítáme na hladině významnosti 0,05.

g) Vypočtěte regresní odhad letošní poptávky při loňské poptávce 110 kusů.

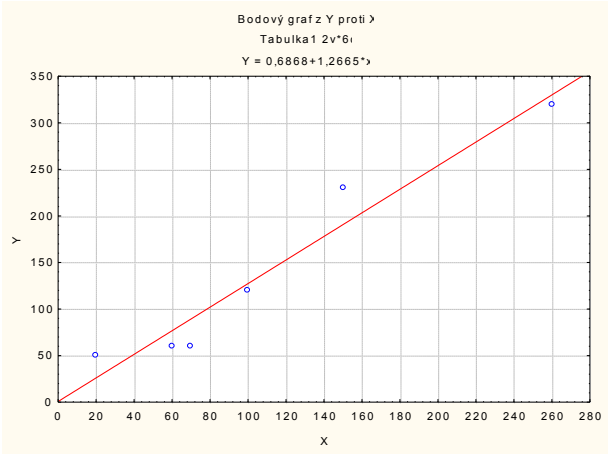
Pro výpočet predikované hodnoty zvolíme Rezidua/předpoklady/předpovědi Předpovědi závisle proměnné X: 110 OK. Ve výstupní tabulce je hledaná hodnota označena jako Předpověď.

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.

h) Nakreslete dvourozměrný tečkový diagram s proloženou regresní přímkou.

Do dvourozměrného tečkového diagramu nakreslíme regresní přímku tak, že v tabulce 2D Bodové grafy zvolíme Typ proložení: Lineární, OK.



i) Vypočtěte střední absolutní procentuální chybu predikce (MAPE)

Ve výsledcích Vícenásobné regrese zvolíme záložku Rezidua / předpoklady / předpovědi – Reziduální analýza – Uložit – Uložit rezidua a předpovědi – Vybrat vše – OK. Ve vzniklé tabulce odstraníme proměnné 5 – 10, přidáme proměnnou chyby a do jejího Dlouhého jména napíšeme
=100*abs(v4/v2)
Pak spočteme průměr této proměnné a zjistíme, že MAPE = 25,17%.

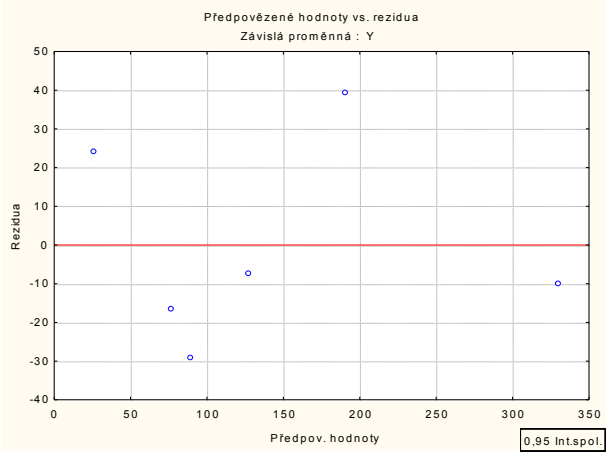
j) Proveďte analýzu reziduí.

Posouzení nezávislosti reziduí pomocí Durbinovy – Watsonovy statistiky:
Statistiky – Vícenásobná regrese – proměnná Závislá: y, nezávislá x – OK – na záložce Residua/předpoklady/předpovědi vybereme Reziduální analýza - Detaily – Durbin-Watsonova statistika:

	Durbin-Watson.d	Sériové korelace
Odhad	2,022847	-0,113505

Hodnota této statistiky je blízká 2, svědčí o tom, že rezidua jsou nekorelovaná.

Posouzení homoskedasticity reziduí
Reziduální analýza – Bodové grafy – Předpovědi vs. rezidua



Rezidua jsou kolem 0 rozmístěna náhodně.

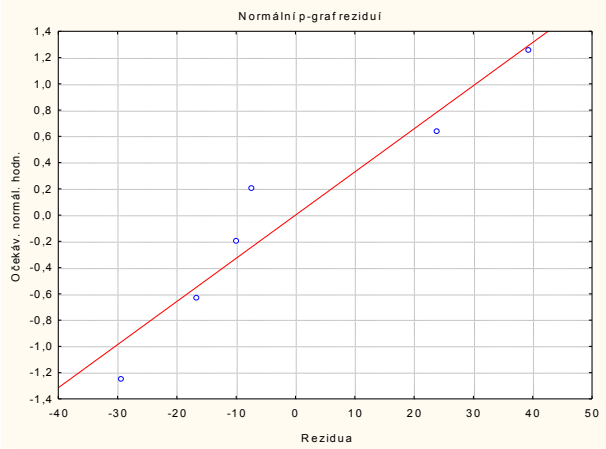
Testování nulovosti střední hodnoty reziduí:

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	-0,000003	26,13469	6	10,66944	0,00	-0,000000	5	1,000000

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:

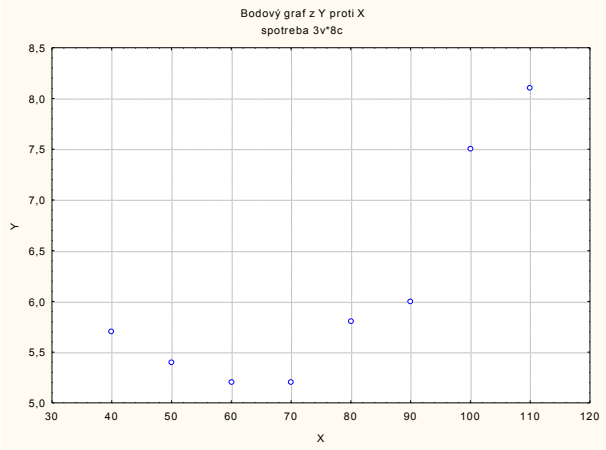


Rezidua se řadí kolem ideální přímky, lze tedy soudit, že se řídí normálním rozložení.

Úkol 2.: (Příklad je převzat z knihy Jiří Anděl: Matematická statistika, SNTL/Alfa, Praha, 1978, str. 111)
U automobilu Škoda 120 byla změřena spotřeba benzínu (v l/100 km) v závislosti na rychlosti (v km/h).

rychlost	40	50	60	70	80	90	100	110
spotřeba	5,7	5,4	5,2	5,2	5,8	6,0	7,5	8,1

a) Data znázorněte graficky dvourozměrným tečkovým diagramem a najděte vhodnou regresní funkci.
Načteme datový soubor spotreba_benzinu.sta se dvěma proměnnými X a Y a 8 případy.
Grafy – Bodové grafy – vypneme Typ proložení – Proměnné X, Y - OK



Z dvourozměrného tečkového diagramu je patrno, že vhodnou regresní funkcí bude parabola: $m(x;\beta_0,\beta_1,\beta_2) = \beta_0 + \beta_1 x + \beta_2 x^2$.
K datovému souboru tedy přidáme novou proměnnou Xkv a do jejího Dlouhého jména napíšeme = X^2

	1 X	2 Y	3 Xkv
1	40	5,7	1600
2	50	5,4	2500
3	60	5,2	3600
4	70	5,2	4900
5	80	5,8	6400
6	90	6	8100
7	100	7,5	10000
8	110	8,1	12100

b) Vypočtěte odhady regresních parametrů a napište rovnici regresní paraboly.
Statistiky – Vícerozměrná regrese – Závisle proměnná Y, nezávisle proměnné X, Xkv - OK – OK – Výpočet: Výsledky regrese.

	Výsledky regrese se závislou proměnnou : Y (spotreba) R= ,98403165 R2= ,96831829 Upravené R2= ,95564561 F(2,5)=76,410 p<,00018 Směrod. chyba odhadu : ,22973					
N=8	Beta	Sm.chyba beta	B	Sm.chyba B	t(5)	Úroveň p
Abs.člen			9,751786	0,945689	10,31183	0,000148
X	-3,38045	0,602292	-0,150536	0,026821	-5,61264	0,002483
Xkv	4,22756	0,602292	0,001244	0,000177	7,01912	0,000905

Rovnice regresní paraboly: y = 9,751786 – 0,150536 x + 0,001244xkv

c) Najděte odhad rozptylu, vypočtěte index determinace a interpretujte ho.

Vrátíme se do Výsledky – vícenásobná regrese – Detailní výsledky – ANOVA.

	Analýza rozptylu (spotreba)				
Efekt	Součet čtverců	sv	Průměr čtverců	F	Úroveň p
Regres.	8,064881	2	4,032440	76,40988	0,000179
Rezid.	0,263869	5	0,052774		
Celk.	8,328750				

Odhad rozptylu najdeme na řádku Rezid., ve sloupci Průměr čtverců, tedy $s^2 = 0,05277$.
Index determinace je uveden v záhlaví původní výstupní tabulky pod označením R2. V našem případě $ID^2 = 0,9683$, tedy variabilita spotřeby benzínu je z 96,8% vysvětlena regresní parabolou.

d) Určete 95 % intervaly spolehlivosti pro regresní parametry.

Ve výstupní tabulce výsledků regrese přidáme za proměnnou Úroveň p dvě nové proměnné dm (pro dolní meze 95% intervalů spolehlivosti pro regresní parametry) a hm (pro horní meze 95% intervalů spolehlivosti pro regresní parametry). Do Dlouhého jména proměnné dm resp. hm napíšeme: =v3-v4*VStudent(0,975;5) resp. =v3+v4*VStudent(0,975;5)

N=8	Výsledky regrese se závislou proměnnou : Y (spotreba) R= ,98403165 R2= ,96831829 Upravené R2= ,95564561 F(2,5)=76,410 p<,00018 Směrod. chyba odhadu : ,22973 Beta	Sm.chyba beta	B	Sm.chyba B	t(5)	Úroveň p	dm =v3-v4*	hm =v3+v4
Abs.člen			9,751786	0,945689	10,31183	0,000148	7,320815	12,18276
X	-3,38045	0,602292	-0,150536	0,026821	-5,61264	0,002483	-0,21948	-0,08159
Xkv	4,22756	0,602292	0,001244	0,000177	7,01912	0,000905	0,000788	0,0017

Vidíme, že
$7{,}320815 < \beta_0 < 12{,}18276$ s pravděpodobností aspoň 0,95,
$-0{,}21948 < \beta_1 < -0{,}08159$ s pravděpodobností aspoň 0,95,
$0{,}000788 < \beta_2 < 0{,}0017$ s pravděpodobností aspoň 0,95

e) Na hladině významnosti 0,05 proveďte celkový F-test.
Testovou statistiku F-testu a odpovídající p-hodnotu najdeme v záhlaví výstupní tabulky regrese. Zde F = 76,41, p-hodnota < 0,00018, tedy na hladině významnosti 0,05 zamítáme hypotézu o nevýznamnosti modelu jako celku. (Výsledky F-testu jsou rovněž uvedeny v tabulce ANOVA.)

f) Na hladině významnosti 0,05 proveďte dílčí t-testy a vypočtěte relativní chyby odhadů regresních parametrů.
Výsledky dílčích t-testů jsou uvedeny ve výstupní tabulce regrese.
Testová statistika pro test hypotézy H_0: $\beta_0 = 0$ je 10,31183, p-hodnota je 0,000148. Hypotézu o nevýznamnosti parametru β_0 tedy zamítáme na hladině významnosti 0,05.
Testová statistika pro test hypotézy H_0: $\beta_1 = 0$ je -5,61264, p-hodnota je 0,002483. Hypotézu o nevýznamnosti parametru β_1 tedy zamítáme na hladině významnosti 0,05.
Testová statistika pro test hypotézy H_0: $\beta_2 = 0$ je 7,01912, p-hodnota je 0,000905. Hypotézu o nevýznamnosti parametru β_2 tedy zamítáme na hladině významnosti 0,05.
K upravené výstupní tabulce s mezemi intervalů spolehlivosti přidáme proměnnou chyba. Do jejího Dlouhého jména napíšeme
=100*abs(0,5*(hm-dm)/v3)

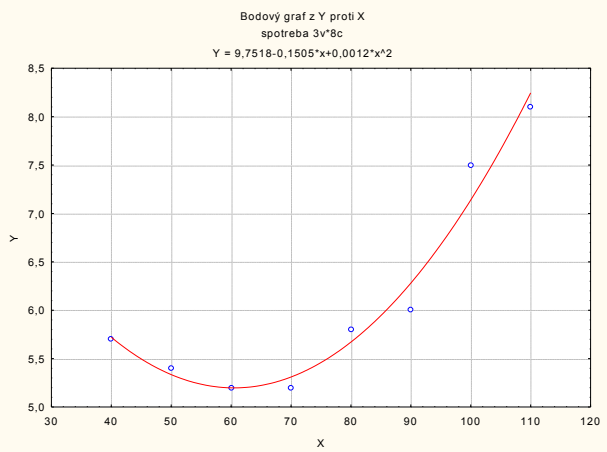
N=8	Výsledky regrese se závislou proměnnou : Y (spotreba) R= ,98403165 R2= ,96831829 Upravené R2= ,95564561 F(2,5)=76,410 p<,00018 Směrod. chyba odhadu : ,22973 Beta	Sm.chyba beta	B	Sm.chyba B	t(5)	Úroveň p	dm =v3-v4*	hm =v3+v4	chyba =100*a
Abs.člen			9,751786	0,945689	10,31183	0,000148	7,320815	12,18276	24,92847
X	-3,38045	0,602292	-0,150536	0,026821	-5,61264	0,002483	-0,21948	-0,08159	45,79987
Xkv	4,22756	0,602292	0,001244	0,000177	7,01912	0,000905	0,000788	0,0017	36,62259

Vidíme, že chyby odhadů jsou velké, v řádu desítek procent.

g) Určete regresní odhad spotřeby benzínu při rychlosti 80 km/h.
Pro výpočet predikované hodnoty zvolíme Rezidua/předpoklady/předpovědi - Předpovědi závisle proměnné X: 80, Xkv 6400 OK. Ve výstupní tabulce je hledaná hodnota označena jako Předpověď: 5,6708

h) Znázorněte data s proloženou regresní funkcí.

Do dvourozměrného tečkového diagramu nakreslíme regresní přímku tak, že v tabulce 2D Bodové grafy zvolíme na záložce Detaily Typ proložení: Polynomiální, OK. Stupeň polynomu je implicitně nastaven na 2, lze změnit na záložce Možnosti 2.



i) Vypočtěte střední absolutní procentuální chybu predikce (MAPE)

Ve výsledcích Vícenásobné regrese zvolíme záložku Rezidua/předpoklady/předpovědi – Reziduální analýza – Uložit – Uložit rezidua a předpovědi – Vybrat X, Y – OK. Ve vzniklé tabulce odstraníme proměnné 5 – 10, přidáme proměnnou chyby a do jejího Dlouhého jména napíšeme
=100*abs(v4/v2)
Pak spočteme průměr této proměnné a zjistíme, že MAPE = 2,15%.