Seminar hemometrije
2009

Multivarijantna linearna regresija

Zadatak 1 Korelacija retencionih parametara sa biološkom aktivnošću

Jedinjenje	Aktivnost	Retencioni parametri			
	A1	C18	Ph	CN-R	Si
1	−0.39	2.90	2.19	1.49	−0.41
2	−1.58	3.17	2.67	1.62	−0.52
3	−1.13	3.20	2.69	1.55	−0.33
4	−1.18	3.25	2.78	1.78	−0.55
5	−0.71	3.26	2.77	1.83	−0.45
6	−1.58	3.16	2.71	1.66	−0.51
7	−0.43	3.26	2.74	1.68	−0.39
8	−2.79	3.29	2.96	1.67	−0.71
9	−1.15	3.59	3.12	1.97	−0.56
10	−0.39	3.68	3.16	1.93	−0.50
11	−0.64	4.17	3.46	2.12	−0.55
12	−2.14	4.77	3.72	2.29	−0.80
13	−3.57	5.04	4.04	2.44	−0.86

1. Izračunajte korelacione matrice za dati set hromatografskih svojstava i jedinjenja. Da li postoje značajne korelacije među pomenutim parametrima (testirajte značajnost).
2. Izvedite višestruku linearnu regresiju koristeći A1 kao zavisnu i retencione parametre kao nezavisne promenljive. Prokomentarišite opšti kvalitet modela i vrednosti koeficijenata (njihovu statističku značajnost). Odredite greške koeficijenata i njihove intervale pouzdanosti.
3. Izvedite analizu varijanse - izračunajte SS_{corr}, SS_{factor} i SS_R, i njihove srednje vrednosti, F_{gof}. Šta na osnovu datih parametara možete reći o kvalitetu dobijenog modela?
4. Analizirajte rezidualne vrednosti, da li postoje elementi trenda, odstupanja od normalne raspodele? Da li postoje spoljašnje vrednosti ili vrednosti uticaja?
5. Kodirajte vrednosti retencionih parametara tako da se nalaze u intervalu od -1 do +1. Ponovo izračunajte multivarijantni model. Prodiskutujte vrednosti koeficijenata (predznak i apsolutna vrednost, kakve su razlike prema modelu dobijenom na osnovu sirovih podataka).
6. Izbacite 13. jedinjenje iz seta podataka. Formirajte model na osnovu sirovih vrednosti. Na osnovu dobijenog modela predvidite biološku aktivnost 13. jedinjenja.

Zadatak 2. Spektroskopsko određivanje komponenti A i B u smeši

Apsorbance na različitim talasnim dužinama

Spektar Br.	λ1	λ2	λ3	λ4	λ5	λ6	λ7	λ8
1	0,227	0,206	0,217	0,221	0,242	0,226	0,323	0,175
2	0,221	0,412	0,45	0,333	0,426	0,595	0,639	0,465
3	0,11	0,166	0,315	0,341	0,51	0,602	0,537	0,246
4	0,194	0,36	0,494	0,588	0,7	0,831	0,703	0,411
5	0,254	0,384	0,419	0,288	0,257	0,52	0,412	0,35
6	0,203	0,246	0,432	0,425	0,483	0,597	0,553	0,272
7	0,255	0,326	0,378	0,451	0,556	0,628	0,462	0,339
8	0,47	0,72	0,888	0,785	1,029	1,233	1,17	0,702
9	0,238	0,255	0,318	0,289	0,294	0,41	0,444	0,299
10	0,238	0,305	0,394	0,415	0,537	0,585	0,566	0,253

Koncentracije jedinjenja A i B

Spektar Br.	Konc. A	Konc. B
1	1	3
2	3	5
3	5	1
4	7	2
5	2	5
6	4	3
7	6	1
8	9	6
9	2	4
10	5	2

1. Izvedite linearnu regresiju za obe komponente, pri tome podrazumevate da su greške propagirane jedino duž spektralnog odgovora, odnosno da su greške koncentracije zanemarljive. Trebalo bi da se dobije po 8 koeficijenata pravca za svako jedinjenje. Šta zaključujete o osetljivosti na pojedinim talasnim dužinama?
   Opišite generalni kvalitet dobijenih regresionih modela.
2. Predvidite koncentracije komponente A koristeći se prethodnim rezultatima, analizirajte reziduale i pokažite koja talasna dužina najbolje odgovara za određivanje ove komponente.
3. Izvedite multilinearnu regresiju uključujući podatke na svim talasnim dužinama u model. Analizirajte reziduale. Da li je predviđanje koncentracije bolje na osnovu multilinearne vrednosti ili jednostruke regresije?

Zadatak 3. Linearna regresija sa jednom promenljivom. Upoznavanje sa osnovnim pojmovima

c	x vrednosti		
1	0,082	0,083	0,079
2	0,174	0,172	0,176
3	0,32	0,322	0,319
4	0,412	0,410	0,415
5	0,531	0,528	0,533
6	0,588	0,592	0,586
7	0,732	0,730	0,731
8	0,792	0,798	0,795
9	0,891	0,888	0,895
10	0,975	0,976	0,978

1. Izvediti jednostavnu regresiju uzimajući koncentraciju analita *c* kao nezavisnu promenljivu, a *x* kao zavisnu - koristite srednje vrednosti.
2. Izračunajte koeficijent pravca i odsečak koristeći se matričnim zapisom.
3. Izračunajte greške pravca i odsečka koristeći se matričnim zapisom.
4. Ponovite regresiju koristeći sve *x* vrednosti. Kakve su razlike u kvalitetu modela kada se koriste srednje, a kada pojedinačne vrednosti?
5. Izračunajte sledeće parametre analize varijanse SS_{corr}, SS_{factor}, SS_R, SS_{lof}, SS_{pe}, F_{gof}, F_{lof}. Prodiskutujte kvalitet modela.
6. Na osnovu modela dobijenog pod 1 analizirajte rezidualne vrednosti.

Zadatak 4. Particioni koeficijent zemljište-voda

Tabela 1.

Naziv	$logK_{OC}$	R_{M0}
Benzyl Alcohol	1,43	1,30
Ethyl-p-hydroxybenzoate	2,21	2,38
1-Naphthol	2,72	2,59
Anthracene	4,31	4,36
4-Chlorophenol	2,24	2,28
Phenol	1,43	1,25
2,4-Dichlorofenol	2,47	2,73
2,4,6-Trichlorophenol	2,94	3,04
Acetophenone	1,60	1,93
p-Nitrophenol	2,37	1,58
Naphtylamine	3,51	2,47
4-Methylphenol	2,70	1,87

Tabela 2.

Naziv	*A*	*B*	*S*	*E*	*V*	R_{M0}
Benzyl Alcohol	0,39	0,56	0,87	0,803	0,916	1,30
Naphtylamine	0,20	0,57	1,26	1,670	1,185	2,47
2-Naphthol	0,61	0,40	1,08	1,520	1,144	2,64
Ethyl-p-hydroxybenzoate	0,69	0,45	1,35	0,860	1,272	2,38
4-Fluoroaniline	0,28	0,40	1,09	0,760	0,834	1,40
1-Naphthol	0,60	0,37	1,05	1,520	1,144	2,59
3-Chloronitrobenzene	0,00	0,25	1,14	1,000	1,013	2,67
p-Nitrophenol	0,82	0,26	1,72	1,070	0,949	1,58
4-Methylphenol	0,57	0,31	0,87	0,820	0,916	1,87
4-Chlorophenol	0,67	0,20	1,08	0,915	0,898	2,28
Anthracene	0,00	0,28	1,34	2,290	1,454	4,36
Methyl-p-hydroxybenzoate	0,69	0,45	1,37	0,900	1,131	1,93
Phenol	0,60	0,30	0,89	0,805	0,775	1,25
Acetophenone	0,00	0,48	1,01	0,818	1,014	1,93
3-Nitrophenol	0,79	0,23	1,57	1,050	0,949	1,76
2,4-Dichlorofenol	0,53	0,19	0,84	0,960	1,020	2,73
2,4,6-Trinitrophenol	0,46	0,42	2,66	1,430	1,298	0,74
4-Hydroxybenzaldehyde	0,85	0,37	1,54	1,010	0,932	1,36
2,4,6-Trichlorophenol	0,42	0,15	0,94	1,070	1,142	3,04
2,4-Dinitrophenol	0,09	0,56	1,49	1,200	1,235	0,49
4-Methoxyphenol	0,57	0,48	1,17	0,900	0,975	1,28
2-Aminophenol	0,60	0,66	1,10	1,110	0,875	0,85
4-t-Buthylphenol	0,56	0,41	0,89	0,810	1,339	3,13
1,3,5-trihidroksibenzen	1,40	0,82	1,12	1,355	0,893	0,09
2,6-Dimethylphenol	0,39	0,39	0,79	0,860	1,057	2,11

1. Na osnovu podataka iz Tabele 1 utvrdite postojanje spoljašnjih, odnosno vrednosti uticaja. Krostite R_{M0} kao zavisnu i $logK_{OC}$ kao nezavisnu promenljivu.
2. Na osnovu podataka iz Tabele 2 uradite linearnu regresiju sa više promenljivih. Odredite vrednosti koeficijenata u modelu, njihovih grešaka i intervala pouzdanosti. Šta možete zaključiti o statističkom značaju svakog od datih koeficijenata?

3. Opišite kvalitet dobijenog modela (izgled i ANOVA reziduala, PRESS, $PRESS_{CV}$, prisustvo spoljašnjih odnosno vrednosti uticaja – odgovor obrazložite računanjem Kukove razdaljine i unutrašnjeg standardnog reziduala).
4. Ukoliko postoje statistički nebitni faktori isključite ih i na osnovu preostalih napravite novi model. Kakve su razlike u odnosu na prethodni?