7

KORELACE

Vzorová úloha 7.1 *Postup vyšetření korelace*

Jako vzorovou použijeme **Úlohu B7.05** *Obsah dehtu, nikotinu a CO v cigaretách* se zadáním: Federální komise obchodu USA posuzuje domácí cigarety dle obsahu dehtu x_1 [mg], nikotinu x_2 [mg] a hmotnosti cigarety x_3 [g] a konečně i obsahu oxidu uhelnatého CO x_4 [mg] v uvolněném cigaretovém kouři. Hlavní hygienik USA totiž považuje faktory x_1, x_2 a x_4 za vysoce nebezpečné pro zdraví člověka. Poslední studie ukázaly, že zvyšující se obsah dehtu a nikotinu spolu nesou i zvýšení obsahu oxidu uhelnatého. Vyšetřete, zda existuje na hladině významnosti α = 0.05 korelace mezi proměnnými (a) x_1 a x_4, dále (b) x_2 a x_4, a (c) x_3 a x_4.

Řešení:

1. Návrh modelu: zařadíme i absolutní člen β_0 a nejprve budeme uvažovat lineární regresní model ve tvaru $x_4 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$.

Proměnná	**Průměr**	**Směrodatná odchylka**	**Pearsonův pár. korel. koeficient**	**Spočtená hladina význ.**	**Spearmanův pár. korel. koeficient**	**Spočtená hladina význ.**
x_1	12.216	5.6658	0.9575	0.0000	0.9448	0.0000
x_2	0.8764	0.3541	0.9259	0.0000	0.8778	0.0000
x_3	0.9703	0.0877	0.4640	0.0195	0.2170	0.2975
x_4	12.528	4.7397	1.0000	-----	1.0000	-----
Vícenásobný korelační koeficient r				: 0.95843		
Koeficient determinace 100 % D				: 91.859		
Predikovaný koeficient determinace D_P				: 0.91326		

Polohu a proměnlivost proměnných x_4, x_1, x_2, x_3 přináší *průměr* a *směrodatná odchylka* hodnot každé proměnné. *Pearsonův párový korelační koeficient* x_4 vs. x_1, x_4 vs. x_2, x_4 vs. x_3 ukazuje na vysokou korelaci, dvě nezávislé proměnné x_1, x_2 jsou se závisle proměnnou x_4 spjaty silnou lineární závislostí, zatímco poslední nezávisle proměnná x_3 slabou. Analogické závěry vyplývají i ze Spearmanova pořadového korelačního koeficientu. *Pearsonův vícenásobný korelační koeficient r* ukazuje, že navržený lineární regresní model je statisticky významný a jeho druhá mocnina, nazvaná *koeficient determinace* $D = r^2$ ukazuje, že 91.86 % bodů dobře koresponduje s modelem. *Predikovaný koeficient determinace* D_p má podobný význam jako *koeficient determinace D*, je však vyčíslen jinak, místo *RSC* se ve vztahu užije *MEP*.

2. Pearsonova korelační matice: vyčíslí se Pearsonovy párové korelační koeficienty parametrického charakteru a Cronbachův korelační koeficient γ.

Pearsonovy párové korelační koeficienty mezi dvojicemi vysvětlujících proměnných		**Spočtená hladina významnosti**
x_1 versus x_2:	0.97661	0.0000
x_1 versus x_3:	0.49077	0.0127
x_1 versus x_4:	0.95749	0.0000
x_2 versus x_3:	0.50018	0.0109
x_2 versus x_4:	0.92595	0.0000
x_3 versus x_4:	0.46396	0.0195
Cronbachův korelační koeficient γ		: 0.6939
Standardizovaný Cronbachův korelační koeficient γ		: 0.9111

Pearsonův párový korelační koeficient x_4 vs. x_1, x_4 vs. x_2, x_4 vs. x_3 ukazuje na vysokou korelaci, dvě nezávislé proměnné x_1, x_2, jsou se závisle proměnnou x_4 spjaty silnou lineární závislostí, zatímco poslední nezávisle proměnná x_3 slabou korelací.

3. Spearmanova korelační matice: vyčíslí se Spearmanovy pořadové párové korelační koeficienty neparametrického charakteru.

Spearmanovy pořadové korelační koeficienty mezi dvojicemi vysvětlujících proměnných		**Spočtená hladina významnosti**
x_1 versus x_2:	0.92843	0.0000
x_1 versus x_3:	0.15539	0.4583
x_1 versus x_4:	0.94480	0.0000
x_2 versus x_3:	0.19623	0.34717
x_2 versus x_4:	0.87781	0.0000
x_3 versus x_4:	0.21697	0.29752

Analogicky jako u Pearsonova korelačního koeficientu i *Spearmanovy párové pořadové korelační koeficienty mezi dvojicemi nezávislých proměnných* ukazují na silnou korelaci mezi prvními dvěma nezávisle proměnnými x_1 vs. x_2. Daleko slabší lineární vztah existuje mezi x_1 vs. x_3 a x_2 vs. x_3.

4. **Matice rozdílů Pearsonových a Spearmanových korelačních koeficientů:** některý software nabízí porovnání těchto dvou typů korelačních matic tak, že se vypočte matice jejich rozdílů.

Matice rozdílů Pearsonových a Spearmanových korelačních koeficientů:					
x_1 versus x_2:	0.04817				
x_1 versus x_3:	0.33538	x_2 versus x_3:	0.30395		
x_1 versus x_4:	0.01269	x_2 versus x_4:	0.04813	x_3 versus x_4:	0.24699

Někteří autoři pak doporučují takto identifikovat, která dvojice proměnných si žádá hlubšího vyšetření. Nám se však ukázalo, že v takovém případě je daleko účinnějším pomocníkem v této knize hodně využívaná regresní diagnostika, která totiž v grafech vlivných bodů snadno a jednoznačně odhalí odlehlé hodnoty.