Poglavlje jedanaest
Diskriminaciona analiza

11.1 ŠTA ĆETE NAUČITI IZ OVOG POGLAVLJA?

Iz ovog poglavlja ćete naučiti kako da klasifikujete jedinku u jednu od dve ili više populacija na bazi vrednosti jedne ili više promenljivih. Tačnije, naučićete:

- Kada se diskriminaciona analiza koristi (11.2, 11.3).
- O osnovnim konceptima u vezi klasifikacije jedinki (11.4).
- Značenje klasičnog metoda klasifikacije, Fišerovu diskriminacionu funkciju, i kako doći do nje (11.5).
- Kako se interpretiraju programi koji se bave diskriminacionom funkcijom (11.6).
- Kako uključiti ranije dobijene podatke u proceduru klasifikacije (11.7).
- Kako proceniti nivo uspešnosti klasifikacijske procedure (11.8).
- Kako proceniti doprinos promenljivih (11.9).
- Kako odrediti promenljive koje treba koristiti u klasifikaciji (11.10).
- O klasifikaciji u više od dve grupe (11.11).
- Kako koristiti kanoničnu korelaciju u analizi diskriminacione funkcije (11.12).
- Kako izabrati odgovarajući kompjuterski program i njegove opcije (11.13).
- Na šta treba obratiti pažnju pri diskriminacionoj analizi (11.14).

11.2 KADA SE KORISITI DISKRIMINACIONA ANALIZA?

Tehnike *Diskriminacione analize* se koriste da bi se jedinka klasifikovala u jednu od dve ili više alternativnih grupa (ili populacija) na bazi niza merenja. Kako je poznato, populacije su različite i svaka jedinka pripada jednoj od njih. Ove tehnike takođe mogu biti korišćene da bi se odredilo koje promenljive doprinose klasifikaciji. Stoga, kao i u regresionoj analizi, imamo dve uloge: predikciju (predviđanje) i deskripciju (opisivanje).

Na primer, zamislite arheologa koji želi da odredi koje od dva plemena je napravilo određenu statuu, nađenu u iskopini. Arheolog će vršiti merenja više karakteristika statue i na osnovu njih treba da odredi da li rezultati merenja ukazuju na distribuciju koja karakteriše statue jednog plemena ili drugog. Ove distribucije su bazirane na podacima dobijenim merenjem karakteristika statua za koje se zna da ih je napravilo tačno određeno od dva plemena. Problem klasifikacije je takav da pogodi ko je napravio novonađenu statuu na bazi merenja koja su dobijena od statua čije je poreklo poznato.

Merenja nove statue mogu se sastojati samo od određivanja jedne karakteristike, na primer njene visine. Međutim, u tom slučaju, očekivali bismo nizak stepen tačnosti u klasifikaciji nove statue obzirom da može biti dosta preklapanja u distribucijama visine statua koje potiču od dva navedena plemena. Ukoliko, s druge strane, klasifikaciju baziramo na više karakterisitika, imaćemo dosta pouzdanije predviđanje. Metode diskriminacione analize opisane u ovom poglavlju su tehnike sa više promenljivih, u smislu da koriste više merenja različitih karakteristika.

Kao drugi primer, zamislite bankarskog službenika na šalteru za izdavanje kredita, koji želi da odredi da li da odobri molbu za kredit automobila. Odluka treba da bude donešena tako što će se odrediti da li karakteristike (osobine) podnosioca molbe više sliče osobama koje su u prošlosti vraćale uspešno kredit ili osobama koje to nisu učinile. Informacije vezane

za te dve grupe, dobijene iz arhive banke, uključuju različite faktore, kao što su: starost, prihodi, bračno stanje i posedovanje nekretnina.

Treći primer, detaljno opisan u sledećoj sekciji, dobijen je iz podataka o depresiji pojedinaca (poglavlje 1 i 3). Želimo da, na bazi podataka o depresiji pojedinaca, predvidimo da li jedinka koja živi u zajednici ima veće ili manje šanse da bude depresivna.

11.3 PRIMER SA PODACIMA

Kako je opisano u poglavlju 1, podaci o depresiji su sakupljeni za pojdince sa mestom prebivališta u okrugu Los Anđelesa. Da bismo ilustrovali ideje opisane u ovom poglavlju, razvićemo metod za utvrđivanje da li je verovatno da osoba bude depresivna. Za potrebe ovog primera, »depresija« je definisana rezultatom 16 ili više na CESD skali (pogledati kodnu knjigu datu u tabeli 3.2). Ova informacija je data u promenljivoj zvanoj »slučajevi«. Određivanje ćemo bazirati na demografskim i drugim karakteristikama individue. Promenljive koje se koriste su obrazovanje i prihod. Takođe ćemo želeti da odredimo da li možemo poboljšati naše predviđanje uključujući informacije o bolesti, polu ili godinama. Dodatne promenljive će činiti prosečno zdravstveno stanje, broj dana provedenih u krevetu tokom protekla dva meseca (0 ukoliko je bilo manje od osam dana, 1 ukoliko je bilo osam ili više), akutne bolesti(1 ukoliko ih je bilo u protekla dva meseca, 0 ukoliko nije) i hronične bolesti (0 ako ih nije bilo i 1 ako ih je bilo).

Prvi korak u analizi podataka je određivanje deskriptivnih mera za svaku od grupa. Tabela 11.1 prikazuje srednje vrednosti i standardne devijacije za svaku promenljivu u obe grupe. Primetite da u depresivnoj grupi, grupi II, imamo veći procenat učešća žena, nižu prosečnu stopu starosti, niži nivo obrazovanja i niže prihode. Standardne devijacije su slične u obema grupama, izuzev za prihode, gde se malo razlikuju. Primetite takođe da je karakteristika zdravstvenog stanja u depresivnoj grupi, uopšte gledano, lošija nego ona u grupi u kojoj nisu depresivci, baz obzira što je grupa depresivaca, prosečno gledano, mlađa od grupe nedepresivaca. Iz razloga što je pol kodiran na način: muškarac = 1 i žena = 2, prosečni pol od 1.80 ukazuje da 80% depresivne grupe čine žene. Slično, 59% nedepresivne grupe čine muškarci.

Pretpostavimo da želimo da predvidimo da li su pojedinci depresivni, ili nisu, na bazi njihovih prihoda. Primer u tabeli 11.1 prikazuje....

Da bismo definisali šta se podrazumeva po "*visokim*" ili "*niskim*", moramo odrediti graničnu tačku. Ako obeležimo ovu tačku sa C, tada bismo klasifikovali individuu u populaciju I ukoliko važi $X \geq C$. Za bilo koju datu vrednost C mi bismo napravili određeni procenat greške. Ako bi jedinka dolazila iz populacije I ali je izmereno X bilo manje od C, mi bismo je pogrešno klasifikovali u populaciju II i obrnuto. Ova dva tipa grešaka su prikazana na slici 11.1. Ukoliko možemo pretpostaviti da dve populacije imaju jednake varijanse, tada je uobičajena vrednost C:

$$C = \frac{\overline{X}_1 + \overline{X}_{11}}{2}$$

Ova vrednost obezbeđuje da verovatnoće obeju grešaka budu jednake.

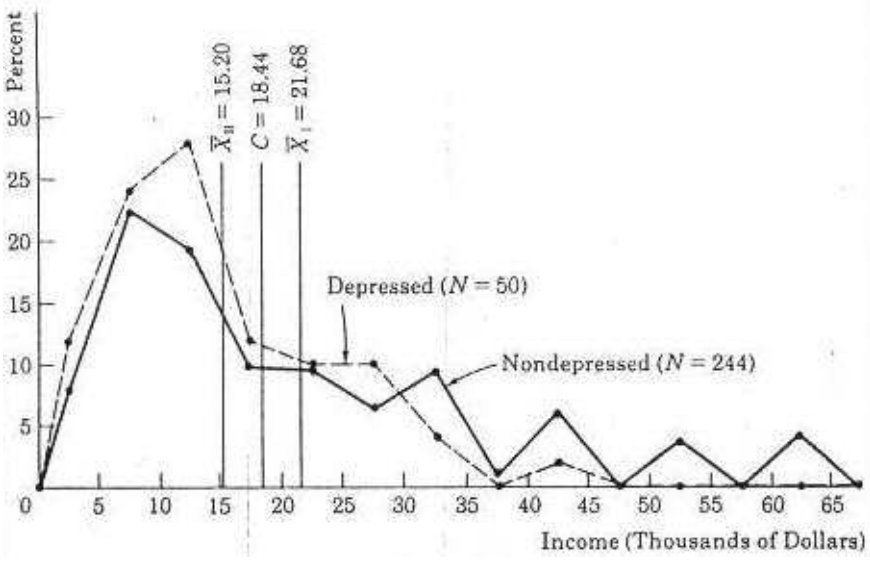
Idealizovana situacija prikazana na slici 11.1 se retko sreće u praksi. U situacijama u realnom životu nivo preklapanja dve distribucije je često velik, a varijanse su retko jednake. Na primer, u podacima o depresivnosti, distribucije prihoda za depresivne i nedepresivne pojedince se preklapaju u velikom stepenu, kako je prikazano na slici 11.2. Uobičajena granična tačka je

$$C = \frac{15.20 + 21.68}{2} = 18.44$$

Kako se može videti na slici 11.2, procenat greške je prilično velik. Tačni podaci o greškama su prikazani u tabeli 11.2.

	Klasifikovani kao		
Aktuelni status	**Nedepresivni**	**Depresivni**	**%Greške**
Nedepresivni (N=244)	123	121	50.4
Depresivni (N=50)	19	31	62.0
Ukupno (N=294)	142	152	52.4

Tabela 11.2. Klasifikacija pojedinaca kao depresivnih ili nedepresivnih bazirana samo na prihodima

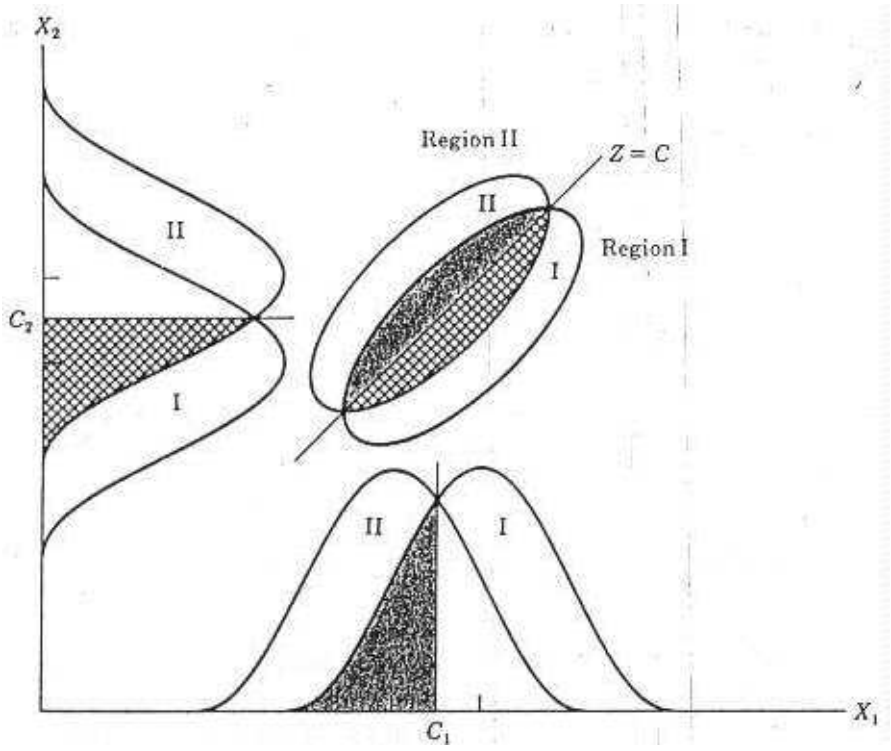


Slika 11.2. Distribucija prihoda za depresivne i nedepresivne osobe pokazuje efekte granične tačke na visini prihoda od 18,440 dolara

Ovi brojevi su dobijeni tako što je prvo utvrđivano da li je prihod svakog pojedinca veći ili jednak prihodu od 18.44 x 10^3 dolara, a zatim utvrđivanjem da li je jedinka korektno klasifikovana. Na primer, od 244 nedepresivne osobe, 123 su imale prihod veći od 18.44 x 10^3 dolara i stoga su bile korektno klasifikovane kao nedepresivne (videti tabelu 11.2). Slično, od 50 depresivnih pojedinaca, 31 je bio korektno klasifikovan. Ukupan broj tačno klasifikovanih pojedinaca iznosi 123 + 31 = 154, što čini 52,4% ukupnog uzorka od 294 osobe, kako je prikazano u tabeli 11.2. Stoga, iako su se proseci prihoda međusobno znatno razlikovali, visina prihoda, sama po sebi, nije vrlo uspešna u određivanju da li je osoba depresivna.

Kombinacija dve ili više promenljivih može doprineti boljoj klasifikaciji. Zapamtite da broj promenljivih koje se koriste mora biti manji od $N_I + N_{II} - 1$. Za dve promenljive, X_1 i X_2 koncentracione elipse mogu biti oblika prikazanog na slici 11.3 (pogledati poglavlje 7.5 za obješnjenje značenja koncentracionih elipsi). Slika 11.3 takođe prikazuje univarijantnu distribuciju X_1 i X_2 odvojeno. Univarijantna distribucija X_1 je ono što se dobije ako se vrednosti X_2 zanemare. Baziravši se samo na X_1 i njegovoj odgovarajućoj graničnoj tački C_1, došlo bi do pojave relativno velike količine grešaka (tj. pogrešne klasifikacije). Slični rezultati se javljaju i za X_2 i njegovu odgovarajuću graničnu tačku C_2. Da bismo koristili obe varijable simultano,

neophodno je da podelimo površ X_1 i X_2 na dva dela, od kojih svaki odgovara jednoj populaciji, i odgovarajuće klasifikuje osobe. Jednostavan način za određivanje ova dva regiona je da se povuče prava linija kroz tačke preseka dve koncetracione elipse, kako je prikazano na slici 11.3.



Slika 11.3. Klasifikacija u dve grupe na bazi dve promenljive

Procenat osoba iz populacije II koje su pogrešno klasifikovane je prikazan u šrafiranoj oblasti. Osenčene površine prikazuju procenat osoba iz populacije I koje su pogrešno klasifikovane. Greške koje nastaju pri korišćenju dve promenljive su često znatno manje nego one koje nastaju korišćenjem bilo koje od dve promenljive samostalno. Na ilustraciji prikazanoj na slici 11.3 radi se, upravo, o ovom slučaju.

Liniju podele je uveo R.A. Fišer (1936) kao jednačinu Z = C, gde je Z linearna kombinacija X_1 i X_2 a C konstanta definisana sledeće:

$$C = \frac{\overline{Z}_I + \overline{Z}_{II}}{2}$$

gde je:

$\overline{Z}_I$ - prosečna vrednost Z u populaciji I

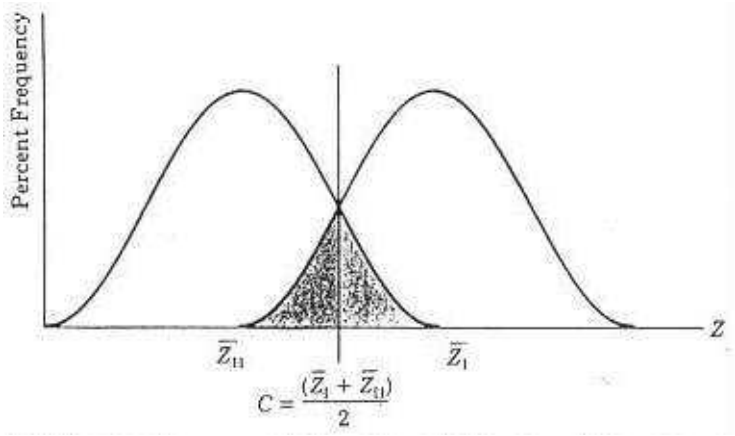
$\overline{Z}_{II}$ - prosečna vrednost Z u populaciji II

U ovoj knjizi Z ćemo zvati *Fišerovom diskriminacionom funkcijom*, napisanom kao:

$$Z = a_1 X_1 + a_2 X_2$$

za slučaj dve promenljive. Formule za izračunavanje koeficijenata a_1 i a_2 mogu biti nađene u radovima Fishera (1936), Lachenbruch-a (1975) i Afifi-ja i Azen-a (1979).

Za svakog pojedinca iz svake populacije, vrednost Z je proračunata. Kada se frekvencije raspodele Z prikažu na grafiku, dobija se rezultat kao što je prikazan na slici 11.4. U ovom slučaju, problem klasifikacije zasnovan na dve promenljive X_1 i X_2, je redukovan na situaciju u kojoj imamo samo jednu promenljivu Z.



Slika 11.4. Frekvencija raspodele Z za populacije I i II

Primer

Kao primer korišćenja ove tehnike na podacima o depresivnim osobama, možda bi bilo bolje koristiti zajedno prihod i starost da bi se klasifikovali depresivni pojedici. Program BMDP7M je korišćen da bi se dobila Fišerova diskriminaciona funkcija. Na nesreću, ova jednačina se ne može koristiti direktno sa izlaza, i neki intermedijarni proračuni moraju biti obavljeni, kako je objašnjeno u odeljku 11.6. Rezultat je:

Z = 0.0209(starost) + 0.0336(prihod)

Srednja vrednost promenljive Z za svaku od grupa može se dobiti na sledeći način, korišćenjem srednjih vrednosti iz tabele 11.1:

srednja vrednost Z = 0.0209(sr. vr. starosti) + 0.0336(sr. vr. prihoda)

Stoga:

$$\bar{Z}_{nedepresivno} = 0.0209(45.2) + 0.0336(21.68) = 1.67$$
$$\bar{Z}_{depresivno} = 0.0209(40.4) + 0.0336(15.20) = 1.36$$

Granična tačka je stoga:

$$C = \frac{1.67 + 1.36}{2} = 1.515$$

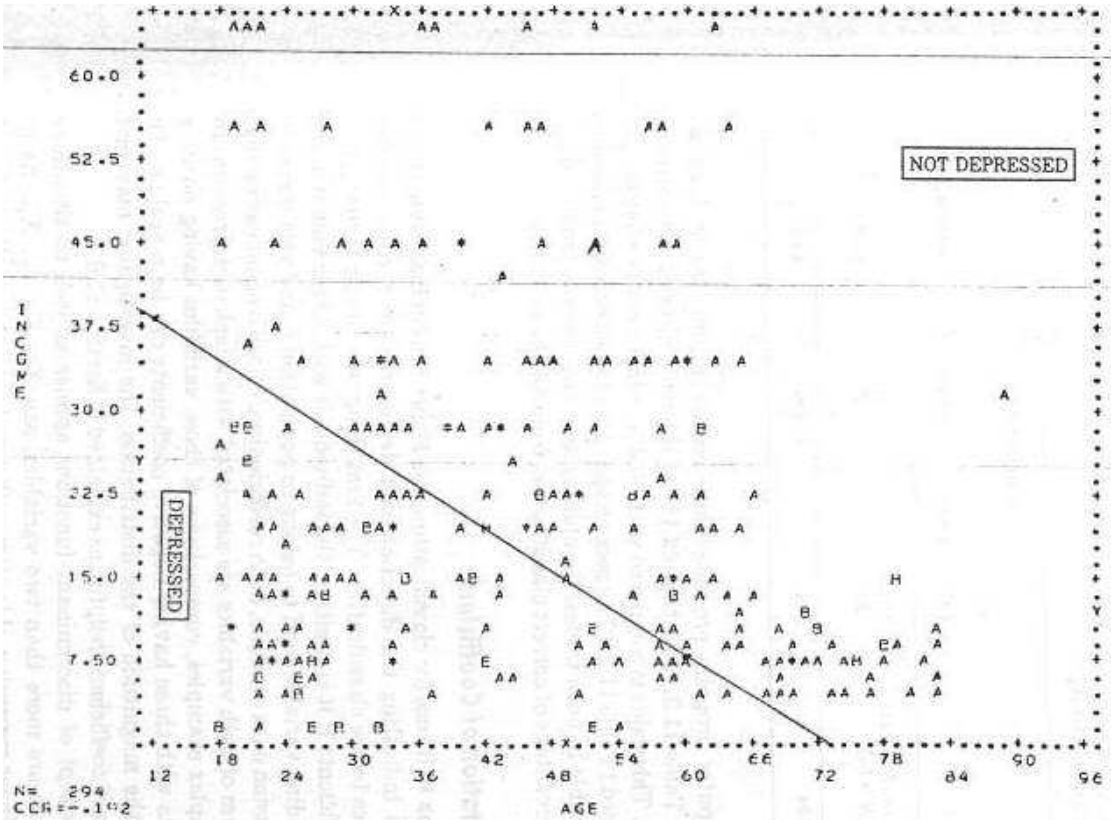
Pojedinac je klasifikovan kao depresivan ako je njegova ili njena vrednost promenljive Z manja od 1.52.

U slučaju korišćenja dve promenljive, moguće je ilustrovati postupak klasifikacije na način prikazan na slici 11.5. Ova slika je dobijena kao izlaz programa BMDP6D (pogledati

poglavlje 6). Na slici 11.5 svako A predstavlja prihod i starost nedepresivne osobe, a svako B predstavlja istu informaciju za depresivnu osobu. Granična linija je graf koji važi za jednačinu $Z = C$, tj.:

0.0209(starost) + 0.0336(prihod) = 1.515

Pojedinac koji pripada regionu iznad granične linije je klasifikovan kao ne depresivan. Primetite, zaista, da se samo nekoliko depresivnih osoba nalazi daleko iznad granične linije.



Slika 11.5 Klasifikacija pojedinaca kao depresivnih ili nedepresivnih, na bazi prihoda i starosti

Da bismo izmerili nivo uspešnosti postupka klasifikacije za ovaj primer, moramo izbrojati koliko pripadnika svake grupe je korektno klasifikovano.

	Klasifikovani kao		
Aktuelni status	**Nedepresivni**	**Depresivni**	**%Greške**
Nedepresivni (N=244)	154	90	63.1
Depresivni (N=50)	20	30	60.0
Ukupno (N=294)	174	120	62.6

Tabela 11.3. Klasifikacija pojedinaca kao depresivnih ili nedepresivnih na bazi prihoda i starosti

Kompjuterski program proračunava ove vrednosti automatski; one su prikazane u tabeli 11.3. Primetite da je 63.1% nedepresivnih osoba korektno klasifikovano. Ova vrednost se može porediti sa rezultatom od 50.4% koji je dobijen kada je korišćen prihod kao jedino merilo (tabela 11.2). Procenat depresivnih koji su korektno klasifikovani je uporediv u obe tabele. Kombinovanje starosti sa prihodom dovelo je do poboljšanja ukupnog procenta uspešno klasifikovanih sa 52.4% na 62.6%.

Interpretacija koeficijenata

Kao dodatak korišćenju pri klasifikaciji, Fišerova diskriminaciona funkcija pruža pomoć pri određivanju pravca i stepena doprinosa svake promenljive korišćene pri klasifikaciji. Prva stvar koju treba ispitati je znak svakog koeficijenta: ukoliko je pozitivan, pojedinici sa većom vrednošću odgovarajuće promenljive imaju tendenciju pripadanja populaciji I, i obrnuto. U primeru podataka o depresivnim osobama, oba koeficijenta su pozitivna. što vodi zaključku da visoke vrednosti oba koeficijenta vode nedostatku depresije. U složenijim slučajevima, poređenje promenljivih koje imaju pozitivne koeficijente sa onima koje imaju negativne, može voditi novim saznanjima. Da bi se kvantifikovala veličina distribucije, osoba koja ispituje pojavu može imati od koristi standardizovane koeficijente, kako je objašnjeno u sekciji 11.6.

Koncept diskriminacione funkcije se takođe koristi u situacijama gde ima više od dve promenljive, recimo X_1, X_2,...,X_p. Kao i u višestrukoj linearnoj regresiji, često je dovoljno odrediti mali broj promenljivih. Selekcija – određivanje promenljivih će dalje biti objašnjeno u sekciji 11.10. U sledećoj sekciji prikazaće se neophodna teorijska osnova potrebna za analizu diskriminacione funkcije.

11.5 TEORIJSKA OSNOVA

U razvoju svoje diskriminacione linearne funkcije, R. A. FIšer (1936) nije morao da pravi bilo kakve distibucijske pretpostavke za promenljive koje se koriSte tokom klasifikacije. FIšer je postavio diskriminacionu funkciju kao:

$$Z = a_1 X_1 + a_2 X_2 + \ldots + a_p X_p$$

Kao i u prethodnoj sekciji, postavljamo dve srednje vrednosti Z kao $\overline{Z}_I$ i $\overline{Z}_{II}$. Takođe imamo ukupnu varijansu uzorka Z kao S_Z^2 (ova statistika je slična ukupnoj varijansi korišćenoj u standardnom dvo-uzročnom *t* testu; videti Dixon i Massey 1983). Da bi se izmerilo koliko su »daleko« dve grupe u smislu vrednosti Z, računamo:

$$D^2 = \frac{\left(\overline{Z}_I - \overline{Z}_{II}\right)^2}{S_Z^2}$$

Fišer je odredio koeficijente a_1, a_2,..., a_p tako da D^2 dobije maksimalnu moguću vrednost.

Pojam D^2 može se interpretirati kao kvadrat rastojanja između srednjih vrednosti standardizovane vrednosti Z. Veća vrednost D^2 dovodi do zaključka da je lakše odlučiti se između dve grupe. Vrednost D^2 se zove *Mahalanobisovo rastojanje (Mahalanobis distance)*. I a_i, kao i D^2 su funkcije grupnih srednjih vrednosti i ukupne varijanse i kovarijanse promenljivih. Korisnička uputstva za pakete statističkih programa često koriste ove formule, a vi ćete ovo naći jednostavno objašnjeno u Lachenbruch (1975) ili Klecka (1980).

Neke pretpostavke u vezi distribucije čine mogućim razvoj dalje statističke procedure u vezi problema klasifikacije. Ove procedure uključuju testove hipoteza u vezi korisnosti nekih ili svih promenljivih i metode za određivanje grešaka pri klasifikaciji.

Varijable korišćene pri klasifikaciji su predstavljene kao X_1, X_2,..., X_p. Standardni model pretpostavlja da za svaku od populacija važi normalna distribucija sa više promenljivih. Dalje se pretpostavlja da je matrica kovarijanse jednaka u obe populacije. Međutim, srednje vrednosti za date promenljive mogu biti različite u dve populacije. Dalje pretpostavljamo da imamo slučajan uzorak za svaku od populacija. Veličine uzoraka su pretstavljene kao N_I i N_{II}.

Alternativno, možemo pretstaviti slučaj kao dve subpopulacije jedne jedine populacije. Na primer, u podacima o depresiji originalna populacija se sastojala isključivo od odraslih, starosti preko 18 godina, sa mestom prebivališta u okrugu Los Anđeles. Njene dve subpopulacije su depresivni i nedepresivni. Jedan uzorak je sakupljen a kasnije je ustanovljeno da se sastoji od dve subpopulacije.

11.6 INTERPRETACIJA

U ovoj sekciji predstavljamo različite metode za interpretaciju diskriminacione funkcije. Tačnije, diskutujemo u vezi analogije sa regresiom, računanja koeficijenata, standardizovanih koeficijenata i kasnijih verovatnoća.

Analogija sa regresijom

Postoji korisna veza između regresije i diskriminacione analize. Kada vršimo interpretaciju regresije, mi predstavljamo klasifikacione promenljive X_1, X_2, ... , X_p kao nezavisne promenljive. Zavisne promenljive su proste promenljive koje daju neke indikacije o populaciji za koju je vršena opservacija. Tačnije,

$$Y = \frac{N_{II}}{N_I + N_{II}}$$

ukoliko je vršena opservacija populacije I i

$$Y = -\frac{N_I}{N_I + N_{II}}$$

ukoliko je vršena opservacija populacije II. Za primer, za podatke o depresiji *Y = 50/(244 + 50)* ako osoba nije depresivna i *Y = -244/(244 + 50)* ako je osoba depresivna.

Kada se koristi uobičajena višestruka regresija, rezultujući regresioni koeficijenti su proporcionalni koeficijentima kod diskriminacione funkcije a_1, a_2, ... , a_p (videti Lachenbruch 1975). Vrednost rezultujućeg višestrukog korelacionog koeficijenta R je u vezi sa Mahalanobisovim D^2 preko sledeće formule:

$$D^2 = \frac{R^2}{1-R^2} \cdot \frac{[N_I + N_{II}] \cdot [N_I + N_{II} - 2]}{N_I \times N_{II}}$$

Dakle, moguće je iz programa višestruke regresije dobiti koeficijente diskriminacione funkcije i vrednost D^2. $\bar{Z}$ za svaku grupu se može dobiti množenjem svakog koeficijenta sa srednjom vrednošću odgovarajuće promenljive uzorka i dodavanjem rezultata. Granična tačka C se može izračunati kao:

$$C = \frac{\bar{Z}_I + \bar{Z}_{II}}{2}$$

Kao i u regresionoj analizi, neke od nezavisnih promenljivih (ili promenljivih klasifikacije) mogu biti proste promenljive (videti sekciju 9.3). U primeru podataka o depresiji mi možemo, na primer, koristiti pol kao jednu od klasifikacionih promenljivih i pri tome je tretirati kao prostu promenljivu. Istraživanja su pokazala da, iako te promenljive ne prate normalnu distribuciju, njihovo korišćenje u linearnoj diskriminacionoj analizi može poboljšati klasifikaciju.

Izračunavanje Fišerove diskriminacione funkcije

U programima za analizu diskriminacione funkcije o kojima će se govoriti u sekciji 11.13, određena izračunavanja moraju biti izvedena da bi se došlo do vrednosti diskriminacionih koeficijenata. Neki od programa (kao BMDP7M i SPSS-X DISCRIMINANT procedura) daju na izlazu nešto što je poznato kao "klasifikaciona funkcija" za svaku grupu. Drugi programi (kao SAS DISCRIM procedura) zovu ove funkcije "linearizovana diskriminaciona funkcija". Za svaku populaciju, koeficijenti su prikazani za svaku promenljivu. Koeficijenti diskriminacione funkcije a_1, a_2, ... , a_p se onda dobijaju oduzimanjem.

Kao primer, ponovo posmatramo podatke o depresiji kod osoba koristeći starost i prihod. Klasifikaciona funkcija je prikazana u tabeli 11.4.

Koeficijent a_1 za starost je *0.1634 – 0.1425 = 0.0209*. Za prihod, a_2 *= 0.1360 – 0.1024 = 0.0336*. Granična tačka C je takođe dobijena oduzimanjem, ali u *obrnutom redosledu*. Stoga *C = -4.3483 – (-5.8641) = 1.5158*. (Ova vrednost je veoma bliska prethodno izračunatoj vrednosti C = 1.515, koju smo koristili tokom ovog poglavlja.) U slučaju više od dve promenljive, koristi se identičan postupak za dobijanje a_1, a_2, ... , a_p i C.

	Klasifikaciona funkcija		
Promenljive	**Grupa I Nedepresivni**	**Grupa II Depresivni**	**Diskriminaciona funkcija**
Starost	0.1634	0.1425	0.0209 = a_1
Prihod	0.1360	0.1024	0.0336 = a_2
Konstante	-5.8641	-4.3483	1.5158 = C

Tabela 11.4 Klasifikaciona funkcija i diskriminantni koeficijenti za starost i prihod dobijeni iz programa BMDP7M

Preimenovanje grupa

Ukoliko želimo da depresivne osobe postavimo u grupu I a nedepresivne u grupu II, to možemo učiniti obrtanjem vrednosti nule i jedinice za promenljive »slučaja«. U podacima koji su korišćeni u našem primeru, ove promenljive su jednake jedinici ako je osoba depresivna i nuli ako je osoba nedepresivna. Međutim, možemo ih pretvoriti i u nulu ako je osoba depresivna i u jedinicu ako nije. Stoga procedura BMDP TRANSFORM za ovu konverziju glasi:

```
/TRANSFORM
X = CASES.
IF (X EQ 0) THEN CASES = 1.
IF (X EQ 1) THEN CASES = 0.
```

Zapamtite da obrtanje ne menja klasifikacione funkcije već jednostavno menja njihov raspored tako da *svi* znaci u linearnoj diskriminacionoj funkciji bivaju promenjeni. Nova konstanta i diskriminaciona funkcija je:

-1.515i – 0.0209(starost) – 0.0336(prihod), respektivno.

Mogućnost određivanja (diskriminacije) je identična kao i do sada, kao i broj osoba koje su tačno klasifikovane.

Standardizovani koeficijenti

Kao i u slučaju regresione analize, vrednosti a_1, a_2, ... , a_p ne mogu se direktno porediti. Međutim, uticaj relativnog efekta svake promenljive na diskriminacionu funkciju može se dobiti iz *standardizovanih diskriminacionih koeficijenata*. Ove tehnike uključuju korišćenje ukupne (ili unutar-grupne) matrice kovarijansi, sa izlaza računara. U originalnom primeru matrica kovarijansi izgleda sledeće:

	Starost	**Prihod**
Godine	324.8	-57.7
Prihod	-57.7	228.6

Stoga ukupne standardne devijacije su koren iz 324.8 što iznosi 18.02 za godine i koren iz 228.6 što iznosi 15.10 za prihod. Standardizovani koeficijenti se dobijaju množenjem a_i sa odgovarajućom ukupnom standardnom devijacijom. Stoga su standardizovani diskriminacioni koeficijenti:

(0.0209)(18.02) = 0.377 za godine

i

(0.0336)(15.10) = 0.505 za prihod

Iz izloženog se može videti da prihod ima malo širi uticaj na diskriminacionu funkciju, nego starost.

Kasnije verovatnoće

Do sada, postupak klasifikacije je pridruživao osobu bilo grupi I ili grupi II. Kako uvek postoji mogućnost pogrešne klasifikacije, mogli bismo želeti da izračunamo *verovatnoću* da osoba pripada jednoj ili drugoj grupi. Tu mogućnost možemo proračunati pomoću višepromenljivog normalnog modela, koji je objašnjen u sekciji 11.5. Formula je:

$$\textit{verovatnoća pripadanja populaciji } I = \frac{1}{1+\exp(-Z+C)}$$

gde exp(-Z + C) znači *e* na (-Z + C), kako je objašnjeno u Truett-u, Cornfield-u i Kannell-u (1967). Verovatnoća pripadnosti populaciji II je jedan minus verovanoća pripadnosti populaciji I.

Na primer, pretpostavimo da je osoba iz studije o depresivnosti, stara 42 godine i da ima prihode od 24 x 10^3 dolara godišnje. Za tu osobu vrednost diskriminacione funkcije je:

Z = 0.0209(42) + 0.0336(24) = 1.718

Kako je C = 1.515 i obzirom da je Z veće od C, klasifikovaćemo osobu kao nedepresivnu (tj. pripadaće populaciji I). Da bismo odredili koliko je verovatno da osoba bude nedepresivna, izračunaćemo verovatnoću:

$$\frac{1}{1+\exp(-1.718+1.515)} = 0.55$$

Verovatnoća da osoba bude depresivna iznosi 1 – 0.55 = 0.45. Stoga, za ovu osobu je malo više verovatno da bude nedepresivna nego depresivna.

Više paketa kompjuterskih programa računa verovatnoću pripadnosti obema grupama za svaku osobu u uzorku. U nekim programima ove verovatnoće se nazivaju *kasnijim verovatnoćama* obzirom da one izražavaju verovatnoću pripadnosti određenoj populaciji kasnije u odnosu na analizu.

Kasnije verovatnoće nude dragocen metod interpretacije rezultata klasifikacije. Istraživač može želeti da klasifikuje samo one osobe čije verovatnoće idu jasno u pravcu jedne ili druge grupe. Ocene za osobe čije su verovatnoće blizu 0.5 mogu biti suspendovane. U sledećoj sekciji drugi tip verovatnoće, nazvan ranija verovatnoća, biće definisan i korišćen da bi se modifikovala granična tačka.

Na kraju, naglašavamo da diskriminaciona funkcija, kako je prikazana ovde, predstavlja primer jednostavne procene diskriminacione funkcije populacije. Izračunali bismo kasnije da smo imali konkretne vrednosti parametara populacije. Da su populacije bile višepromenljive normalne sa jednakim matricama kovarijansi, tada bi postupak diskriminacione klasifikacije populacija bio optimalan; tj. nijedan drugi postupak klasifikacije ne bi proizveo manju ukupnu grešku klasifikacije (videti Anderson 1984).

11.7. PODEŠAVANJE VREDNOSTI GRANIČNE TAČKE

U ovom poglavlju navodimo kako izračunavanje ranije verovatnoće i gubici zbog pogrešne klasifikacije mogu uticati na izbor granične tačke C.

Uključenje ranijih verovatnoća u izbor tačke C

Do sada, granična tačka C je bila korišćena kao tačka koja je uticala na podjednak procenat grešaka oba tipa, tj. verovatnoće pogrešne klasifikacije osobe iz populacije I u populaciju II i obrnuto. Ova uloga tačke C može biti viđena na slici 11.4. Ali izbor vrednosti tačke C može biti takav da rezultat bude bilo koji odnos navedenih verovatnoća grešaka. Da bismo objasnili način na koji se takav izbor vrši, moramo uvesti koncept *ranije verovatnoće*. Obzirom da dve populacije zajedno čine ukupnu populaciju, nas zanima da ispitamo njihovu relativnu veličinu. Ranija verovatnoća populacije I je verovatnoća da osoba koja je slučajno izabrana zaista dolazi iz populacije I. Drugim rečima, to je onaj udeo osoba ukupne populacije, koji pripada populaciji I. Ovaj udeo se obeležava sa q_I.

U podacima o depresivnim osobama, definicija depresivne osobe se zasnivala na tome da je 20% osoba definisano kao depresivno i 80% kao nedepresivno. Stoga, ranija verovatnoća da osoba nije depresivna (populacija I) je $q_I = 0.8$. Slično, $q_{II} = 1 – q_I = 0.2$. Bez poznavanja bilo koje karakteristike osobe koja je izabrana, mi bismo mogli tvrditi da je on ili ona nedepresivna osoba, obzirom da toj grupi pripada 80% osoba. U tom slučaju bili bismo u pravu u 80% slučajeva. Navedeni primer nudi intuitivnu interpretaciju ranijih verovatnoća. Primetite, međutim, da bismo uvek pogrešili pri identifikaciji depresivne osobe.

Teorija koja stoji iza izbora tačke C je tako postavljena da ukupna verovatnoća pogrešnih klasifikacija bude minimalna. Ova ukupna verovatnoća je definisana kao q_I (verovatnoća pogrešne klasifikacije osobe iz populacije I u populaciju II) plus q_{II} (verovatnoća pogrešne klasifikacije osobe iz populacije II u populaciju I), ili:

$$q_I \cdot Prob(II\ dato\ I) + q_{II} \cdot Prob(I\ dato\ II)$$

U slučaju normalnog modela sa više promenljivih, koji je spomenut u sekciji 11.5, optimalni izbor granične tačke C je:

$$C = \frac{\overline{Z}_I + \overline{Z}_{II}}{2} + \ln \frac{q_{II}}{q_I}$$

gde je ln – prirodni logaritam. Primetite da ukoliko je $q_I = q_{II} = ½$, tada je $q_{II}/q_I = 1$ i $\ln(q_{II}/q_I) = 0$. U tom slučaju C je:

$$C = \frac{\overline{Z}_I + \overline{Z}_{II}}{2}$$

ako je $q_I = q_{II}$

Zbog toga smo u prethodnim sekcijama implicitno podrazumevali da je $q_I = q_{II} = ½$.

Za podatke o depresivnim, osobama videli smo da je $q_I = 0.8$, i stoga teorijska granična tačka treba da bude:

C = 1.515 + ln(0.25) = 1.515 – 1.386 = 0.129

Tokom ispitivanja podataka, vidimo da korišćenje granične tačke klasifikuje sve nedepresivne osobe tačno ali, takođe, sve depresivne osobe netačno. Stoga je verovatnoća klasifikacije nedepresivnih osoba (populacija I) kao depresivnih (populacija II) jednaka nuli. S druge strane, verovatnoća klasifikacije depresivne osobe (populacija II) kao nedepresivne (populacija I) je 1. Iz tog razloga je ukupna verovatnoća pogrešne klasifikacije (0.8)(0) + (0.2)(1) = 0.2. Kada smo koristili C = 1.515, dve verovatnoće pogrešne klasifikacije su bile 0.369 i 0.400, respektivno (videti tabelu 11.3). U tom slučaju, ukupna verovatnoća pogrešne klasifikacije je (0.8)(0.379) + (0.2)(0.400) = 0.383. Rezultat dokazuje da je teorijski izabrana granična tačka proizvela manju vrednost za ukupnu verovatnoću pogrešne klasifikacije.

U praksi, međutim, nije od pomoći ukoliko se nijedna od depresivnih osoba ne identifikuje. Ukoliko je cilj klasifikacije bila preventivna zaštita, verovatno bismo želeli da pogrešno označimo neke od nedepresivnih osoba kao depresivne da ne bismo propustili isuviše onih koji su zaista depresivni. U praksi, birali bismo razne vrednosti C i za svaku od vrednosti određivali dve verovatnoće pogrešne klasifikacije. Željena vrednost C bi bila postignuta kada bi se postigao balans te dve verovatnoće.

Uključenje gubitaka u izbor tačke C

Jedna od metoda određivanja težine grešaka ide preko određivanja relativnih gubitaka dva tipa pogrešnih klasifikacija. Na primer, pretpostavimo da je četiri puta ozbiljnije pogrešno označiti depresivnu osobu kao nedepresivnu, nego što je označiti nedepresivnu kao depresivnu. Ovi gubici se mogu izraziti kao:

gubitak(II dato I) = 1

i

gubitak(I dato II) = 4

Granična tačka C tada može biti izabrana tako da minimizira ukupne gubitke pogrešne klasifikacije, tj.:

$q_I \cdot Prob(II\ dato\ I) \cdot gubitak(II\ dato\ I) + q_{II} \cdot Prob(I\ dato\ II) \cdot gubitak(I\ dato\ II)$

Izbor tačke C koja postiže datu minimizaciju je:

$$C = \frac{\overline{Z}_I + \overline{Z}_{II}}{2} + K$$

gde je:

$$K = \ln \frac{q_{II} \cdot gubitak(IdatoII)}{q_I \cdot gubitak(IIdatoI)}$$

U primeru sa podacima o depresivnim osobama, vrednost K iznosi:

$$K = \ln \frac{0.2(4)}{0.8(1)} = \ln 1 = 0$$

Drugim rečima, ovaj izbor numeričke vrednosti gubitaka usled pogrešne klasifikacije i korišćenje ranije verovatnoće su suparnički nastrojeni jedno prema drugom tako da je C = 1.515, što je identična vrednost koja se dobija i bez uključenja gubitaka i ranijih verovatnoća.

Na kraju, bitno je primetiti da uključenje ranijih verovatnoća i gubitaka usled pogrešne klasifikacije menja samo izbor granične tačke C. Ono ne utiče na proračun koeficijenata a_1, a_2, ... , a_p u diskriminacionoj funkciji. Ukoliko računarski program ne dozvoljava opciju uključenja ovih kvantifikatora, granična tačka se lako može modifikovati kao što je učinjeno u gornjem primeru. U sekciji 11.13, pokazaćemo kako gubici usled pogrešne klasifikacije mogu biti uključeni u program koji dozvoljava samo unošenje ranijih verovatnoća.

11.8 KOLIKO JE DOBRA DISKRIMINACIONA FUNKCIJA?

Mera dobrote za postupak klasifikacije sastoji se od dve verovatnoće za pogrešnu klasifikaciju, verovatnoće (II dato I) i verovatnoće (I dato II). Postoje različiti metodi za utvrđivanje ove dve funkcije. Jedan metod, poznat kao *empirijski metod*, korišćen je u prethodnim primerima, tj. korisitili smo diskriminacionu funkciju na istim primerima korišćenim za izračunavanja te funkcije i odredili udeo pogrešno klasifikovanih za svaku grupu (vidi tabele 11.2 i 11.3). Ovaj proces predstavlja način provere diskriminacione funkcije. Iako je ovaj metod intuitivno jasan i jednostavan, on ipak proizvodi pristrasne procene. U suštini, rezulrujući udeli koji se dobijaju na kraju, podcenjuju pravu verovatnoću pogrešne klasifikacije, jer je isti uzorak korišćen za računanje i proveru diskriminacione funkcije.

U idealnom slučaju, želeli bismo da izvedemo funkciju iz jednog uzorka, a zatim je primenimo na drugom da bismo utvrdili udeo pogrešno klasifikovanih. Ovaj postupak je poznat pod nazivom *unakrsna provera*, i proizvodi nepristrasne procene. Istraživač može ostvariti unakrsnu proveru tako što, na slučajan način, podeli originalni uzorak na dva poduzorka: jedan za računanje diskiminancione funkcije i drugi za njenu unakrsnu proveru.

Istraživač, doduše, može i oklevati da podeli uzorak ako je on isuviše mali. Alternativni metod koji se katkada koristi u ovom slučaju, a imitira podelu uzorka na poduzorke, poznat je pod nazivom *jackknife procedura*. U ovom metodu, isključujemo jednu od observacija iz prve grupe i računamo diskriminacionu funkciju na osnovu preostalih opservacija. Zatim vršimo klasifikaciju izostavljenih opservacija. Ova procedura se ponavlja za svaku opservaciju u prvom uzorku. Udeo pogrešno klasifikovanih osoba je *jackknife* procena *Prob(II dato I)*. Slična procedura se koristi za utvrđivanje Prob(I dato II). Ovaj metod daje približno nepristrasne procene. Takođe, neki od računarskih programa nude i ovu opciju.

Ukoliko prihvatimo normalni model sa više promenljivih, teorijski način procene verovatnoća postaje takođe moguć. U tom slučaju zahteva se da je poznato samo D^2. Formule su:

procenjeno Prob(II dato I) = površ levo od $\left(\frac{K-1/2D^2}{D}\right)$ *pod uslovom standardne normalne krive*

i

procenjeno Prob(I dato II) = površ levo od $\left(\frac{-K-1/2D^2}{D}\right)$ *pod uslovom standardne normalne krive*

gde je:

$$K = \ln\frac{q_{II}\cdot gubici(IdatoII)}{q_I\cdot gubici(IIdatoI)}$$

Ukoliko je K = 0, ove dve promene su međusobno jednake površini koja se nalazi levo od (-D/2) pod uslovom standardne normalne krive. Na primer, u primeru sa podacima o depresivnim osobama, D^2 = 0.319 i K = 0. Stoga D/2 = 0.282, i površina levo od –0.282 je 0.389. Dalje ustanovljavamo i Prob(II dato I) i Prob(I dato II) kao 0.39. Ovaj postupak je naročito koristan ukoliko diskriminacionu funkciju dobijamo iz regresionog programa, obzirom da D^2 lako može biti proračunato iz R^2 (videti 11.6).

Na nesreću, ovaj metod takođe podcenjuje prave verovatnoće pogrešne klasifikacije. *Nepristrasna procena* populacije Mahalanobis D^2 je:

$$nepristrasnoD^2 = \frac{N_I+N_{II}-P-3}{N_I+N_{II}-2}D^2 - P\left(\frac{1}{N_I}+\frac{1}{N_{II}}\right)$$

U primeru depresije imamo:

$$nepristrasnoD^2 = \frac{50+244-2-3}{50+244-2}(0.319) - 2\left(\frac{1}{50}+\frac{1}{244}\right) = 0.316-0.048 = 0.268$$

Rezultujuća površina je proračunata na sličan način kao i u prethodnom metodu. Obzirom da je nepristrasno D/2 = 0.259, rezultujuća površina mora biti levo od –D/2 i iznosi 0.398. Kada ovaj rezultat poredimo sa rezultatom baziranim na nepristrasnom D^2, primećujemo da su razlike male jer (1) samo dve promenljive su korišćene i (2) veličine uzoraka su poprilično velike. Sa druge strane, ukoliko bi broj promenljivih bio blizu ukupnoj veličini uzorka ($N_I + N_{II}$), dve procene bi se mogle veoma razlikovati.

Preporučljivo je da, kad god je to moguće, istraživač odredi bar neke od gorenavedenih grešaka usled pogrešne klasifikacije i izvrši korekciju odgovarajućih predviđanja.

Da bi procenio kako se određena diskriminaciona funkcija ponaša, istraživač može naći za korisno da izračuna verovatnoću tačnog predviđanja na osnovu čistog *pogađanja*. Taj postupak je sledeći: Pretpostavimo da je ranija verovatnoća pripadanja populaciji I poznata i iznosi q_I. Tada je $q_{II} = 1 – q_I$. Jedan od načina da klasifikujemo osobe korišćenjem samo ovih verovatnoća je da zamislimo novčić koji, kada padne glava, daje verovatnoću q_I, a kada padne pismo daje q_{II}. Svaki put kada treba klasifikovati neku osobu, baca se novčić. Osoba se klasifikuje u populaciju I ako novčić padne na glavu, a ako padne pismo, klasifikuje se u populaciju II. Ukupno gledano, udeo q_I će predstavljati osobe koje će biti klasifikovane u populaciju I.

Zatim, istraživač računa ukupnu verovatnoću tačne klasifikacije. Podsetimo da je verovatnoća da osoba pripada populaciji I q_I, a da je verovatnoća da osoba bude klasifikovana u populaciju I takođe q_I. Stoga, verovatnoća da osoba koja pripada populaciji I bude *tačno* klasifikovana u populaciju I iznosi q_I^2. Slično, q_{II}^2 je verovatnoća da osoba koja pripada populaciji II bude tačno klasifikovana u tu populaciju. Stoga je ukupna verovatnoća tačne klasifikacije korišćenjem samo ranijih verovatnoća $q_I^2 + q_{II}^2$. Primetite da se najniža moguća vrednost ove verovatnoće javlja kada je $q_I = 0.5$, tj. kada je podjednako verovatno da osoba pripada obema populacijama. U tom slučaju $q_I^2 + q_{II}^2 = 0.5$.

Korišćenjem ovog metoda na podacima o depresivnim osobama, sa $q_I = 0.8$ i $q_{II} = 0.2$, dobijamo $q_I^2 + q_{II}^2 = 0.68$. Stoga bismo očekivali da više od dve trećine osoba bude tačno klasifikovano ukoliko samo bacamo novčić koji u 80% slučajeva pada na glavu. Primetite, ipak, da bismo pogrešili u 80% slučajeva u vezi depresivnih osoba, a to je situacija preko koje se možda ne može preći. (U vezi ovoga možete se podsetiti uloge gubitaka usled pogrešne klasifikacije, o čemu smo pričali u sekciji 11.7)

11.9 TESTIRANJE DOPRINOSA KLASIFIKACIONIH PROMENLJIVIH

Da li možemo klasifikovati osobe korišćenjem raspoloživih promenljivih bolje, nego što to možemo klasifikujući ih slučajno? Jedan od odgovora na ovo pitanje podrazumeva normalni višepromenljivi model prikazan u sekciji 11.4. Ovo pitanje može biti formulisano kao problem za testiranje hipoteze. Nulta hipoteza koja se testira može biti da nijedna od promenljivih neće poboljšati rezultate tzv. slučajne klasifikacije. Ekvivalnetna nulta hipoteza je da su srednje vrednosti obe populacije za svaku promenljivu jednake, odnosno da je populacioni D^2 nula. Statistički test za nultu hipotezu je:

$$F = \frac{N_I + N_{II} - P - 1}{P(N_I + N_{II} - 2)} \times \frac{N_I N_{II}}{N_I + N_{II}} \times D^2$$

sa stepenom slobode P i $N_I + N_{II} - P - 1$ (videti Rao 1973). Vrednost P je »repna« površina udesno proračunatog statističkog testa. Naglašavamo da je $N_I N_{II} D^2/(N_I + N_{II})$ poznata kao dvo-uzročna *Hotelling* T^2, originalno razvijena za testiranje jednakosti dva seta srednjih vrednosti (videti Morrison 1976).

Za primer podataka o depresivnim osobama, proračunata vrednost F je:

$$F = \frac{244 + 50 - 2 - 1}{2(244 + 50 - 2)} \times \frac{244 \times 50}{244 + 50} \times 0.319 = 6.60$$

sa 2 i 291 stepena slobode. Vrednost P za ovaj test je manja od 0.005. Stoga ove dve promenljive osetno poboljšavaju predviđanje koje je bazirano samo na slučaju. Ekvivalentno, postoji statistički dokaz da usrednjene vrednosti populacije nisu jednake u obema grupama. Treba zapamtiti da većina računarskih programa ne izbacuje vrednost D^2. Međutim, iz izračunate vrednosti gorenavedenog F, može se izračunati D^2 na sledeći način:

$$D^2 = \frac{P(N_I + N_{II})(N_I + N_{II} - 2)}{(N_I N_{II})(N_I + N_{II} - P - 1)} F$$

Sledeći koristan test je da li dodatna promenljiva poboljšava diskriminaciju. Pretpostavimo da je populaciona D^2 bazirana na promenljivima $X_1, X_2, \ldots, X_p$ predstavljena preko pop D^2_P. Ako želimo da testiramo da li dodatna promenljiva X_{P+1} može značajno da

poveća pop D^2, tj. testiramo hipotezu da je $D^2_{P+1} = D^2_P$. Statistički test, uzimajući u obzir višepromenljivi normalni model, daje statističko F na sledeći način:

$$F = \frac{(N_I + N_{II} - P - 2)(N_I N_{II})(D^2_{P+1} - D^2_P)}{(N_I + N_{II})(N_I + N_{II} - 2) + N_I N_{II} D^2_P}$$

sa jednim i $(N_I + N_{II} - P - 2)$ stepena slobode (videti Rao 1965).

Na primer, zamislimo da u podacima o depresivnim osobama želimo da testiramo hipotezu da starost poboljšava diskiminacionu funkciju kada je u kombinaciji sa prihodom. D_1^2 za prihod samo iznosi 0.183, a D_2^2 za prihod i starost iznosi 0.319. Stoga, za P = 1

$$F = \frac{(50 + 244 - 1 - 2)(50 \times 244)(0.319 - 0.183)}{(294)(292) + 50 \times 244 \times 0.183} = 5.48$$

sa jednim i 291 stepena slobode. Vrednost P za ovaj test je jednaka 0.02. Stoga možemo zaključiti da starost značajno poboljšava klasifikaciju kada je u kombinaciji sa prihodima.

Generalizacija poslednjeg testa nam dozvoljava ispitivanje doprinosa više dodatnih promenljivih, i to simultano. Tačnije, ukoliko započnemo sa X_1, X_2, ... , X_p promenljivih, možemo testirati da li X_{P+1}, ... , X_{P+Q} promenljivih može poboljšati predviđanje. U ovom slučaju testiramo hipotezu da je pop D^2_{P+Q} = pop D^2_P. Za višepromenljivi normalni model, statistički test je:

$$F = \frac{(N_I + N_{II} - P - Q - 1)}{Q} \times \frac{N_I N_{II}(D^2_{P+Q} - D^2_P)}{(N_I + N_{II})(N_I + N_{II} - 2) + N_I N_{II} D^2_P}$$

sa Q i $(N_I + N_{II} - P - Q - 1)$ stepena slobode.

Poslednje dve formule su korisne pri određivanju promenljivih, što će biti prikazano u sledećoj sekciji.

11.10 ODREĐIVANJE PROMENLJIVIH

Setite se da postoji analigija između regersione analize i analize diskriminacione funkcije. Stoga, veliki deo izloženog u poglavlju 8 u vezi selekcije promenljivih, može se primeniti za selekciju promenljivih pri klasifikaciji u dve grupe. U suštini, računarski programi o kojima smo pričali u poglavlju 8 mogu biti isto tako i ovde korišćeni. Ovi programi su često tipa stepwise regresioni programi i subset regresioni programi. Kao dodatnu mogućnost, neki programi nude i izvođenje stepwise diskriminacione analize, na osnovu istih koncepata koji se koriste kod stepwise regresione analize.

Pri analizi diskriminacione funkcije, umesto da se testira da li se vrednost višestrukog R^2 menja kada se dodaje (ili oduzima) promenljiva, mi testiramo da li se vrednost pop D^2 menja pri dodavanju ili oduzimanju promenljive. Statističko F dato u sekciji 11.19 se koristi za ovu svrhu. Kao i ranije, korisnik može odrediti vrednosti F-za-unos i F-za-izbacivanje. Za F-za-unos Constanza i Afifi (1979) preporučuju korišćenje vrednosti koja odgovara za P = 0.15. Za sada ne postoji preporučena vrednost za F-za-izbacivanje, ali za neku razumnu vrednost može se uzeti P = 0.30.

11.11. KLASIFIKACIJA U VIŠE OD DVE GRUPE

Detaljna diskusija u vezi klasifikacije u više od dve grupe je van područja koje ova knjiga obrađuje. Međutim, više knjiga uključuju u sebi detaljnu diskusiju na ovu temu, a neke su Tatsuoka (1988), Lachenbruch (1975), Morrison (1976) i Afifi i Azen (1979). U ovoj sekciji ćemo sumirati klasifikacioni postupak za višepromenljivi normalni model i diskutovati primer.

U ovom slučaju, pretpostavljamo da osoba treba da bude klasifikovana u jednu od *k* populacija, $k \geq 2$, na bazi vrednosti P promenljivih X_1, X_2, ... , X_p. Pretpostavljamo da u svakoj od k populacija P promenljive imaju višepromenljivu normalnu distribuciju, sa istom matricom kovarijanse. Tipični kompjuterski program bi proračunao, iz uzorka svake populacije, klasifikacionu funkciju. U aplikacijama, vrednosti promenljive za datu osobu se zamenjuju u svakoj klasifikacionoj funkciji, i osoba se klasifikuje u populaciju na osnovu najviše klasifikacione funkcije.

Sada ćemo razmotriti primer. U podacima o depresivnim osobama, vrednosti CESD mogu biti između 0 i 60 (pogledati poglavlje 3). U ranijim proračunima, osobe bi bile određene kao depresivne ako bi njihov CESD rezultat bio veći ili jednak 16; u suprotnom, bile bi klasifikovane kao nedepresivne. Druga mogućnost je da podelimo osobe na k = 3 grupe: one koje potpuno pobijaju mogućnost da su depresivne (CESD rezultat = 0), one koje imaju CESD rezultat između 1 i 15 inkluzivno, i one koje imaju CESD rezultat 16 ili viši. Ove grupe možemo nazvati *lowdep*(1), *meddep*(2) i *highdep*(3).

Promenljive koje se razmatraju pri unosu na stepwise način su pol, starost, obrazovanje, prihod, zdravlje, broj dana provedenih u krevetu i hronične bolesti. Ove promenljive se koriste u sekciji 11.13, gde su unete u program BMDP7M. Metod unosa promenljivih koji se koristi u datom programu je sličan metodu opisanom za dalju stepwise regresiju u poglavlju 8.

Delimični rezultati za ovaj primer dati su u tabeli 11.5. Primetite da nisu sve promenljive unete u diskiminacionu funkciju.

Aproksimirano statističko F dato u tabeli 11.5 testira nultu hipotezu da su srednje vrednosti jednake za sve promenljive simultano. Iz tabele A.4, date u dodatku, možemo videti da je vrednost P veoma mala, što ukazuje na to da nultu hipotezu treba odbaciti.

Matrica F, koja je deo tabele 11.5, daje statističko F za testiranje jednakosti srednjih vrednosti za svaki par u grupi. Vrednost F (2.36) za *lowdep* i *meddep* grupe je značajno jedva na nivou od 5%. Ostala statistička F ukazuju na značajnu razliku između *highdep* grupe i ostale dve. Ovaj rezultat delimično potvrđuje našu raniju analizu dve grupe, gde je *highdep* grupa depresivna grupa, a *lowdep* i *meddep* čine nedepresivnu grupu.

Klasifikacione funkcije su takođe prikazane u tabeli 11.5. Da bismo klasifikovali novu osobu u jednu od ove tri grupe, prvo određujemo svaku od tri klasifikacione funkcije, koristeći rezultat osobe vezan za promenljive koje ulaze u funkciju. Zatim se osoba dodeljuje grupi za koju proračunata klasifikaciona funkcija ima najvišu vrednost. Ako smo zainteresovani za dve grupe istovremeno, odgovarajući par klasifikacionih funkcija treba oduzeti jednu od druge da bi se dobila diskriminaciona funkcija, kako je objašnjeno u sekciji 11.6.

U ovom primeru, mi smo delili određenu promenljivu na tri podseta, kako bismo dobili tri grupe. Često, kada radimo sa nominalnim podacima, javljaju se tri ili više odvojenih grupa. Zbog toga, mogućnost analize diskriminacione funkcije kada se radi sa više od dve grupe, predstavlja korisnu osobinu u nekim aplikacijama.

APPROXIMATE F-STATISTIC 4.347 DEGREES OF FREEDOM 12.00 572.0

F-MATRIX DEGREES OF FREEDOM = 6 286

	lowdep	meddep
meddep	2.36	
highdep	7.12	5.58

CLASSIFICATION FUNCTIONS

GROUP =	lowdep	meddep	highdep
VARIABLE			
2 sex	7.07529	7.49617	8.14165
3 age	.16774	.13935	.11698
5 educat	2.54993	2.82551	2.68116
7 income	.10533	.09005	.06537
32 health	2.13954	2.75024	3.10425
35 beddays	-.97394	-.80246	.46685
CONSTANT	-17.62107	-18.54811	-18.81630

VARIABLE	COEFFICIENTS FOR CANONICAL VARIABLES	
2 sex	-.73103	-.01977
3 age	.03167	.02531
5 educat	-.00617	-.65481
7 income	.02758	-.00067
32 health	-.57524	-.68822
35 beddays	-1.13644	1.13117
CONSTANT	.49631	2.17769

Tabela 11.5. Delimični štampani prikaz iz programa BMDP7M za slučaj klasifikacije u više od dve grupe korišćenjem podataka o depresiji za k = 3 grupe

11.12 KORIŠĆENJE KANONIČNE KORELACIJE U ANALIZI DISKRIMINACIONE FUNKCIJE

Prisetite se da postoji analogija između regresije i diskriminacione analize za dve grupe. Nezgodno je to što ova anlogija *ne važi* za sličaj kada je k veće od dva. Međutim, u tom slučaju, postoji korespodencija između kanonične korelacije i klasifikacije u više populacija.

Počinjemo tako što definišemo novi set promenljivih Y_1, Y_2, ... , Yk_{-1}. Ovo su proste ili indikatorske promenljive koje pokazuju iz koje grupe svaki član uzorka dolazi. Zapamtite da, kao što je već diskutovano u odeljku 9, nama treba k-1 prostih promenljivih da bismo opisali k grupa. Na primer, pretpostavimo da ima k = 4 grupe. Tada proste promenljive Y_1, Y_2 i Y_3 formiramo na sledeći način:

Grupe	**Y_1**	**Y_2**	**Y_3**
1	1	0	0
2	0	1	0
3	0	0	1
4	0	0	0

Stoga će osobi koja dolazi iz grupe 1 biti dodeljena vrednost 1 za Y_1, 0 za Y_2 i 0 za Y_3 itd.

Ako imamo Q = k – 1, tada imamo uzorak sa dva seta promenljivih, Y_1, Y_2, ... , Y_Q i X_1, X_2, ... , X_P. Sada ćemo izvesti analizu kanonične korelacije nad ovim promenljivima. Ova analiza će rezultirati setom od U_i promenljivih i setom od V_i promenljivih. Kako je objašnjeno u poglavlju 10, broj parova ovih promenljivih je manji od P i Q. Zato je ovaj broj manji od brojeva P i k – 1.

Promenljiva V_1 je linearna kombinacija promenljivih X sa makimalnom korelacijom sa U_1. U tom smislu, V_1 maksimizuje korelaciju sa prostim promenljivima koje predstavljaju grupe, i stoga prikazuje maksimalne razlike koje postoje između grupa. Slično, V_2 prikazuje maksimalne razlike pod uslovom da V_2 nije u korelaciji sa V_1; itd.

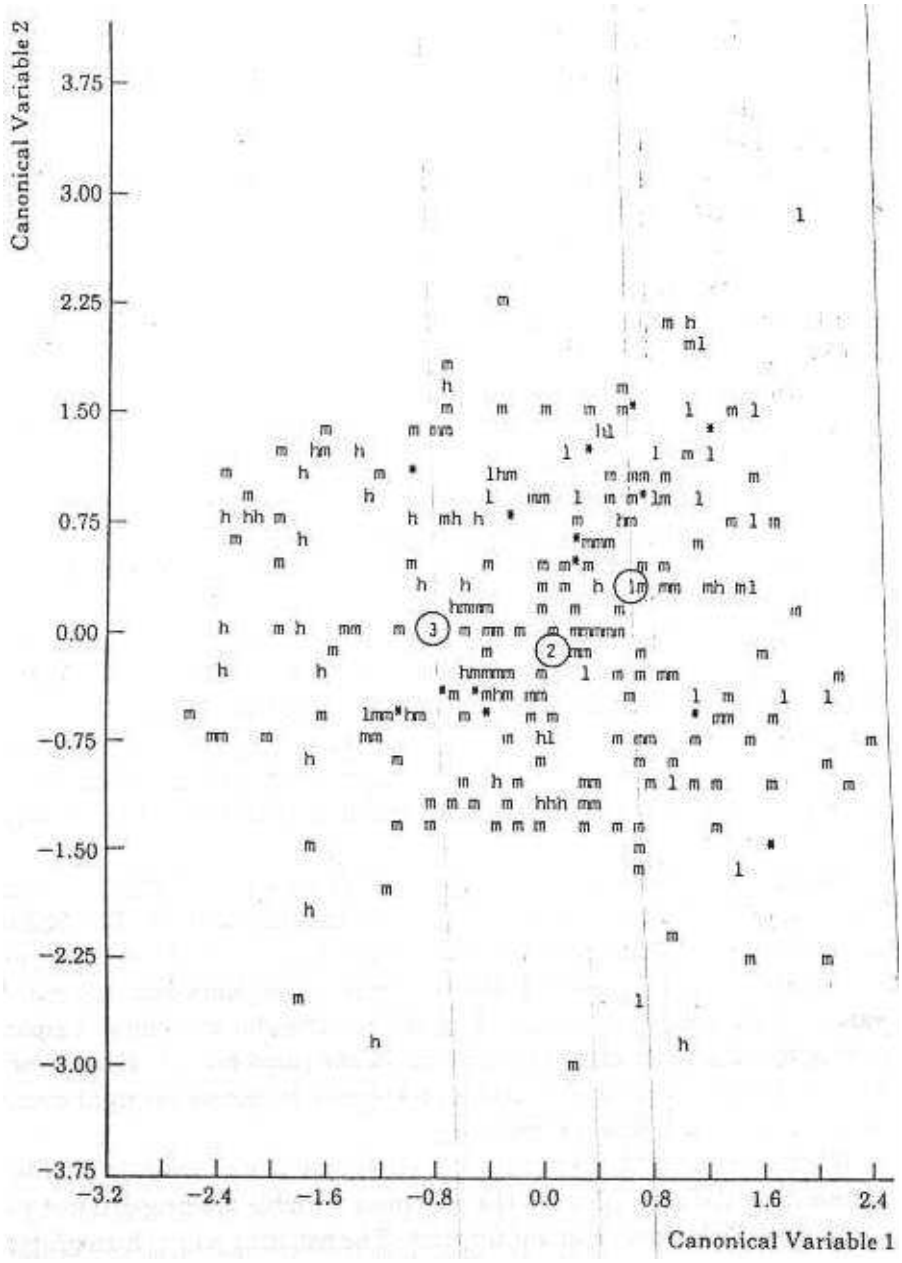
Jednom izvedene, promenljive V_i treba ispitivati kako je objašnjeno u poglavlju 10, u pokušaju da im se da smisleno značenje. Ova interpretacija može biti bazirana na veličini standardizovanih koeficijenata koji su dobijeni na izlazu. Promenljive koje imaju najveće standardizovane koeficijente mogu dati »imena« kanoničnim diskriminacionim funkcijama. Da bi dalje pomogli korisniku u ovoj interperetaciji, SPSS-X DISCRIMINANT procedura nudi opciju rotiranja promenljive V_i pomoću varimax metode (objašnjeno u poglavlju 15, faktorska analiza).

Promenljive V_i su nazvane *kanonične diskriminacione funkcije* ili *kanonične promenljive* na izlazu programa. Za primer podataka o depresivnim osobama, rezultati su dati na dnu tabele 11.5. Obzirom da postoje tri grupe, dobili smo za k – 1 = 2 kanonične (diskriminacione funkcije) promenljive. Numeričke vrednosti koje bismo dobili da smo koristili program za izračunavanje kanonične diskriminacione funkcije su proporcionalne onima iz tabele 11.5. Kako je pomenuto ranije, V_1 je linearna kombinacija X promenljivih sa maksimalnom korelacijom sa U_1, linearnom funkcijom prostih promenljivih koje prikazuju pripadnost grupi. U ovom slmislu, V_1 maksimizuje korelaciju sa prostim promenljivima koje predstavljaju grupe, i stoga ističe maksimum razlika između grupa. Slično, V_2 ističe maksimum razlike pod uslovom da nije u korelaciji sa V_1 itd.

Ono je bila istina u vezi kanonične korelacione analize, važi i ovde: često je prvu kanoničnu diskriminacionu funkciju lakše interpretirati nego one koje slede. Za primer podataka o depresivnim osobama, promenljive pol, zdravlje i broj dana provedenih u krevetu imaju velike negativne koeficijente u prvoj kanoničnoj promenljivoj. Stoga, pacijenti koji su žene sa slabim zdravstvenim stanjem i velikim brojem dana provedenih u krevetu imaju tendenciju da budu depresivne. Drugu kanoničnu promenljivu je daleko teže interpretirati.

Kada postoji dve ili više kanoničnih varijabli, neki od programa prave grafik koji prikazuje prve dve kanonične promenljive za svaku osobu u ukupnom uzorku. Na ovom grafiku svaka od grupa predstavljena je različitim slovom ili brojem. Ovaj grafik je dragocen pri podeli u grupe jer prikazuje maksimalne moguće načine za podelu između grupa. Isti grafik može koristiti i za utvrđivanje odbačenih rezultata i slučajnih grešaka.

Slika 11.6 prikazuje graf dve kanonične promenljive za naš primer. Simboli l, m i h prikazuju tri grupe. Slika takođe prikazuje poziciju srednjih vrednosti (1, 2 i 3) kanoničnih promenljivih za svaku od tri grupe. Ove srednje vrednosti pokazuju da je glavna varijacija prikazana u kanoničnoj promenljivoj 1. Ovaj crtež takođe prikazuje veliki deo preklapanja između osoba – pripadnika tri grupe. Na kraju, nema naročitih ekstremnih vrednosti koje bi trebalo odbaciti, mada slučaj u gornjem desnom uglu može zahtevati još istraživanja.



Slika 11.6. Graf kanoničnih promenljivih za slučaj dve promenljive za podatke o depresiji sa k = 3 grupe

Tamo gde postoje samo dve grupe, postoji samo jedna kanonična varijabla. U ovom slučaju, koeficijenti te kanonične varijable su proporcionalni onima u Fišerovoj diskriminacionoj funkciji. Rezultujući par histograma za dve grupe treba ispitati, pre nego rasejani dijagram. Drugim rečima, nema naročite prednosti pri korišćenju uopštene kanonične procedure. Prednost kanonične procedure se ogleda u njenoj pomoći pri interpretiranju rezultata kada postoji tri ili više grupa.

Razne opcije i dodaci su omogućeni pri klasifikaciji u standardnim programskim paketima, od kojih su neki prikazani u sledećoj sekciji. Detaljnije, stepwise postupak selekcije promenljivih je moguć u većini standardnih programa. Međutim, ukoliko istraživač nije

upoznat sa složenošću date opcije, preporučujemo korišćenje standardnih (default) opcija. Za najveći broj standardnih programskih paketa, promenljiva koja treba da bude uneta je selektovana tako da daje maksimum za statistički test pod nazivom Vilksova lambda. Izuzetak koji mi predlažemo u vezi ove procedure je modifikovanje F-za-unos i F-za-izbacivanje, kako je objašnjeno u prethodnoj sekciji.

Kada je broj grupa, k, veći od dva, istraživač može hteti da ispita diskriminaciju između grupa uzimajući dve odjednom. Ovo ispitivanje će omogućiti da se jasno vide specifične razlike između bilo koje dve grupe, a rezultati mogu postati lakši za interpretaciju. Sledeća mogućnost je da dovedemo u kontrast jednu grupu nasuprot svih ostalih zajedno.

11.13 DISKUSIJA U VEZI KOMPJUTERSKIH PROGRAMA

U ovoj sekciji, prikazaćemo mogućnosti programskih paketa za diskriminacionu analizu i daćemo primer preko programa koji će obraditi podatke o depresivnim osobama.

Osobine programskih paketa

Tabela 11.6 prikazuje stavke koje na izlazu daju standardni programski paketi. BMDP sadrži u sebi jedan program koji može biti korišćen da bi se vršila regularna analiza diskriminacione funkcije ili stepwise analiza diskriminacione funkcije. On takođe omogućava kanonične koeficijente i kanonične rezultate za svaku osobu. To je bazično stepwise program, pa da bi uneo sve promenljive, korisnik može podesiti vrednost za F kao vrlo nisku (kako je prikazano u primeru kasnije u ovoj sekciji) ili odrediti FORCE = 2 bez komande o nivou u paragrafu pod nazivom DISCRIMINANT.

Izlaz	**BMDP**	**SAS**	**SPSS-X**
Sr. vred. i stand. devijacije po grupi	7M	Svi	DISCRIMINANT
Pooled kovarijansa i korelacije	7M	Svi	DISCRIMINANT
Klasifikaciona funkcija	7M	DISCRIM	DISCRIMINANT
D^2	7M	DISCRIM, STEPDISC	
Statističko F	7M	Svi	DISCRIMINANT
Vilks-ova lambda	7M	Svi	DISCRIMINANT
Stepwise opcije	7M	STEPDISC	DISCRIMINANT
Klasifikacione tabele	7M	DISCRIM	DISCRIMINANT
Jackknife klasifikacione tabele	7M	DISCRIM	
Unakrsna provera sa poduzorkom	7M		
Kanonični koeficijenti	7M	CANDISC	DISCRIMINANT
Standardizov. kanonični koeficijenti	7M	CANDISC	DISCRIMINANT
Kanonični grafikoni	7M	CANDISC	DISCRIMINANT
Ranije verovatnoće	7M	DISCRIM	DISCRIMINANT
Kasnije verovatnoće	7M	DISCRIM	DISCRIMINANT

Tabela 11.6. Suma računarskih izlaza iz programa BMDP, SAS I SPSS-X za analizu diskriminacione funkcije

SAS ima tri programa. DISCRIM se preporučuje za analizu diskriminacione funkcije kada ima tri ili više grupa. On takođe uključuje opciju za izvršavanje neparametrijske analize diskriminacijone funkcije koja nije obrađivana tokom ove knjige (videti Lachenbruch 1975; Hand 1981). STEPDISC procedura je namenjena za stepwise analizu diskriminacione funkcije. CANDISC procedura prikazuje podatke na izlazu sa stanovišta kaninične korelacije.

U paketu SPSS procedura DISCIMINANT takođe vrši analizu diskriminacione funkcije po unapred podrazumevanim opcijama ili stepwise analizu korišćenjem opcione METHOD

komande. Ovaj program prikazuje i crta kanonične rezultate, koji se u uputstvu za SPSS nazivaju »diskriminacijskim rezultatima«.

Ispitivanje srednjih vrednosti promenljivih je korisno u određivanju »osećaja« kako se grupe međusobno razlikuju. Kada postoje samo dve grupe, može biti korisno izračunati Mahalanobis-ovo D^2 za pojedinačne promenljive ili grupe promenljivih. Na nesreću, najveći broj programa ne prikazuje ove vrednosti D^2. Za pojedinačnu promenljivu X, vrednost D^2 se lako računa iz grupnih srednjih vrednosti $\overline{X}_I$ i $\overline{X}_{II}$ a ukupna varijansa S^2 je:

$$D^2(za\_jednu\_promenljivu\_X) = \frac{\left(\overline{X}_I - \overline{X}_{II}\right)^2}{S^2}$$

Ove grupne srednje vrednosti i ukupne varijanse se mogu dobiti iz svih programa. Za podsetove promenljivih svaki program prikazuje statističko F preko testa jednakosti.Vrednost ovog F može biti korišćena da bi se izračunalo D^2, kako je prikazano u sekciji 11.9. Napomenimo da BMDP7M prikazuje Mahalanobis-ovo D^2 za svaku osobu. Ova vrednost predstavlja meru rastojanja između tačke koja predstavlja osobu i tačke koja predstavlja ustanovljenu srednju vrednost populacije. Niska vrednost D^2 znači da osoba verovatno pripada toj populaciji.

U sekciji 11.7 govorili smo kako ranije verovatnoće i gubici usled pogrešne klasifikacije mogu biti korišćeni za određivanje vrednosti granične tačke C, u slučaju dve grupe. Ukoliko su gubici jednaki, možemo prepustiti programu da izvrši automatska podešavanja tako što ćemo mu uneti ranije verovatnoće kao ulaz. Ukoliko gubici usled pogrešne klasifikacije nisu jednaki, možemo »zavarati« program tako što ćemo ove gubitke uključiti u ranije verovatnoće, kao što sledeći primer ilustruje. Pretpostavimo da je $q_I = 0.4$, $q_{II} = 0.6$, gubitak(II dato I) = 5 i gubitak(I dato II) = 1. Tada:

podešeno $q_I = q_I \cdot$ *gubici(II dato I) = (0.4)(5) = 2*

i

podešeno $q_{II} = q_{II} \cdot$ *gubici(I dato II) = (0.6)(1) = 0.6*

Obzirom da se ranije verovatnoće mogu sabirati, dalje vršimo podešavanja q_I i q_{II} tako da je njihova suma 2.6:

$$podeseno\_q_I = \frac{2}{2.6}$$

i

$$podeseno\_q_{II} = \frac{2.6}{2}$$

Paketi programa BMDP7M, SAS CANDISC i SPSS-X DISCRIMINANT računaju kanonične vrednosti. Kako je već pomenuto u sekciji 11.12, ove promenljive predstavljaju linearnu kombinaciju promenljivih koje su izabrane da predstavljaju maksimalni nivo separaciju između grupa. Kada je broj grupa dva, samo jedna kanonična promenljiva postoji, i njena vrednost za datu osobu je proporcionalna vrednosti Fišerove diskriminacione funkcije. Program može prikazati histogram ove promenljive za svaku od dve grupe.

Za slučaj postojanja samo dve grupe, podsećamo vas na analogiju između regresije i diskriminacije. Stoga je moguće izvršiti diskriminacionu analizu za dve grupe korišćenjem bilo kojeg programa za analizu regresije, prikazanog u delu 2 ove knjige. Tačnije, svi potprogrami za analizu regresije, kao BMDP9R i SAS REG mogu biti korišćeni za selekciju promenljivih pri

analizi diskriminacione funkcije. Zapamtite da koeficijenti diskriminacione funkcije koji se dobiju iz dva različita programa mogu biti različiti, ali će obavezno biti međusobno proporcionalni.

Primer

Primer ulaznih komandi korišćenih za pokretanje programa za proračun stepwise diskriminacione funkcije BMDP7M je dat ovde za podatke o depresivnim osobama. U ovom primeru, osam promenljivih je razmatrano kao moguće promenljive za diskriminacionu funkciju: pol, starost, obrazovanje, prihod, zdravstveno stanje, broj dana provedenih u krevetu, akutne bolesti i hronične bolesti. Opis ovih promenljivih je dat u kodnoj knjizi u tabeli 3.2. Primetićemo da prve četiri promenljive predstavljaju tipične demografske podatke, dok su poslednje četiri mera zdravlja i bolesti. Na bazi prihoda i starosti isključivo, bili smo u stanju da klasifikujemo 62.6% osoba tačno (videti tabelu 11.3), a sada ćemo videti da li možemo u ovom pokušaju poboljšati taj procenat.

Postoji jedna konfuzija u korišćenju terminologije. U BMDP programima termin *slučaj (case)* se koristi da prikaže broj opservacija. Ali, setite se da smo koristili isti termin da označimo da li je osoba depresivna ili nije, prateći uobičajenu medicinsku terminologiju u kojoj termin *slučaj* predstavlja nekoga ko je bolestan.

Ulazne komande za program BMDP7M izgledaju sledeće:

```
/input      file is 'depress'.
            format is '(8x,f1.0,f2.0,1x,fl.0,1 x,
            f2.0,23x,f1.0,1x,fl.0,2x,f1.0,fl.0,fl.0)'
            variables are 9.
/variable   names are sex,age,educat,income,cases,health,
            beddays,acuteill,chronill.
            group is cases.
/group      code(5) =0,1.
            names(5) = notdep,depress.
/disc       enter=1,1.
            remove = 0,0.
/end
```

Naredba »group« mora se koristiti da bi se označilo kako su dve različite grupe definisane. Ovo je peta promenljiva i uzima vrednost 0 ako je osoba nedepresivna i 1 ako je osoba depresivna. F-za-unos i F-za-izbacivanje vrednosti su postavljene veoma nisko da bi se videlo šta će se desiti. Takođe je moguće dobiti dodatni prikaz korišćenjem paragrafa »print«, ali ćemo ovde korisititi standardne opcije za prikaz.

Program prvo prikazuje srednje vrednosti i standardne devijacije za obe grupe (50 depresivnih i 244 nedepresivne osobe). Ovi rezultati su prikazani u tabeli 11.7.

	Nedepresivna		**Depresivna**	
Promenljiva	**Sred. vrednost**	**St. devijacija**	**Sred. vrednost**	**St. devijacija**
Pol	1.59	0.49	1.80	0.41
Starost	45.24	18.15	40.38	17.40
Obrazovanje	3.55	1.33	3.16	1.17
Prihod	21.68	15.98	15.20	9.84
Zdravlje	1.71	0.80	2.06	0.98
Dani u krevetu	0.17	0.38	0.42	0.50
Akutne bolesti	0.29	0.45	0.38	0.49
Hronične bolesti	0.48	0.50	0.62	0.49

Tabela 11.7. Srednje vrednosti i standardne devijacije za podatke o depresiji

Iz tabele se može primetiti da se standardne devijacije ne razlikuju puno između grupa, ali se za podatke ne može reći da se smatraju višepromenljivim normalnim. Neke od promenljivih uzimaju samo dve vrednosti, 1 i 2 ili 0 i 1. Takođe, prihod deluje »ukošeno«, što je manifestovano preko velike standardne devijacije relativno u odnosu na srednju vrednost i nedostatkom negativnih prihoda. Letimičan pogled na srednje vrednosti pokazuje da su depresivne osobe u većem broju slučajeva žene, mlađe, sa manjim prihodima i nižim obrazovanjem, i slabijeg zdravlja.

U koraku 0, najveći F-za-unos je za broj dana provedenih u krevetu, tako da se on unosi u koraku 1. Prihod se unosi sledeći u koraku 2, pol u koraku 3, starost u koraku 4 i zdravlje u koraku 5. Ostale promenljive se ne unose obzirom da je njihov F-za-unos manji od 1. U poslednjem koraku F-za-izbacivanje je 3.91 za pol, 6.02 za starost, 6.95 za prihod, 4.20 za zdravlje i 8.65 za broj dana provedenih u krevetu. Nije moguće pridružiti određene P vrednosti ovim F-za-izbacivanje nivoima, usled nedostatka normalnosti i korišćenja stepwise procedure, međutim, P vrednosti izgleda da su dovoljno velike da bi promenljive bile ostavljene.

	Klasifikaciona funkcija		
Promenljive	**Nedepresivni**	**Depresivni**	**Diskriminaciona f-ja**
Pol	7.268	7.962	-0.694
Starost	0.132	0.108	0.024
Prihod	0.181	0.151	0.030
Zdravlje	1.793	2.240	-0.447
Dani u krevetu	-0.140	1.119	-1.259
Kontrast	-12.932	-13.724	-0.792

Tabela 11.8. Klasifikacione i diskriminacione funkcije

Klasifikaciona funkcija je data u tabeli 11.8 i predstavlja izlaz programa, a diskriminaciona funkcija se dobija oduzimanjem (videti sekciju 11.6).

Primetićemo da su veličine koeficijenata za starost i prihod prilično slične onima u tabeli 11.4 čak i kada se dodaju tri druge promenljive.

Vrednost F u poslednjem koraku je data kao 7.60 sa 5 i 288 stepena slobode. Ova vrednost testira hipotezu da je populacijsko $D^2 = 0$ i vrlo je značajno ($P < 0.001$). D^2 možemo izračunati (videti sekciju 11.9) za pet promenljivih za $N_I = 244$ i $N_{II} = 50$ kao:

$$D^2 = \frac{(5)(294)(292)}{(50)(244)(294-5-1)}(7.60) = 0.9285$$

Stoga je –D/2 jednako –0.4818, što rezultuje oblašću za oko 0.31 ulevo u odnosu na ovu vrednost u tabeli A.1 dodatka. Nepodešena procena D^2 je:

$$\frac{244+50-5-3}{244+50-2}(0.9285) - 5\left(\frac{1}{50}+\frac{1}{244}\right) = 0.7889$$

(videti sekciju 11.8). Ova procena modifikuje gorenavedenu oblast na 0.33, što je dosta blizu ranijem 0.31.

Program prikazuje kasniju verovatnoću pripadanja depresivnoj i nedepresivnoj grupi za svaku osobu. Za prvu osobu u uzorku, koja je 68-godišnja žena, sa prijavljanim prihodom od 4000 dolara godišnje, zdravstvenim statusom 2 i bez dana provedenih u krevetu (videti tabelu 3.3):

$$\Pr ob(nedepresivna) = \frac{1}{1+\exp(-Z+C)} = \frac{1}{1+\exp(0.530-0.792)} = 0.565$$

gde je C = -0.792 i

Z = -0.694(2) + 0.024(68) + 0.030(4) – 0.447(2) – 1.259(0) = -0.530

Verovatnoća da je osoba nedepresivna je blizu jedne polovine, ali je program ipak klasifikuje kao nedepresivnu. Prva osoba koje biva klasifikovana kao depresivna je peta osoba u tabeli 3.3, koja je 33-godišnja žena, sa godišnjim prihodom od 35000 dolara, zdravstvenim statusom 1 i brojem dana provedenih u krevetu 1. Ova žena je imala, konkretno, ukupan CESD rezultat od samo 6, tako da nije bila depresivna! Međutim, ukupno gledano, program je tokom izvršavanja korektno klasifikovao 71.1% osoba, što prikazuje tabela 11.9. Ovi rezultati su bolji od onih prikazanih u tabeli 11.3.

		Klasifikovani kao		
Stvarni status	**N**	**Nedepresivni**	**Depresivni**	**% Tačno**
Nedepresivni	244	175	69	71.7
Depresivni	50	16	34	68.0
Ukupno	294	191	103	71.1

Tabela 11.9. Klasifikacija osoba kao depresivne ili nedepresivne na bazi pola, starosti, prihoda, zdravstvenog stanja i broja dana provedenih u krevetu

Korišćenjem metoda opisanog u sekciji 11.9, možemo testirati da li je populaciono D^2 za svih pet promenljivih jednako populacionom D^2 za dve promenljive (starost i prihod). Za P = 2 i Q = 3 mi računamo, korišćenjem $D^2_5 = 0.929$ i $D^2_2 = 0.319$:

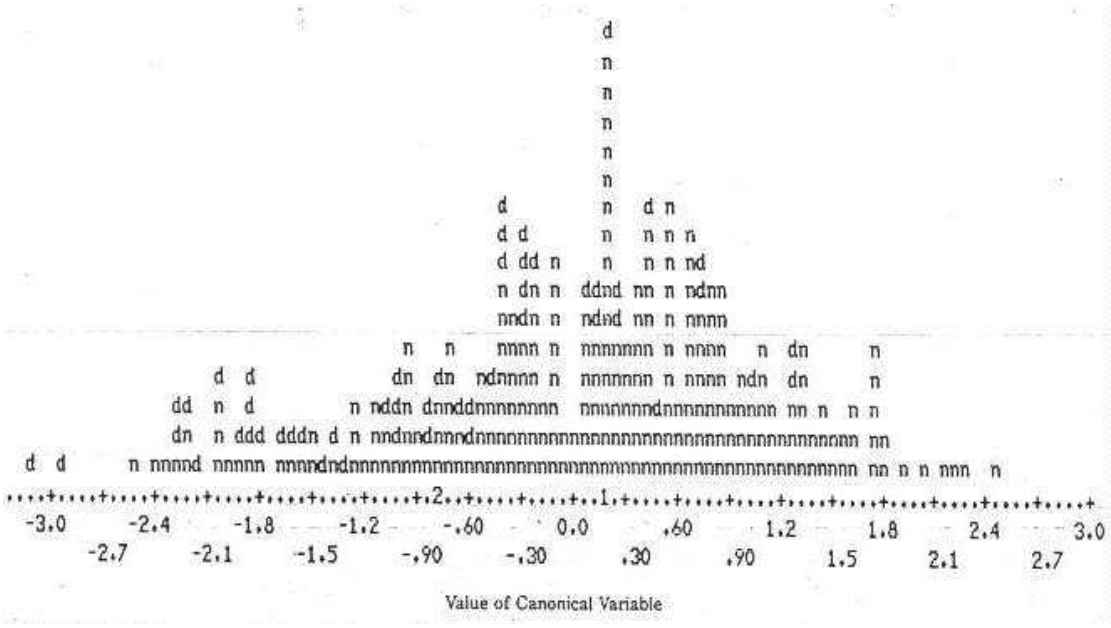
$$F = \frac{244+50-2-3-1}{3} \times \frac{(244)(50)(0.929-0.319)}{(244+50)(244+50-2)+(244)(50)(0.319)} = 7.96$$

sa 3 i 288 stepeni slobode. Obzirom da ovaj rezultat odgovara P koje je manje od 0.005, zaključujemo da pol, zdravlje i broj dana provedenih u krevetu značajno poboljšavaju predviđanje koje je bazirano samo na starosti i prihodu.

Program zatim prikazuje kanonične promenljive za svaku osobu i pravi histogram vrednosti, sa nedepresivnim označenim kao n i depresivnim označenim kao d, kako je prikazano na slici 11.7. Primetite da je u gornjem delu grafika dominantno n, što i treba da bude slučaj. Kada bi bila omogućena bolja diskriminacija, slika bi se sastojala iz dve preklapajuća, ali jasno razdvojiva histograma.

Osoba koju program klasifikuje kao najdepresivniju, obzirom na kanonične varijable i kasniju verovatnoću je 18-godišnja žena sa prijavljenim prihodom od 2000 dolara, zdravstvenim statusom 3 i jednim danom provedenim u krevetu. Ova žena ima CESD rezultat od 39, i stoga ju je program korektno klasifikovao kao depresivnu.

U daljim izvršavanjima programa možemo uvesti druge promenljive, razmotriti transformaciju nekih od promenljivih, na primer logaritamski oblik prihoda, i podešavanja vrednosti gubitaka i ranijih verovatnoća prema željenim vrednostima.



Slika 11.7. Histogram kanoničnih promenljivih dobijen iz programa BMDP7M

11.14 NA ŠTA TREBA PAZITI

Delimično zato što postoji jedna grupa, analiza diskriminacione funkcije ima još neke aspekte, na koje treba obratiti pažnju, a koji nisu spomenuti kada smo govorili o regresionoj analizi. Diskusija o mogućim problemima je data u Lachenbruch-u (1977). Lista važnih i problematičnih oblasti je sledeća:

1. Teorijski, razmatra se jednostavan slučajni uzorak iz svake populacije. Kako ovo u praksi često nije slučaj, uzorak treba ispitati zbog mogućnog potrebnog uvođenja podešavajućih faktora.
2. Kritičnu tačku predstavlja potreba da se napravi tačna identifikacija grupe. Na primer, u primeru datom u ovom poglavlju, smatrano je da sve osobe imaju tačan rezultat na CESD skali tako da mogu biti tačno identifikovane kao normalne ili depresivne. Ukoliko bismo neke od osoba identifikovali pogrešno, ovo bi povećalo nivo greške koju bi prijavio računar. Takođe je posebno problematično ukoliko jedna grupa sadrži više pogrešno identifikovanih osoba nego druga.
3. Izbor promenljivih je takođe veoma bitan. Analogija važi u odnosu na regresionu analizu. Slično kao i u regresionoj analizi, važno je otkloniti outliers, izvršiti neophodne transformacije promenljivih, i proveriti nezavisnost slučajeva. Takođe treba brinuti o tome da li postoji značajno više nedostajućih vrednosti u jednoj grupi nego u drugoj.
4. Višepromenljiva normalnost se podrazumeva u analizi diskriminacione funkcije kada se računaju kasnije verovatnoće ili vrše statistički testovi. Korišćenje prostih promenljivih nije dovodilo do nekih problema, ali vrlo ukošene ili distribucije sa velikim »repom« za neke promenljive mogu povećati ukupan nivo greške.
5. Sledeće što se podrazumeva je jednakost matrica kovarijansi u grupama. Ukoliko se jedna matrica kovarijanse veoma razlikuje od druge, tada treba razmatrati analizu kvadratne diskriminacione funkcije za dve grupe (videti Lachenbruch 1975) ili treba izvršiti transformacije promenljivih koje imaju najveće razlike u svojim varijansama.

6. Ukoliko je veličina uzorka dovoljna, istraživači ponekad dobijaju diskriminacionu funkciju iz jedne polovine ili dve trećine tih podataka i primenjuju je na ostatak podataka da bi videli da li je isti procenat slučajeva klasifikovan tačno u oba poduzorka. Često, kada se rezultati primenjuju na drugačijem uzorku, udeo tačno klasifikovanih je manji. Ukoliko diskriminaciona funkcija treba da služi za klasifikaciju osoba, važno je da početni (originalni) uzorak dolazi iz iste populacije kao onaj na kojeg će se diskriminaciona funkcija primenjivati u budućnosti.
7. Ukoliko su neke od promenljivih dihotomne i jedan od rezultata se retko pojavljuje, tada treba razmisliti o analizi logisitičke regresije (o ovome govorimo u sledećem poglavlju).

ZAKLJUČAK

U ovom poglavlju smo govorili o analizi diskriminacione funkcije, o tehnici koja datira bar još iz 1936. godine. Međutim, njena popularnost je započela uvođenjem moćnih računara tokom 60-ih godina. Originalna primena metoda je bila klasifikacija osobe u jednu od više populacija. Takođe je korišćena za potrebe određivanja relativnih doprinosa pojedinih promenljivih ili grupa promenljivih pri klasifikaciji.

U ovom poglavlju koncentrisali smo se na slučaj dve populacije. Dali smo određenu teorijsku osnovu i prikazali primer korišćenja paketa programa za ovu situaciju. Takođe smo govorili o slučaju kada imamo više od dve grupe.

Čitaoci koje ova tema dalje zanima mogu pogledati reference date u bibliografiji. U suštini, Lachenbruch (1975) i James (1985) govore veoma široko o ovoj temi.

BIBLIOGRAFIJA

Afifi, A. A., and Azen, S. P.1979. *Statistica1l ana1ysis: A computer oriented approach.* 2nd ed. New York: Academic Press.

*Anderson, T. W. 1984. *An introduction to multivariate statistical analysis.* 2nd ed. New York: Wiley.

Costanza, M. C., and Afifi, A. A. 1979. Comparison of stopping rules for forward stepwise discriminant analysis. *Journal of the American Statistical Association* 74:777-785.

Dixon, W. J., and Massey, F. J. 1983. *Introduction to statistical analysis.* 4th ed. New York: McGraw-Hill.

Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7:179-188.

Hand, D. J. 1981. *Discrimination and classification.* New York: Wiley.

James, M. 1985. *Classification algorithms.* New York: Wiley.

Klecka, W. R. 1980. *Discriminant analysis.* Beverly Hills: Sage.

Lachenbruch, P. A. 1975. *Discriminant analysis.* New York: Hafner Press.

Lachenbruch, P. A. 1977. Some misuses of discriminant analysis. *Methods of information in medicine* 16:255258.

*Morrison, D. F. 1976. *Multivariate statistical methods.* 2nd ed. New York: McGraw-Hill.

*Rao, C. R. 1973. *Linear inference and its application.* 2nd ed. New York: Wiley.

*Tatsuoka, M. M. 1988. *Multivariate analysis: Techniques for educational and psychological research.* 2nd ed. New York: Wiley.

Truett, J., Cornfield, J., and Kannell, W. 1967. Multivariate analysis of the risk of coronary heart disease in Framingham. *Journal of Chronic Diseases* 20:511-524.