SEMINÁŘ VÝPOČETNÍ STATISTIKY C5

2008-04-21

REGRESE:

Regresní funkce:

- ✓ Není jen: $y' = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_n x_n$
- ✓ Ale také: $y' = a_0 + a_1 f_1(x) + a_2 f_2(x) + \ldots + a_n f_n(x)$

Jednoduchý regresní model:

- ✓ $y' = a + bx$
- ✓ $y' = a_0 + a_1 x_1$

Příklad:

- ✓ Průzkum mobilních operátorů – závislost počtu odeslaných zpráv na ceně
- ✓ Vstupní data:

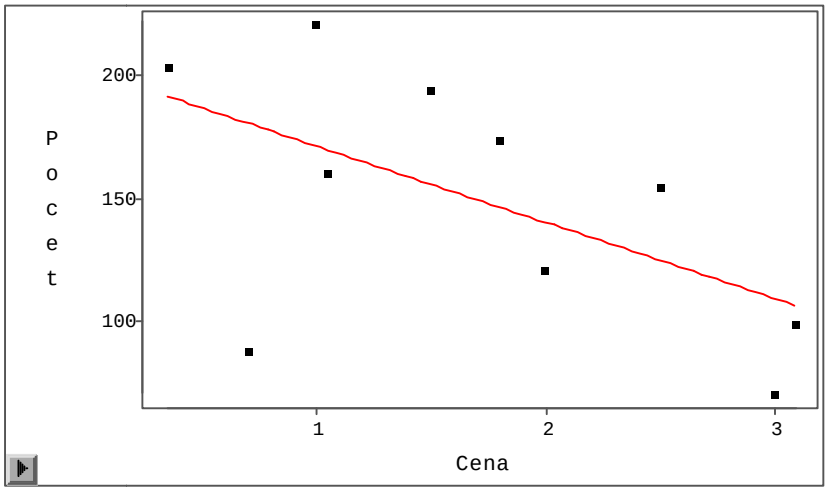
WORK.A

	Cena	Pocet
1	0.35	205
2	0.70	89
3	1.00	223
4	1.05	162
5	1.50	196
6	1.80	175
7	2.00	122
8	2.50	156
9	3.00	71
10	3.10	100

- ✓ Analyze -> Fit (Y X) – Cena je X, počet Y.
- ✓ Výstup:

```
Pocet        =        Cena
Response Distribution:  Normal
Link Function:      Identity
```

```
          Model Equation
Pocet   =   203.024   -   31.2492  Cena
```



Parametric Regression Fit

		Model		Error				
Curve	Degree(Polynomial)	DF	Mean Square	DF	Mean Square	R-Square	F Stat	Pr > F
	1	1	7973.2422	8	2068.4572	0.3252	3.85	0.0852

Summary of Fit

Mean of Response	149.9000	R-Square	0.3252
Root MSE	45.4803	Adj R-Sq	0.2408

Analysis of Variance

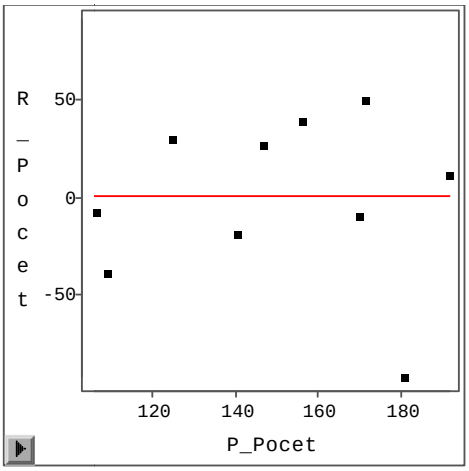
Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
Model	1	7973.2422	7973.2422	3.85	0.0852
Error	8	16547.6578	2068.4572		
C Total	9	24520.9000			

Type III Tests

Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
Cena	1	7973.2422	7973.2422	3.85	0.0852

Parameter Estimates

Variable	DF	Estimate	Std Error	t Stat	Pr >|t|	Tolerance	Var Inflation
Intercept	1	203.0237	30.6427	6.63	0.0002	.	0
Cena	1	-31.2492	15.9164	-1.96	0.0852	1.0000	1.0000



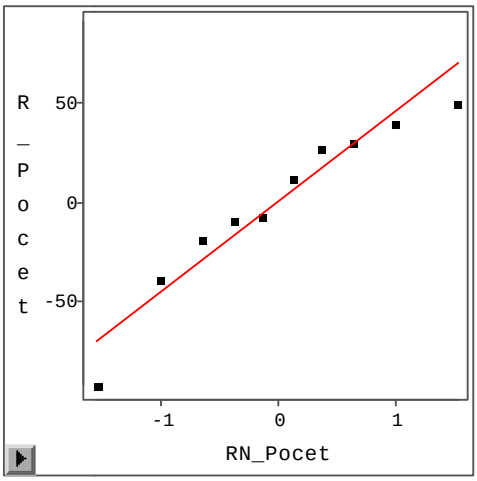
✓ **Hodnocení:**

- Jsou-li hodnoty u sebe a je možné je proložit přímkou, je závislost přímá. Zde závislost nepřímá, střední.
- Pomocí jezdce lze proložit polynomem x-tého stupně a rezidua jsou pak minimální. Větší polynom bychom neměli používat. Existuje reziduální náhodná složka, kterou nejsme schopni matematicky postihnout. Pokud použijeme vyšší polynom, říkáme, že rezidua mají velkou váhu pro model a zjistíme, jak se dělají odchylky. My však chceme popsat skutečnost, ne chyby měření
- R-square – koeficient determinace – z kolika % vysvětlující proměnné vysvětlují vysvětlovanou. Popsaná závislost se dá cenou vysvětlit ze 32,5%. Odmocninou je koeficient korelace, znaménko se určí klesající přímkou => záporné.
- Místo procedury glm je tabulka Analysis of Variance.
- Odhady parametrů (Parameter Extimates) – lze otestovat významnost parametrů, v 1. řádku regresní konstanta, parametr b je směrnice přímky a významný není, takže nemůže být významný model
- Poslední je **graf reziduí** – reziduum je odchylka: $e_i = y_i - y'_i$, rezidua mají být co nejmenší, např. součet metodou nejmenších čtverců. Fungují na základě měření dvou a více veličin, reziduální graf lze nakreslit vždy. Rezidua jsou náhodné veličiny, které mají normální rozdělení s nulovou střední hodnotou a konstantním rozptylem. Hodnoty by se měly pohybovat kolem střední hodnoty, která je nulová, konstantnost je daná tak, že nemůžeme najít tendenci.

  Možnosti:
  - Zde je nulová střední hodnota, protože jsou odchylky nad a pod, ale konstantní rozptyl ne, jedno pozorování se výrazně odchyluje, mohlo by se jednat o chybu měření.

- Může být také megafonový efekt, kdy jsou podmínky čím dál horší.
- Mohou také být dva hloučky, což mohou způsobit změny podmínek.
- Rezidua jsou sdružená, odchylky jsou minimální, ale například jsou rostoucí, mají trend.

✓ **Ověření normality:** Graphs -> Residual normal QQ:

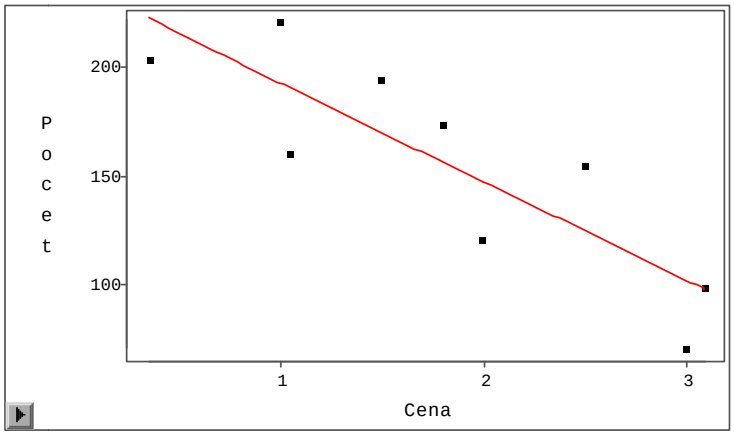


✓ **Odlehlé pozorování:**

- Můžeme ho vypustit, ale nemusí se jednat o chybu.
- Máme malý počet měření a malý počet vysvětlujících funkcí.
- Horší chyba – měření je málo, ačkoli je jich cca 2000, ale tarifů je pouze 10. Nutné provést více měření třeba v různých časech.
- Smazání pozorování: Edit -> Delete:

Pocet = Cena	
Response Distribution:	Normal
Link Function:	Identity

Model Equation				
Pocet	=	239.551	-	45.7644 Cena



Parametric Regression Fit

		Model		Error				
Curve	Degree(Polynomial)	DF	Mean Square	DF	Mean Square	R-Square	F Stat	Pr > F
	1	1	14773.4961	7	803.7863	0.7242	18.38	0.0036

Summary of Fit

Mean of Response	156.6667	R-Square	0.7242
Root MSE	28.3511	Adj R-Sq	0.6848

Analysis of Variance

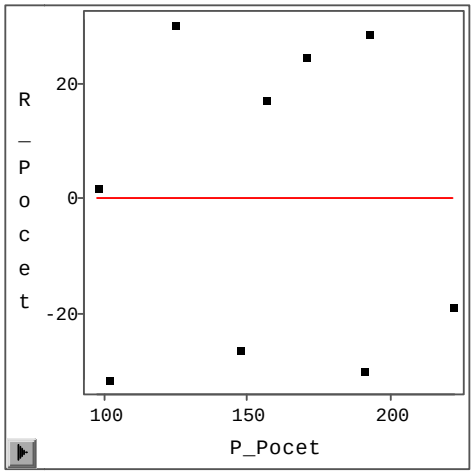
Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
Model	1	14773.4961	14773.4961	18.38	0.0036
Error	7	5626.5039	803.7863		
C Total	8	20400.0000			

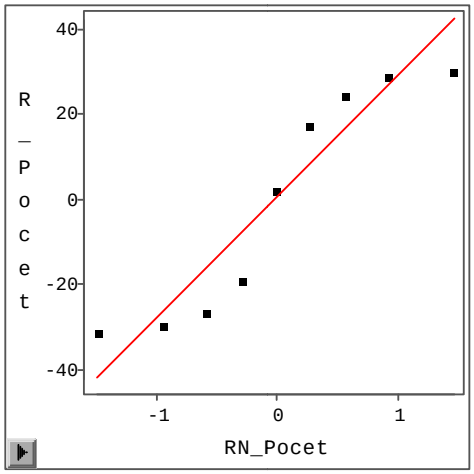
Type III Tests

Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
Cena	1	14773.4961	14773.4961	18.38	0.0036

Parameter Estimates

Variable	DF	Estimate	Std Error	t Stat	Pr >|t|	Tolerance	Var Inflation
Intercept	1	239.5510	21.5192	11.13	<.0001	.	0
Cena	1	-45.7644	10.6747	-4.29	0.0036	1.0000	1.0000





- Nyní je model dostačující, determinace je přes 75%. Byla změněna rovnice, směrnice je -45,76, při nakreslení ve skutečném měřítku by byla přímka na malém úseku téměř strmá (kolmá), jelikož na osách grafu jsou jednotky a stovky.
- Pokud by se nám nezdálo další pozorování, není možné pozorování vyloučit, protože nelze upravovat upravený model.

✓ **Typy divností:**

- **Odlehlost** – pozorování je odlehlé
- **Vlivnost** – pozorování má vliv na tvar (sklon) přímky, ale u vícenásobných modelů není možné vlivnost takto poznat.

✓ **Regresní diagnostika:**

- **Procedura:**

```
proc reg data=svs;
model pocet = cena;
run;
```

- **Výstup:**

```
The SAS System          08:51 Monday, April 21, 2008   1

                        The REG Procedure
                          Model: MODEL1
                    Dependent Variable: Pocet

                Number of Observations Read          10
                Number of Observations Used          10

                        Analysis of Variance

                                  Sum of           Mean
Source                    DF     Squares         Square    F Value    Pr > F

Model                      1     7973.24219     7973.24219       3.85    0.0852
Error                      8       16548        2068.45723
```

```
Corrected Total          9          24521


                Root MSE            45.48029    R-Square     0.3252
                Dependent Mean     149.90000    Adj R-Sq     0.2408
                Coeff Var           30.34042


                          Parameter Estimates

                                Parameter       Standard
          Variable     DF       Estimate          Error    t Value    Pr > |t|

          Intercept     1      203.02370       30.64271       6.63      0.0002
          Cena          1      -31.24923       15.91641      -1.96      0.0852
```

- Lze vyčíst rovnici: $y' = 203{,}02 - 31{,}25x$
- **Procedura diagnostiky:**

```
proc reg data=svs;
model pocet = cena/r influence;
run;
```

- **Výstup:**

```
                          Model: MODEL1
                     Dependent Variable: Pocet

                          Output Statistics

    Dependent Predicted    Std Error           Std Error  Student                    Cook's
Obs  Variable     Value Mean Predict  Residual  Residual Residual   -2-1 0 1 2           D

  1  205.0000  192.0865      25.8562   12.9135   37.415    0.345 |       |       |    0.028
  2   89.0000  181.1492      21.4518  -92.1492   40.103   -2.298 |  ****|       |    0.755
  3  223.0000  171.7745      18.1928   51.2255   41.683    1.229 |       |**     |    0.144
  4  162.0000  170.2120      17.7166   -8.2120   41.888   -0.196 |       |       |    0.003
  5  196.0000  156.1498      14.7302   39.8502   43.029    0.926 |       |*      |    0.050
  6  175.0000  146.7751      14.4699   28.2249   43.117    0.655 |       |*      |    0.024
  7  122.0000  140.5252      15.1541  -18.5252   42.881   -0.432 |       |       |    0.012
  8  156.0000  124.9006      19.2088   31.0994   41.225    0.754 |       |*      |    0.062
  9   71.0000  109.2760      25.1988  -38.2760   37.861   -1.011 |     **|       |    0.226
 10  100.0000  106.1511      26.5213   -6.1511   36.947   -0.166 |       |       |    0.007

                          Output Statistics

                    Hat Diag         Cov                 -------DFBETAS-------
Obs  RStudent            H       Ratio      DFFITS     Intercept        Cena

  1    0.3253      0.3232      1.8728      0.2248       0.2236     -0.1868
  2   -3.6861      0.2225      0.1942     -1.9717      -1.9122      1.4629
  3    1.2763      0.1600      1.0233      0.5571       0.5079     -0.3412
  4   -0.1838      0.1517      1.5250     -0.0778      -0.0697      0.0454
  5    0.9169      0.1049      1.1631      0.3139       0.2037     -0.0678
  6    0.6294      0.1012      1.3017      0.2112       0.0780      0.0232
  7   -0.4089      0.1110      1.4015     -0.1445      -0.0242     -0.0455
  8    0.7322      0.1784      1.3716      0.3412      -0.0798      0.2262
  9   -1.0125      0.3070      1.4339     -0.6739       0.3081     -0.5534
 10   -0.1560      0.3400      1.9654     -0.1120       0.0546     -0.0941

            Sum of Residuals                           0
            Sum of Squared Residuals               16548
          Predicted Residual SS (PRESS)            26195
```

- Odlehlost:
  - *Studentizovaná rezidua* – Student Residual, je-li |SR|>2, jedná se o outliers, tedy odlehlé pozorování. Pozorování číslo 2 je určitě odlehlé. Totéž zobrazují hvězdičky.
  - *Leverage* – projekční matice, tzv. klobouková:

$$\hat{H} = \begin{pmatrix} \ldots & & \\ & h_{ii} & \\ & & \ldots \end{pmatrix}$$

Na diagonále hii: $h_{ii} > 2\frac{p}{n}$

p – počet všech funkcí, které se zúčastňují, počet regresních parametrů je vždy větší, jelikož je funkce $y' = a_0 + a_1x_1 + a_2x_2 + \ldots + a_nx_n$, kde je n+1, tím jedním je a_0.

$$h_{ii} > 2\frac{p}{n} = 2\frac{2}{10}$$

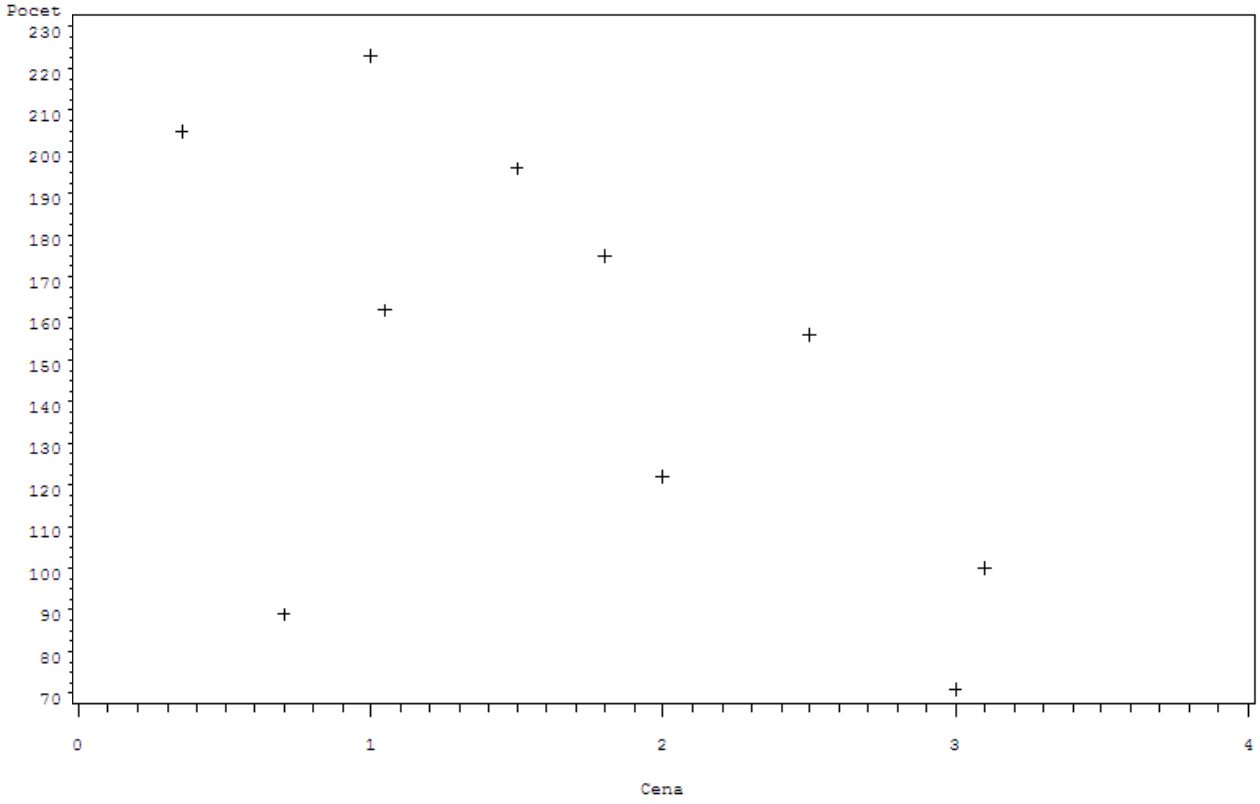
- Vlivnost:
  - *Cook's D* – je-li $D > \frac{4}{n} = \frac{4}{10} = 0{,}4$, jak se ovlivní celý model
  - *Welsch-Kuhova vzdálenost* – DFFITS: $|\text{DFFITS}| > 2\sqrt{\frac{p}{n}}$, jak pozorování změní vyrovnanou konkrétní hodnotu

✓ **Graf korelačního pole:**

- **Procedura:**

```
proc gplot data=svs;
plot pocet*cena;
run;
```

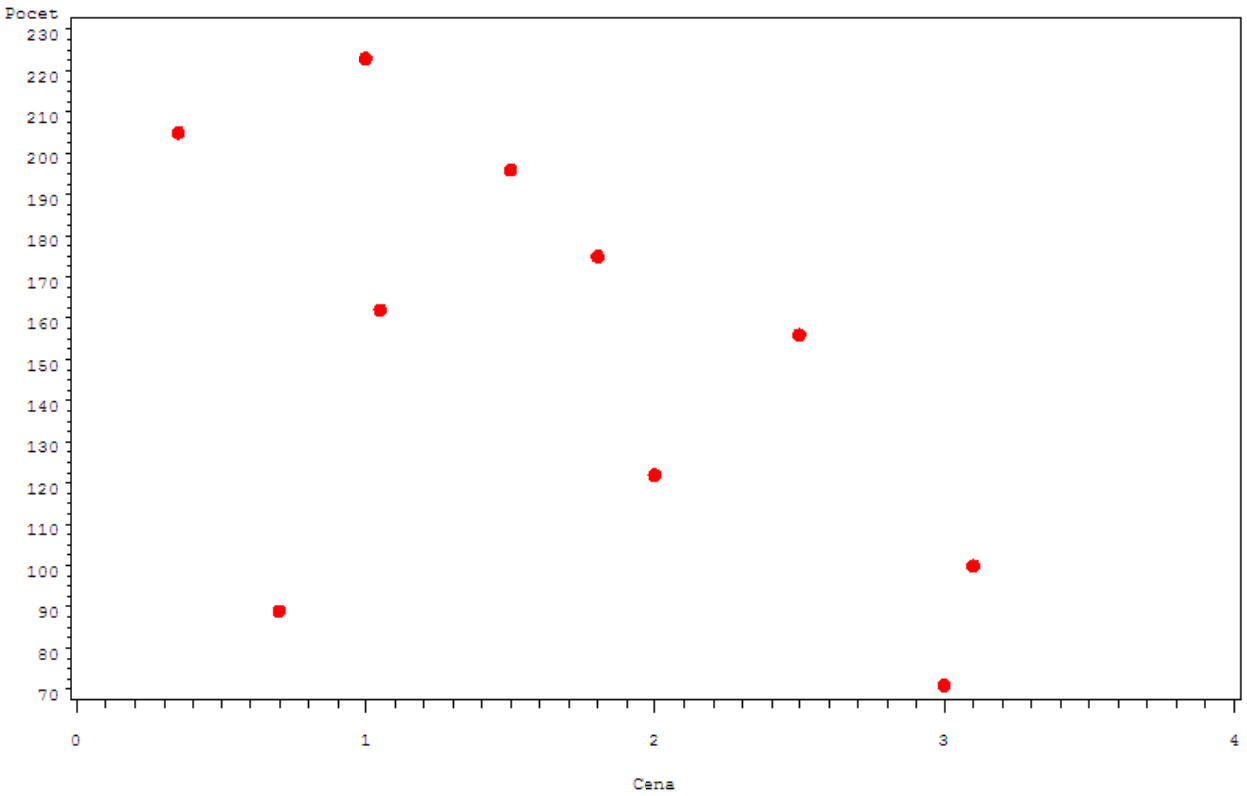
- **Výstup:**



- **Jiné znaky a barva bodů:**

```
proc gplot data=svs;
plot pocet*cena;
symbol v=dot c=red;
run;
```

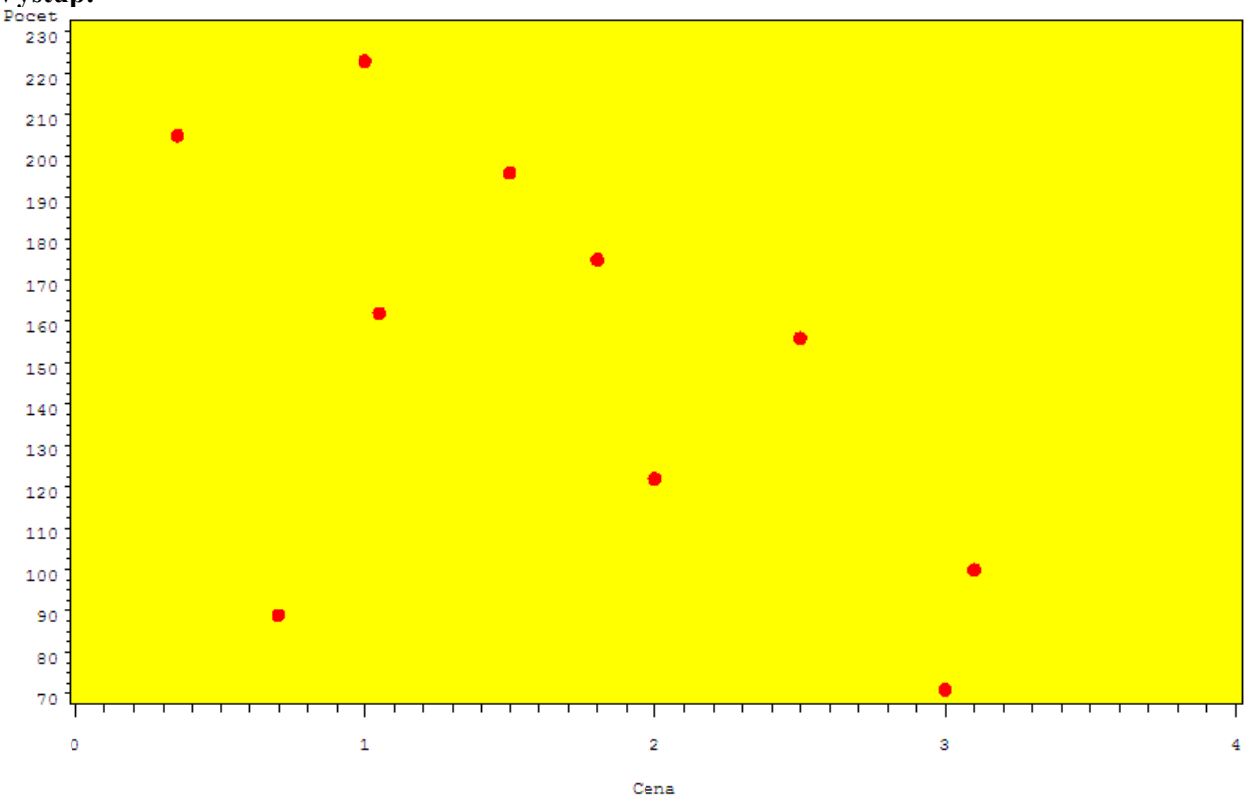
- **Výstup:**



- **Jiná barva pozadí**

```
proc gplot data=svs;
plot pocet*cena/cframe*yellow;
symbol v=dot c=red;
run;
```

- **Výstup:**



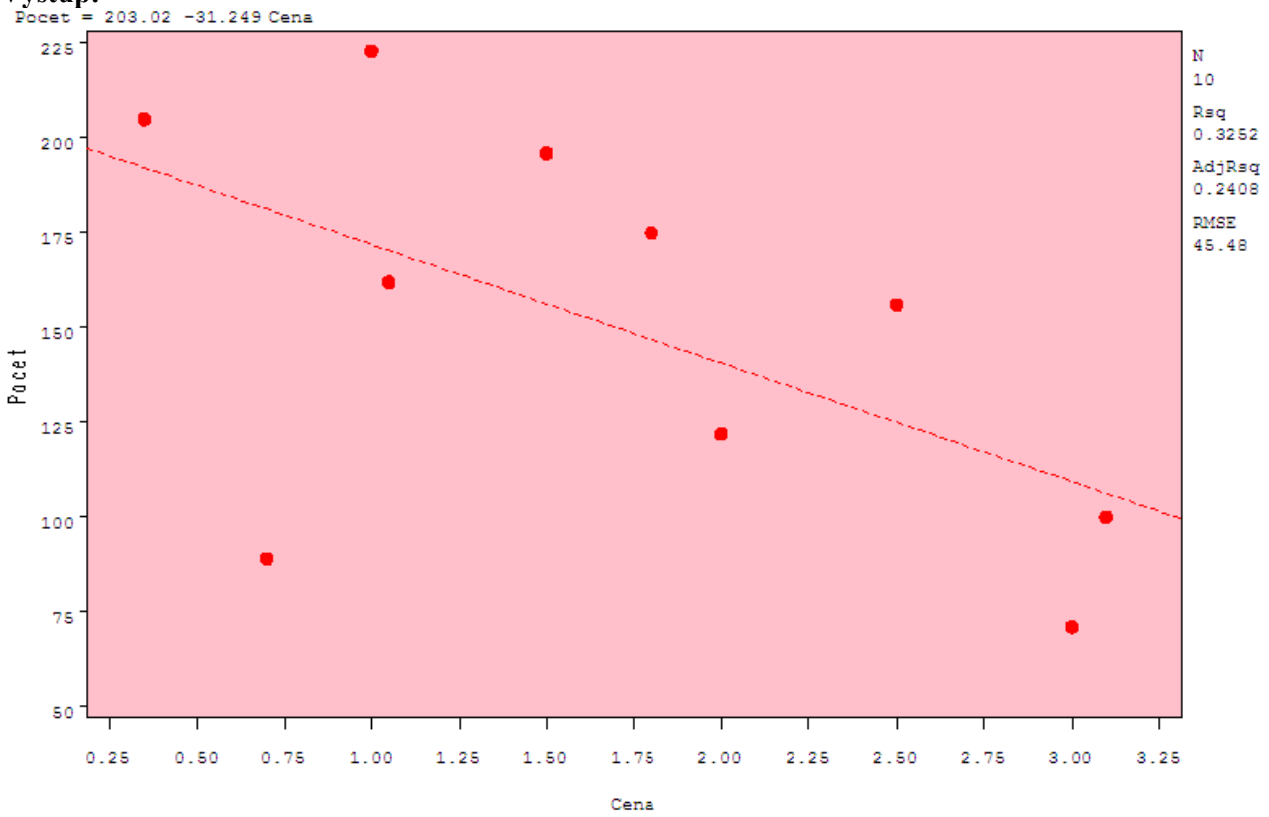
✓ **Diagnostika:**

- **Procedura reg:**

```
proc reg data=svs;
model pocet=cena/r influence;
```

```
plot pocet*cena/cframe=pink;
run;
```

- **Výstup:**



  - Zakreslení grafu v proceduře zobrazí i přímku, což je lepší než grafická procedura, vypíše se také rovnice
  - RMSE – ukazatel kvality chyby, čím chyba menší, tím model kvalitnější.

✓ **Cookova vzdálenost:**

- **Procedura:**

```
proc reg data=svs;
model pocet*cena/r influence;
plot pocet*cena/cframe=pink;
symbol v*dot c*green h=2;
plot cookd.*p./cframe=grey;
run;
```

✓ **Rezidua:**

- **Procedura:**

```
proc reg data=svs;
model pocet*cena/r influence;
plot pocet*cena/cframe=pink;
symbol v*dot c*green h*2;
plot cookd.*p./cframe=yellow;
plot r.*p./cframe=ligr;
run;
```