SEMINÁŘ VÝPOČETNÍ STATISTIKY P9

2008-04-14

VÍCENÁSOBNÁ KORELACE A REGRESE:

Regrese a korelace:

- ✓ **Jednoduchá:** Y, X
- ✓ **Vícenásobná:** Y, X_1, X_2, …, X_k

Příklad:

Student	Hodiny	IQ	Body
1	9	99	56
2	6	100	45
3	12	119	80
4	14	95	73
5	11	110	71
6	6	117	55
7	19	98	95
8	16	101	86
9	3	100	34
10	9	115	66

Zkoumáme:

- ✓ Hledáme **regresní funkci** ve tvaru $y' = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n$. Koeficientu b0 říkáme absolutní člen (intercept), koeficienty b_1 až b_k jsou parciální (dílčí) regresní koeficienty.
- ✓ **Koeficient mnohonásobné korelace**, značíme ho velkým R, interval $0 \leq R \leq 1$, stupnice stejná jako u jednoduché. Vynásobíme-li ho 100, říká, z kolika procent je veličina Y vysvětlována (ovlivňována) nezávislými veličinami X_1 až X_k
- ✓ **Koeficient mnohonásobné determinace** – R^2.

Předpoklady použitelnosti mnohonásobné regrese a korelace:

- ✓ **Normalita** rozdělení analyzovaných proměnných.
- ✓ Vysvětlující proměnné by měly být navzájem **nezávislé**, pokud by byly závislé, opakovaly by hodnoty jedné veličiny hodnoty druhé veličiny a nedozvěděli bychom se nic nového. V reálném světě bude závislost vždy, ale chceme, aby byla co nejmenší.

  Ověření nezávislosti:

  - **Výpočet korelační matice:**
    - Obecný tvar korelační matice:

	X_1	X_2	…	X_k
X_1	1	r_{x1x2}	…	r_{x1xk}
X_2		1		
…			…	
X_k				1

    - Na diagonále jsou jedničky, neboť korelace mezi jednou a tou samou je maximálně silná
    - Na ostatních pozicích jsou korelační koeficienty
    - Pokud jsou korelační koeficienty v absolutní hodnotě menší než 0,75 ($\left|r_{x_jx_k}\right| < 0{,}75$), považujeme podmínku nezávislosti za splněnou.
    - Je-li hodnota překročena, vzniká zde multikolinearita, předpoklad nezávislosti byl tedy porušen
  - **VIF – Variance Inflation Factor**
    - Faktor zvětšení rozptylu
    - Pokud VIF>10, znamená to multikolinearitu, tedy nežádoucí stav.

- ✓ **Rezidua**, tzn. rozdíly $y_i - y_i'$, i = 1, 2, …, n, by měla mít normální rozdělení s nulovou střední hodnotou a konstantním rozptylem. Nulová střední hodnota vyjadřuje, že model je umístěn natolik ideálně, že představuje osu korelačního pole.

SAS:

- ✓ **Posouzení normality rozdělení analyzovaných proměnných:**

```
ods exclude Moments BasicMesures TestsForLocation Quantiles ExtremeObs;
proc univariate data=svs normal plot;
run;
quit;
```

- ✓ Máme-li pouze 10 pozorování, je testování normality riskantní, testy fungují přesně až u větších datových souborů (od 30 pozorování výše). Testy většinou řeknou, že data mají normální rozdělení.
- ✓ **Výpočet korelační matice všech sledovaných proměnných:**
  - **Procedura:**

```
proc corr data=svs pearson spearman;
run;
quit;
```

  - **Příkazy:**
    - Pearson - spočítá tradiční korelační koeficienty, funguje, pokud je normální rozdělení a neobjevují se problematické údaje, pokud by normalita nebyla splněna, je lepší použít koeficienty pořadové korelace
    - Spearman – koeficienty pořadové korelace, Spearmanovy koeficienty, je neparametrický, takže je univerzálnější, což je zaplaceno menší kvalitou (přesností).
    - Příkazy psát nemusíme, pokud je nenapíšeme, vypočítají se pouze pearsonovy koeficienty.
    - Před příkaz run lze vsunout příkaz var (variables) a jména proměnných, výsledek by však byl stejný, pokud příkaz neuvedeme, jelikož automaticky se počítá pro všechny proměnné.
- ✓ **Model mnohonásobné regrese, doplněný o korelační matici všech proměnných a o regresní diagnostiku:**
  - **Procedura:**

```
proc reg data=svs corr;
model body=hodiny IQ/r influence vif spec;
plot r.*p;
plot cookd.*p;
symbol v=dot c=green;
output out=diag r=rezid;
run;
quit;
```

  - **Příkazy:**
    - Procedura corr nabízí detailnější výstupy než pouze reg.
    - Model – píše se jméno vysvětlované proměnné a za rovná se se píší vysvětlující proměnné. Lomítkem naznačujeme, že budeme vypisovat další doplňující příkazy, tedy regresní diagnostiku. Ta slouží ke kontrole jednotlivých předpokladů
    - r influence – počítá rezidua a modifikace reziduí, spočítají se také studentizovaná rezidua, která využijeme k tomu, abychom posoudili, zda ve vstupních datech nejsou nějaké problémy. Bereme absolutní hodnotu studentizovaných reziduí a porovnáváme jí s hodnotou 2: |SR| > 2, pokud je hodnota větší, značí to odlehlé pozorování (outlier).
    - Odlehlost znamená, že údaj se vychýlil z množiny Y podstatným způsobem, ale odlehlost nemusí znamenat negativní dopad na výsledný model, je nutné zjistit, zda pozorování je vlivné, tedy zda má vliv na výsledný model => Cookova vzdálenost (D), $D_i > \frac{4}{n}$ - pozorování je vlivné.
    - Hledáme také odlehlou hodnotu mezi proměnnými X, charakteristika leverage umožňuje zhodnotit, zda v množině vysvětlujících proměnných není odlehlý údaj, $h_{ii} = 2\frac{p}{n}$, p je počet parametrů regresního modelu, tedy b_0, b_1 atd., u příkladu výše jsou 3. Pokud je hranice překročena, je obsažena problematická hodnota.

- DFFITS – charakteristika nazývaná Velschova-Kuhova vzdálenost, slouží k posouzení vlivnosti. Může být kladná i záporná, bereme jí tedy v absolutní hodnotě: $|DFFITS| = 2\sqrt{\frac{p}{n}}$, pokud je hodnota překročena, je pozorování vlivné.
- Rozdíl mezi Cookovou a Velschvou-Kuhovou vzdáleností:
  - Cookova je obecnější, říká, jak je ovlivněn celý model.
  - Velschova-Kuhova vzdálenost – do jaké míry ovlivňuje pozorování jednu hodnotu veličiny Y, hodnotu, u které byla spočtena.
- Spec – počítá tzv. Whiteův test, který hodnotí jednu z důležitých vlastností reziduí, umožňuje posoudit, zda rezidua mají konstantní rozptyl.
- plot r.*p – konstrukce reziduálního grafu, lze posoudit, zda jsou vlastnosti splněny a model je kvalitní.
- plot cookd.*p – grafické zobrazení Cookových hodnot.
- symbol v=dot c=green – provedení grafů, v=dot znamená, že body budou jako tečky, barva bude zelená.
- output out=diag r=rezid – vytvoříme pomocný soubor nazvaný podle slova diagnostika „diag“, bude obsahovat jedinou proměnnou, tedy „rezid“, která je tvořen a reziduálními proměnnými.

✓ **Testování normality reziduí a ověření, zda mají nulovou střední hodnotu:**

- **Procedura:**

```
ods exclude Moments Quantiles ExtremeObs;
proc univariate data=diag normal plot;
var rezid;
run;
quit;
```

- **Popis:**
  - Využíváme soubor diag a zaměřujeme se na rezidua

✓ **Výstupy:**

- **Výstup:**

Procedura CORR

3 Proměnné: hodiny IQ body

Jednoduché statistiky

Proměnná	N	Průměr	Std odch	Medián	Minimum	Maximum
hodiny	10	10.50000	4.92725	10.00000	3.00000	19.00000
IQ	10	105.40000	8.90942	100.50000	95.00000	119.00000
body	10	66.10000	18.84705	68.50000	34.00000	95.00000

Pearsonovy korelační koeficienty, N = 10
Prob > |r| pro H0: Rho=0

	hodiny	IQ	body
hodiny	1.00000	-0.23539 0.5127	0.96138 <.0001
IQ	-0.23539 0.5127	1.00000	0.04010 0.9124
body	0.96138 <.0001	0.04010 0.9124	1.00000

Spearmanovy korelační koeficienty, N = 10
Prob > |r| pro H0: Rho=0

	hodiny	IQ	body
hodiny	1.00000	-0.25382 0.4792	0.98173 <.0001
IQ	-0.25382 0.4792	1.00000	-0.08511 0.8152
body	0.98173 <.0001	-0.08511 0.8152	1.00000

- **Procedura CORR** – korelační matice, ve druhém řádku jsou p-hodnoty, které říkají, do jaké míry vypočtené koeficienty jsou statisticky významné, tedy zobecnitelné pro základní soubor, jsou-li menší než 5%, je koeficient statisticky významný, tedy zobecnitelný. Pearsonovy i Spearmanovy korelační koeficienty jsou podobné, nepotvrdila se domněnka ohledně normálního rozdělení. Chceme posoudit, zda jsou hodnoty mezi sebou vázány.
- **Výstup:**

Procedura REG

Number of Observations Read	10
Number of Observations Used	10

Korelace

Proměnná	hodiny	IQ	body
Hodiny	1.0000	-0.2354	0.9614
IQ	-0.2354	1.0000	0.0401
body	0.9614	0.0401	1.0000

Procedura REG
Model: MODEL1
Závislá proměnná: body

Number of Observations Read	10
Number of Observations Used	10

Analýza rozptylu

Zdroj	DF	Součet čtverců	Průměr Kvadrát	F hodnota	Pr > F
Model	2	3194.93279	1597.46639	5684.32	<.0001
Chyba	7	1.96721	0.28103		
Korigovaný součet	9	3196.90000			

Odmocnina MSE	0.53012	R-kvadrát	0.9994
Závislý průměr	66.10000	Přizp R-kv	0.9992
Koef prom	0.80200		

Odhady parametrů

Proměnná	DF	Odhad parametru	Standardní chyba	t hodnota	Pr > |t|	Inflace proměnné
Regresní	1	-38.05965	2.27970	-16.70	<.0001	0
Hodiny	1	3.93127	0.03690	106.54	<.0001	1.05866
IQ	1	0.59660	0.02041	29.23	<.0001	1.05866

Procedura REG
Model: MODEL1
Závislá proměnná: body

Test první a druhé specifikace momentu

DF	Chí-kvadrát	Pr > ChíKv
5	3.87	0.5681

- **Procedura REG:**
  - R-kvadrát je vysoký, bodový výsledek je popsán časem a IQ
  - Individuální p-hodnoty (Pr>t) hodnotí zobecnitelnost každého koeficientu, všechny jsou statisticky významné
  - Inflace proměnné = VIF – jsou malé, není přítomna multikolinearita
  - Test první a druhé specifikace momentu – Whiteův test, p-hodnota menší než 5% => konstantnost rozptylů není
- **Další výstupy:**

Procedura UNIVARIATE
Proměnná: rezid (Reziduum)

Základní statistické míry

Poloha		Variabilita	
Průměr	0.00000	Std odchylka	0.46752
Medián	-0.10582	Rozptyl	0.21858
Modus	.	Rozpětí	1.55787
		Mezikvartilové rozpětí	0.51980

Testy polohy: Mu0=0

Test	-Statistika-		----p hodnota-----	
Studentovo t	t	0	Pr > |t|	1.0000
Znaménko	M	-1	Pr >= |M|	0.7539
Znam. pořadí	S	-3.5	Pr >= |S|	0.7695

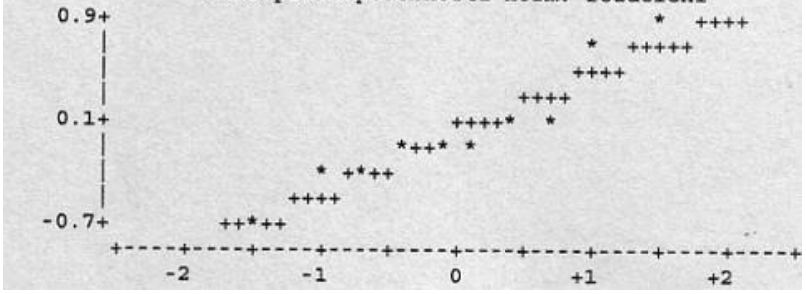
Testy normality

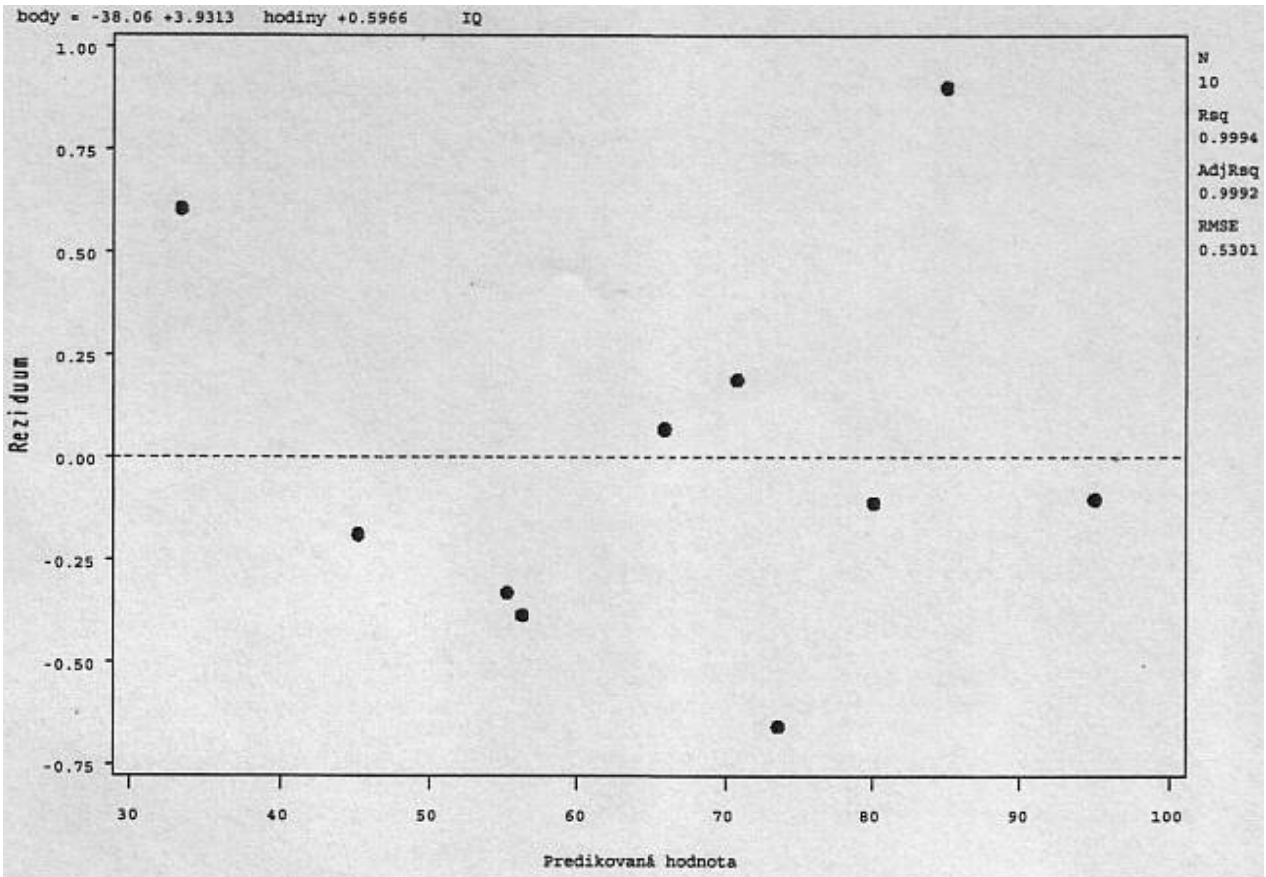
Test	--Statistika--		----p hodnota-----	
Shapiro-Wilk	W	0.944629	Pr < W	0.6056
Kolmogorov-Smirnov	D	0.185534	Pr > D	>0.1500
Cramer-von Mises	W-Kv	0.052778	Pr > W-Kv	>0.2500
Anderson-Darling	A-Kv	0.31207	Pr > A-Kv	>0.2500

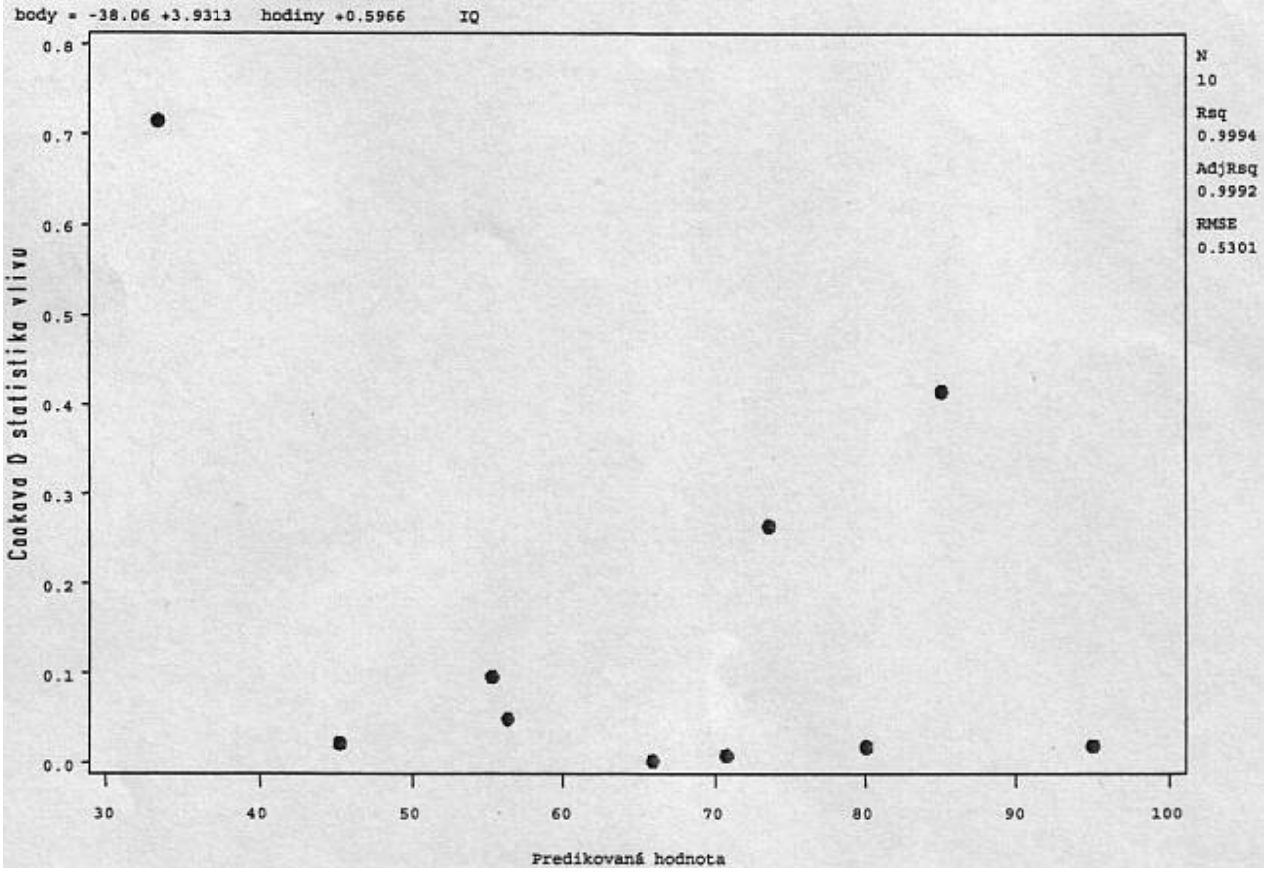
```
Kmen List                    #        Krb.graf
   8 0                       1           |
   6 1                       1           |
   4                                     |
   2                                     |
   0 79                      2         +---+---+
  -0 910                     3         *-------*
  -2 83                      2         +-------+
  -4                                     |
  -6 5                       1           |
     ----+----+----+----+
Násobit listy větve číslem 10**-1
```

Procedura UNIVARIATE
Proměnná: rezid (Reziduum)

Graf pravděpodobnosti norm. rozdělení




body = -38.06 +3.9313 hodiny +0.5966 IQ
Reziduum
1.00
0.75
0.50
0.25
0.00
-0.25
-0.50
-0.75
30
40
50
60
70
80
90
100
Predikovaná hodnota
N
10
Rsq
0.9994
AdjRsq
0.9992
RMSE
0.5301


body = -38.06 +3.9313 hodiny +0.5966 IQ
Cookova D statistika vlivu
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
30
40
50
60
70
80
90
100
Predikovaná hodnota
N
10
Rsq
0.9994
AdjRsq
0.9992
RMSE
0.5301