1.1 Zakaj naj bi uporabljal multivariatno analizo?

Večina stvari ima več vzrokov, zakaj obstaja in kako je nastala, ki so med seboj povezani. Multivariatna analiza je statistično orodje za določanje relativnih prispevkov različnih vzrokov za en sam dogodek ali izid. (primer zakaj potrebujemo multivariatno analizo je primer bolezni, večina bolezni ima več razlogov za nastanek in prognoza je po navadi določena z upoštevanjem velikega števila faktorjev.) Multivariatna analiza nam omogoča razvrstitev faktorjev in njihovega relativnega doprinosa k izidu. (Multivariatna analiza= analiza odnosa med več spremenljivkami)

Obstaja več podobnih analiz, za ugotavljanje odnosov, vendar je z njimi mogoče računati vseh faktorjev in odnosov na izid skupaj, vendar samo vsakega posebej. Ena izmed teh je statificirana analiza (razdelitev po skupinah) ocenjuje učinek faktorja tveganja na izid z upoštevanjem ostalih variabilnih konstant. Je bivariatna. (Bivariatna analiza= analiza odnosa med dvema spremenljivkama (korelacija), napovedovanje vrednosti Y na podlagi X (regresija))

1.2 Kaj je »zmeda« (confounder) in kako mi multivariatna analiza pomaga pri reševanju tega problema?

Sposobnost multivariatne analize da lahko hkrati oceni neodvisne prispevke velikega števila faktorjev tveganja za izid je posebej pomembna ko imamo »zmedo«. »Zmeda« se zgodi imajo na očitne povezave med faktorji in izidom vpliv tudi zveze z tretjo spremenljivko faktorja tveganja na izid. Tretja spremenljivka se imenuje »zmeda.«

Da je spremenljivka »zmeda«, mora biti povezana z faktorjem tveganja in vzročno povezana z izidom. (primer zmede je povezanost med nošenjem vžigalic in pljučnim rakom. Oseba, ki nosi (ima) vžigalice ima večjo možnost za razvoj pljučnega raka; spremenljivka »zmeda« je kajenje. In očitno je da nošnja vžigalic ne povzroča raka, zato je to najboljša razlaga kaj »zmeda« je.) Stratificirana analiza se uporablja za ocenjevanje in odpravljanje »zmede«. Če stratificiraš status kajenja (razdeliš po skupinah, vzrokih…) ugotoviš, da nošnja vžigalic ni povezana z pljučnim rakom. (Torej ni povezave med vžigalicami in rakom ko posebej preučiš kadilce in nekadilce.) Statistični dokaz »zmede« je razlika med stratificirano in nestratificirano analizo.

Razlike med bivariatnimi in multivariatnimi analizami kažejo na to, da je spremenljivka »zmeda« prisotna. Prednost multivariatnih analiz pred stratificiranimi je ta, da ni potrebno razslojevanje vseh faktorjev, vzrokov, povezav..

1.3 Kaj je »zaviralec« (suppressor) in kako mi multivariatna analiza pomaga pri reševanju tega problema?

Zaviralec« je poseben tip spemenljivke »zmede«. Kot pri »zmedi« je tudi »zaviralec« povezan z faktorjem tveganja in izidom. Razlika je ta, da pri bivariatni analizi ni vidnega vpliva faktor tveganja in izid. Vendar ko se prilagodiš spremenljivki »zaviralec«, zveza med faktorjem tveganja in izidom postane pomembna. Prepoznavanje in prilagajanje »zaviralcu« lahko pripelje do pomembnih ugotovitev. (primer jemanja zdravila, ki naj bi z statistično nepomembnostjo zaviralo HIV pri delavkah, ki so v stiku z HIV okuženimi iglami. Jemanje zdravila in učinkovitost po rezultatih študije ni v sorazmerju z izpostavljenostjo iglam v bivariatni analizi, vendar ta podatek spremeni in postane pomemben ko se raziskava prilagodi spremenljivki resnosti poškodbe z iglo – spremenljivka »zaviralec«, učinek zdravila se statistično poveča)

Čeprav je bil ta primer multivariatne analize predstavljen na učinkovitosti zdravila po prilagoditvi spremenljivki »zaviralec«, je potrebno vedeti, da se multivariatne analize ne morejo prilagajati potencialnim pristranskostim v analizi. (Na izid tega primera študije je najverjetneje vplivala razlika v populaciji med testno in kontrolno skupino, saj povečana uporaba pri skupini ki so se ranili dvigne uporabo zdravila in posledično učinkovitost zdravila.)

1.4 Kaj so interakcije in kako nam multivariatne analize pomagajo se soočiti z le-temi?

Interakcija se zgodi, ko se spremeni vpliv dejavnika na izid z vrednostjo tretje spremenljivke. Včasih jo označujemo tudi kot »učinek« sprememba, saj na vpliv dejavnika tveganja vpliva tudi rezultat druge spremenljivke. Torej je učinek faktorja tveganja odvisen od vrednosti interakcij spremenljivke (medsebojen vpliv obeh). V skrajnih primerih lahko interakcija popolnoma spremeni razmerje med faktorjem in rezultati. To se zgodi, ko dejavnik tveganja poveča verjetnost rezultata na različnih vrednostih variabilne interakcije. Efekt faktorja tveganja na rezultat je večji (ali manjpi) na določenih vrednostih tretje spremenljivke.

Definicija: Interakcija se zgodi, ko faktor tveganja vpliva na rezultat in je le-ta spremenjen zaradi vrednosti tretje spremenljivke.

Razslojevanje nam je lahko v pomoč pri identificiranju interakcije. Z razslojevanjem interakcije spremenljivke, lahko opazujemo efekt faktorja tveganja na različnih vrednostih spremenljivke (ki je v interakciji). To lahko tudi preverimo s stratifikacijo – se pravi – ali je povezava med dejavnikom

(faktorjem) tveganja in rezultatom na različnih ravneh interakcijske spremenljivke je statistično drugačen od tiste, pri kateri uporabimo hi-kvadrat preizkus homogenosti. A vendar ima tudi stratifikacija kot uporabo za odpravo motenj – ima svoje pomanjkljivosti. Težavno je razslojiti po več kot ene ali dve spremenljivke, ker imamo ponavadi več podatkov. Statifikacija bo natačno ocenila vpliv dejavnikov tveganja na izid na različnih ravneh interakcijskih spremenljivk, ta analiza ne bo prilagodljiva za druge spremenljivke v modelu (npr. nagnjenost), ki lahko vplivajo na odnos med dejavniki tveganja in rezultati. Multivariatne analize nam tako omogočajo da vključimo pogoje interakcije in jih ocenimo, medtem ko se prilagajamo na druge spremenljivke.

2.poglavje

2.1. Kateri so najpogostejši načini uporabe multivariabilnih modelov v kliničnih raziskavah?

Štiri najbolj pogosti načini:
*opazovalne študije etiologije
* intervencijske študije (randomizirane in nerandomizirane)
*Študije diagnoze
* Študije prognoze

2.2. Kako se multivariabilna analiza uporablja v opazovalnih študijah etiologije?

Cilj etioloških študij je ugotoviti vzroke za izid, po navadi z odstranjevanjem škodljivih snovi (npr. tobak) ter spodbujanjem zdravih aktivnosti (npr. vadba). Čeprav opazovalne študije ne morejo dokazati vzročnosti, lahko multivariabilna analiza okrepi argument vzročnosti, z izključitvijo zavajajočih razlag o povezavi.

Razlog, zakaj je multivariabilna analiza postala nepogrešljiva za opazovalne študije o etiologiji je, da ker smo izvedeli več o nekaterih več faktorskih boleznih (kot so srčne bolezni), smo lahko identificirali tudi vse večje število dejavnikov tveganja bolezni. Ker je veliko teh spremenljivk povezanih med seboj, statifikacija postane toga tehnika odpravljanja motečih spremenljivk.

Na žalost pa imajo multivariabilne analize tudi svoje omejitve kadar se uporabljajo za opazovalno študijo o etiologiji. Največja pomanjkljivost pa bi lahko bila, da je nemogoče prilagoditi tiste spremenljivke, ki niso izmerljive ali pa so neznane.

Še eno omejitev predstavlja, da tudi vključena potencialna zavajajoča spremenljivka (»confounder«) v modelu ne zagotavlja odprave pristranskosti (zaradi) le te. Da se multivariabilnim modelom ustrezno prilagodijo morajo zavajajoče spremenljivke v različnih skupinah ali izidih sovpadati.
Primer: Če so v eni skupini skoraj vsi kadilci v drugi pa skoraj vsi nekadilci, prilagajanje kajenju kljub temu ne bo odstranilo motečih/zavajajočih dejavnikov, ki jih povzroča kajenje. Zato je pomembno, da uporabimo bivariantno analizo, da preverimo ali je na voljo dovolj prekrivanja med potencialnimi zavajajočimi spremenljivkami, pred izvedbo multivariabilne analize.
Čeprav je na voljo dovolj prekrivanja, prilagoditev ni popolna. Tako kot je napaka pri merjenju odvisnih in neodvisnih spremenljivk, je napaka tudi pri zavajajočih spremenljivkah (confounder).

2.3. Kako se multivariabilna analiza uporablja v intervencijskih študijah (randomizirana in nerandomizirana)?

Pri analizi podatkov intervencijskih poskusov, če so poskusi randomizirani ali nerandomizirani, če gre za intervencijo droge, svetovanja ali spremembo v zakonodaji, je ključnega pomena, da se prilagodi osnovne razlike v skupinah. Zakaj? Ker, če obstajajo razlike med skupinami že na začetku študije, je nemogoče vedeti ali so razlike na koncu študije nastale zaradi intervencije ali zaradi razlik med skupinami. Primer: kaj, če je bila oseba, ki je prejela zdravilo starejša od osebe, ki je prejela placebo... Ne moremo vedeti ali so razlike v skupinah na koncu študije nastale zaradi, npr. zdravila, ali zaradi osnovnih razlik v skupinah (npr. starost).

V nerandomiziranih intervencijskih študijah so bistvene izhodiščne razlike med skupinami (so bolj pravilo kot izjema), saj se v življenju ljudje ne razporedijo kar sami (v enake skupine-brez izhodiščnih razlik). Za ljudi, ki jemljejo zdravila je bolj verjetno, da so bolni, ali imajo boljši dostop do zdravstvene oskrbe,... Ljudje, ki se prijavijo na svetovanje so verjetno bolj motivirani za zdravje,...

2.4 Kako je analiza več spremenljivk uporabna v študijah diagnoze:

Modeli več spremenljivk lahko razkrijejo najboljšo kombinacijo diagnostičnih informacij za določitev, ali oseba trpi za določeno boleznijo. V največji meri so diagnostične algoritme oblikovali v namen diagnoze miokardnega infarkta pri pacientih, ki se na zdravniški pregled prijavijo zaradi bolečin v prsih – to pa zaradi visokega tveganja. Bolečine v prsih so pogost simptom na urgentnih oddelkih, pojavijo se pa lahko zaradi lažjih

težav, kot je zgaga, ali nevarnejših, kot je srčni napad. Na urgentnih oddelkih vsepovsod se zdravniki redno odločajo, koga poslati domov in koga na koronarni oddelek. Čeprav je slednji učinkovit pri reševanju življenj bolnikov z akutno ishemijo, manj kot polovica pacientov dejansko trpi za tem stanjem. Trenutno ni na voljo nobenega testa, ki bi učinkovito ločil paciente, ki drago terapijo potrebujejo od tistih, ki je ne.

Pozen in kolegi so razvili tak diagnostični model, ki preverja verjetnost, da bolnik, ki toži nad bolečinami v prsih, tudi trpi za akutno ishemijo. Od 59 kliničnih znakov so jih izbrali 7 – ti, ko se jih uporabi skupaj v logističnem regresijskem modelu (*Opomba: Logistična regresija je metoda, ki nam, na podlagi več neodvisnih spremenljivk, omogoča napovedati verjetnosti izida, ki lahko zavzame vrednosti 0 ali 1*) – tvorijo napoved prisotnosti ishemije na intervalu od 0 do 1,0. Da bi določili uporabnost modela, so raziskovalci zdravnikom na urgencah za pomoč pri odločanju posredovali rezultate modela. Ugotovili so, da so se zdravniki z dostopom do rezultatov modela odločali občutno bolje: primerov, napotenih na koronarni oddelek je bilo 30% manj brez spregledanih pojavov akutne ishemije.

Na žalost kljub temu zdravniki na urgencah še ne uporabljajo statističnih izračunov verjetnosti za akutno ishemijo pri odločanju. Corey in Merenstein sta testirala sprejemljivost modela na urgentnih oddelkih tako, da sta jih preoblikovala v pregledne delovne liste in osebju na oddelkih omogočila priročen dostop do njih. Zdravniki so jih uporabili le v 2.8% primerov. Podobna neuporaba preverjenih diagnostičnih orodij je bila opažena tudi drugod.

Razlogi za neuporabo diagnostičnih pravil oziroma orodij so zapleteni. Velik vpliv na to ima omejenost zdravnikov s časom. Pozenov algoritem se da izračunati v borih 20 sekundah, a mora biti pri tem uporabljen posebej prednastavljeno računalo – česar pa zdravniki po navadi ne nosijo s sabo okrog. Uporaba tega algoritma v obliki delovnega lista bi vzela pol minute do minuto časa za izračun verjetnosti prisotne akutne ishemije. Minuta se nam ne zdi dolga, a na urgentnem oddelku je vsaka sekunda dragocena.

Drug razlog za neuporabo diagnostičnih algoritmov je najverjetneje psihološkega izvora – zdravniki opravljajo prakso z bolj izkušenim od sebe do trenutka, ko so sami dovolj izkušeni, da lahko delo opravljajo sami. Takrat pa jih je večina tako prepričana v svoje sodbe, da se jim posluževanje diagnostičnih algoritmov ne zdi potrebno. V računalniških modelih za diagnozo pa vseeno obstaja potencial.

Diagnostični modeli morajo biti zanesljivi, z visoko napovedno vrednostjo, zaradi česar je razvoj takih algoritmov izziv. Kljub temu pa jih je lažje

oblikovati kakor pojasnjevalne modele. Simptom oziroma znak, povezan z določeno diagnozo, lahko nima kavzalne povezave z boleznijo samo, je pa še zmeraj uporaben pri diagnozi (oblika ušesne mečice pri napovedi verjetnosti srčnega infarkta, na primer).

2.5 Uporaba analize več spremenljivk v študijah prognoze:

Na pacientova vprašanja, kot so: »Se bo rak spet pojavil?«, »koliko časa bom še živel?« in podobna je zdravniku težko odgovoriti, hkrati pa ima bolnik pravico do iskrenega odgovora. Analiza več spremenljivk nosi potencial za oblikovanje prognoz s pomočjo znanih prognostičnih faktorjev.

Schuchter je s kolegi izvedel študijo na bolnikih s kožnim rakom. Zanimalo jih je, kateri faktorji napovedujejo preživetje bolnikov po desetih letih. Po desetih letih je preživelo 78 odstotkov ljudi v vzorcu (N=488). Na podlagi sledenja vzorcu bolnikov v obdobju desetih let so raziskovalci z logistično regresijo oblikovali prognostični model, ki je izid napovedal pravilno v 74 odstotkih. Pri napovedi so se izkazali pomembni štirje faktorji: starost, spol, lokacija lezije na telesu in debelina le te.

Na videz se zdi, da lahko brez kakršnihkoli prognostičnih informacij oblikujemo bolj učinkovit model, kot so ga Schuchter in kolegi – preprosto bi lahko rekli, da bolnik po dobi desetih let še živi, in prav bi imeli v 78 odstotkih primerov. To je sicer res, a ne bi mogli pravilno napovedati nobene od smrti.

Prognostični modeli omogočajo ustrezne napovedi tveganja le za paciente s podobnimi značilnostmi, kot so jih imeli ti v preučevani populaciji – če je bil model osnovan na vzorcu 50-letnikov, nam ne bo pomagal pri napovedi preživetja 45-letne ženske. Poleg tega prognostični modeli delujejo le v primeru, ko so nam znani prognostični faktorji, splošno uporabni pa so le v primeru, ko imamo enostaven dostop do teh spremenljivk (če med prognostične faktorje spada genetska predispozicija posameznika, bi morali na pacientih izvesti genetske analize, kar pa ni praktično). Prognostični modeli tudi niso nikoli tako učinkoviti pri napovedi za paciente iz populacije, kot so za paciente iz vzorca, na katerem je bil model oblikovan.

3. »Outcome« spremenljivke v multivariabilni analizi

3.1. kako narava »outcome«v spremenljivke vpliva na to, kateri tip multivariabilne analize bomo uporabili?

na izbiro multivariabilne analize prvotno vpliva tip spremenljivke, ki so lahko: intervalne, dihotomne, ordinalne, nominalne in časovno odvisne. V

nekaterih primerih je mogoče tudi spremeniti naravo spremenljivke, ali pa testirati naenkrat več kot eno obliko spremenljivke (v poglavju 3.12)

Posebne metode pa so potrebne za longitudinalne študije ko je predmet preverjanja ugotovljen preko različnih rezultatov/zaključkov. (primer: krvni pritisk na 6 mesecev, na 1 leto)

primeri »outcome« spremenljivk (za osvežitev znanja)

*intervalne (teža, temperatura)
*dihotomne (smrt, rak)
* ordinalne (stopnja bolezni, resnost simptomov)
*nominalne (vzrok smrti)
*časovno odvisne (čas smrti)

3.2. KAKŠEN TIP VEČDIMENZIONALNE ANALIZE BI UPORABILI PRI INTERVALNIH IZIDIH

Pri intervalnih spremenljivkah ima vsaka enota (interval) na lestvici enako kvantitativno spremembo. Primeri intervalov: krvni pritisk, telesna teža in temperatura. Pri teh primerih je sprememba za eno enoto enaka povsod na lestvici npr. mililiter živega srebra na termometru. Če bi izvedli bivariantno analizo z intervalnim izidom bi uporabljali t test (z dihotomno neodvisno spremenljivko), korelacijski koeficient ali linearno regresijo (z intervalno-neodvisno spremenljivko), analizo variance (z nominalno neodvisno spremenljivko) ali Spearmanov rank korelacijski koeficient (z ordinalno ali ne-normalno intervalno-neodvisno spremenljivko). Med temi tehnikami lahko samo linearna regresija in analiza variance vključujejo multiple neodvisne spremenljivke: opisane so spodaj.

3.2.A -MULTIPLA LINEARNA REGRESIJA

Multipla linearna regresija določi najboljšo napovedovanje končne spremenljivke na podlagi vrednosti neodvisnih spremenljivk.
-KONČNE SPREMENLJIVKE V ANALIZI Z VEČ SPREMENLJIVKAMI
Linearna regresija z najmanj kvadratki (najbolj uporabljena metoda linearne regresije) določi linijo, ki zmanjša razdaljo med točkami podatkov in samo črto/linijo. Statistični modeli so v najboljšem primeru približki podatkov. Uporaba bivraiantne statistike je zadovoljiva če na odnos, ki ga preučujemo ne vpliva še kakšna spremenljivka. Zato je bolje uporabiti multiplo analizo, da zaobjamemo vse spremenljivke, ki bi lahko vplivale na rezultate
ANALIZA VARIANCE ANOVA

tip	indikacija
ANOVA-analiza variance	Primerja dve ali več skupin na intervalnem izzidu. Lahko inkorporira kategorično neodvisne

	spremenljivke in interakcijo kategoričnih spremenljivk z glavnim učinkom
ANCOVA-analiza kovariance	Podobno kot analiza variance ampak vključuje nepretrgane in tudi kategorično neodvisne spremenljivke.
MANOVA-multivariatna analiza variance	Podobno kot analiza variance ampak se uporablja, ko imamo več kot eno odvisno spremenljivko. Zmanjša možnost napake tipa I.
MANCOVA-multivariatna analiza kovariance	Podobno analizi kovariance ampak se uporablj,a ko imamo vek kot eno odvisno spremenljivko. Zmanjša možnost napake tipa I.
Analiza ponavljajočih opazovanj variance/kovariance	Podobno kot analiza variance/kovariance ampak vključuje tudi ponavljajoča opazovanja istih subjektov (poglavje 11.3).
Multivariatna analiza ponavljajočih opazovanj variance/kovariance	Podobno kot multivariatna analiza variance/kovariance ampak vključuje tudi ponavljajoča opazovanja istih subjektov (poglavje 11.3).

Analiza variance (ANOVA) se lahko uporablja za analizo odnosa med več neodvisnimi spremenljivkami in končno spremenljivko. Beseda varianca v imenu tehnike se nanaša na razliko med vrednostmi individualnega subjekta in povprečja.

- »med-skupinska«varianca je osnovana na razliki med subjektom in povprečjem vzorca
- »znotraj skupinska« varianca je osnovana na razliki med člani skupine in skupinskim povprečjem

Če so povprečja posameznih skupin zelo različna med sabo, bo varianca, izračunana na podlagi povprečij celotnega vzorca (medskupinska varianca) večja od variance, izračunane na podlagi povprečja posamezne skupine (znotrajskupinska varianca). Če predvidevamo da je vzorec dovolj velik bo

razlika statistično pomembna. če so povprečja posamezne skupine podobna, bo tudi varianca, izračunana na podlagi celotnega vzorca podobna varianci, izračunani na podlagi posamezne skupine.
Več tipov analize variance poznamo. Najenostavnejši tip ANOVA je eno-smerni (enofaktorski) pri katerem primerjamo dve ali več skupin za eno intervalno spremenljivko. ANOVA in podobni postopki se lahko uporabljajo tudi pri multivariantnih vprašanjih. Dvo-smerna ANOVA se lahko uporabi za determiniranje učinka (1) po skupinah, (2) drugih kategoričnih spremenljivkah (npr. spol, narodnost) in (3) interakcijo med skupino in kategorično spremenljivko na intervalnem izidu.
Če so neodvisne spremenljivke, ki jih moramo vključiti intervalne (starost, teža) potrebujemo analizo kovariance (ANCOVA). Z ANOCVA lahko inkorporiramo intervalne in kategorične spremenljivke.
Zaradi podobnosti med analizo variance in analizo kovariance lahko vaš računalnik avtomatsko izbere katero bo uporabil, na podlagi tega ali vnesete intervalno-neodvisne spremenljivke.
MANOVA-multivariantna analiza variance in MANCOVA- multivariantna analiza kovariance so razširitve analize variance in kovariance in se uporabljajo za preučevanje več kot ene odvisne spremenljivke. Spremenljivke, ki jih dobimo so večinoma v korelaciji. Ti postopki se uporabljajo, da bi zmanjšali možnost napake tipa I (napačna zavrnitev nulte hipoteze). Postopek MANOVA je multivarianten F. Če so testi multivariance F statistično pomembni izračunamo tudi bivariantne F testne vrednosti. V nasprotju če multivariantni F testi niso pomembni ignoriramo individualne F teste. Multivariatni dizajn zahteva, da so interkorelacije izidov spremenljivk homogene v vseh celicah dizajna.
Prilagoditve analize variance in analize kovariance, multivariatne analize variance in multivariatne analize kovariance se uporabljajo tudi za ponavljajoče opazovanje posameznikov.

3.2.C OSNOVNE PREDPOSTAVKE MULTIPLE LINEARNE REGRESIJE IN ANOVA

osnovne predpostavke multiple linearne regresije in ANOVA so enake. Pri obeh je lahko številčni izid negativen ali pozitiven. Obe tehnike predpostavljajo, da ima izid normalno distribucijo, enako varianco okoli povprečja za katerokoli vrednost neodvisne spremenljivke. Da bi izpolnili ta pogoj mora imeti končna spremenljivka krivuljo v obliki zvona za katerokoli vrednost neodvisne spremenljivke. S histogramom odvisne in neodvisne spremenljivke lahko očistimo svoje podatke in ta nas tudi obvesti o možnih prekršitev predpostavk o normalni distribuciji in enakovredni varianci. Histogrami omogočajo tudi detekcijo malo verjetnih vrednosti (npr. starost 120let) in pomagajo najti luknje v vrednostih (če imaš maloštevilen vzorec starostnikov nad 60 let ne moreš generalizirati na to starostno skupino.

Univaritna statistika tudi pomaga zaznati ekstremne vrednosti (outlajerje), ki bi vplivali na rezultate (poglavje 9.5)
Če je vzorec vsake od neodvisnih spremenljivk večji od 100 lahko predvidimo normalno distribucijo (če ni ekstremnih vrednosti). Samo ekstremne vrednosti, ki odstopajo od enotne variance lahko vplivajo na rezultate to pa bi zmanjšalo moč analize, ki prikazuje povezavo med neodvisno in odvisno spremenljivko.
IZBIRANJE MED MULTIPLO LINEARNO REGRESIJO IN ANOVA
Obe tehniki podata enake rezultate če nastavimo modele na podoben način. Predpostavimo, da imamo več kot dve skupini za analizo podatkov z multiplo linearno regresijo zato bomo morali ustvariti multiple dihotomne spremenljivke, ki predstavljajo skupine (poglavje 4.2). Večinoma se multipla linearna regresija uporablja pri podatkih opazovanja (v medicini), analiza variance pa pri eksperimentih (psihologija).

3.3. katera vrsto multivariabilne analize uporabimo pri DIHOTOMNIH spremenljivkah?

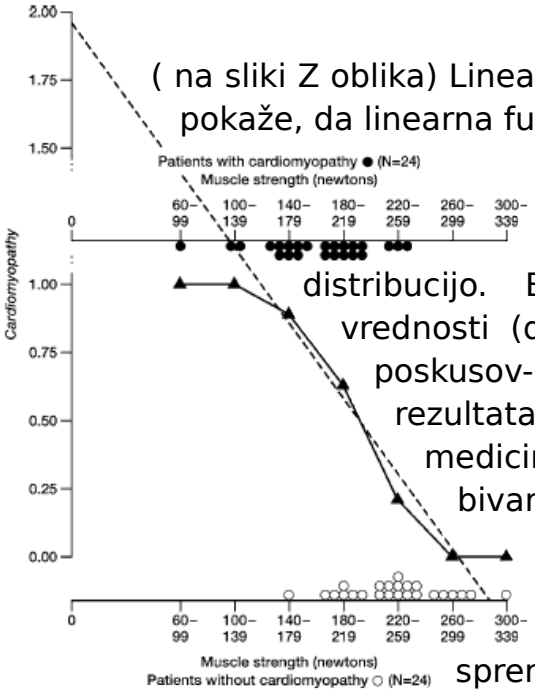
- dihotomne spremelnjivke so spremenljivke ki imajo dve diskretni vrednosti (kategoriji): živ/mrtev, da/ne

*za bivariantno analizo dihotomnih spremenljivk upotabljamo hi-kvadrat test, pri neodvisnih dihotomnih in nominalnih Fisherjev eksaktni test, intervalne neodvisne – t-test, hi-kvadrat in Mann-Whitney test za ordinalne in ne normalne ordinalne neodvisne spremenljivke

A vse te teste se neda uporabiti, če imamo več kot eno neodvisno spremenljivko! Za multivariablino analizo dihotomnih spremenljivk uporabimo **multiplo (binarno) logistično regresijo**.

primer: povezava trdnosti skeletnih mišic in prisotnostjo kardiomiopatije pri alkoholikih *(za posebej vedoželne: Kardiomiopatije so bolezni srčne mišice (miokarda), ki so posledica degenerativnih ali/in vnetnih procesov)*
Pri nižji mišični moči je veliko bolnikov z kardiomiopatijo in nič pacientov z normalno srčno funkcijo. Pri visokem livelu mišične moči je tudi malo bolnikov z to boleznijo. Krivulja kaže na to da se z zmanjševanjem deleža bolnikov povečuje moč mišic, verjetnost kardiomiopatije je kot pri vseh drugih primerih večja od 0 in manjša od 1. vrednost logistične regresije je da vsebuje to predpostavko.

LOGIT: Logistična regresija modelira »logit« (logaritemska funkcija) rezultatov. Logit je naravni logaritem možnih izidov.Delež izzida je verjetnost da imamo izzid deljeno z verjetnostjo da nimamo izida. Izzidi se gibljejo v razponu od 0-1. To nam poda logistično funkcijo v obliki S ali Z.



(na sliki Z oblika) Linearna premica ki je črtkano narisana, pa nam pokaže, da linearna funkcija ni dovolj ustrezna.

Z logistično regresijo domnevamo da ima odvisna spremenljivka binomsko distribucijo. Binomska porazdelitev opisuje številne vrednosti (da/ne, smrt/preživetje) v seriji neodvisnih poskusov-rezultat enega poskusa je neodvisen od rezultata drugega. (npr. niso izbrani iz istih družin, medicinskih praks,..) . Model na sliki predstavlja bivariantno odvisnost med mišično močjo in kardimiopatijo. Prednost logistične regresije nad bivariantnimi tehnikami je da lahko dostopamo do razmerja neodvisne spremenljivke na dihotomni izid z prilagoditvijo na druge faktorje.

Primer: logistična regresija je uporabna za prikaz kardiopulmonalnem oživljanju (CPR) saj lahko afektivno opravimo le z opritiskom na prsi brez umetnega dihanja.

3.4 3.3. Katera vrsto multivariabiLne analize uporabimo pri ORDINALNIH spremenljivkah?_

Ordinalne spremenljivke imajo vče kategorij, za bivariantno analizo le teh uporabimo: hi kvadrat test, Mann-Whitney test, Kruskal-Wallis test, Spearmanovo korelacijo med intervalno in ordinalno spremenljivko.

Za multivariabilno analizo uporabimo **sorazmerno verjetnostno regresijo**. Ocenjuje odnos med neodvisno spremenljivko in ordinalno spremelnjivko. Če imamo npr. 5 ordinalnih kategorij obstajajo 4 različne presečne točke. Če uporabimo ta pristop lahko ocenimo logit (naravni logaritem možnih izidov) in posledično povprečje logitov za posamezne presečne točke(cut points). prednost tega pristopa je torej, da imamo eno oceno učinka vsake neodvisne spremenljivke namesto 4 različnih ocen, ki ustrezajo 4 različnim presečnim točkam.
primer: identificiranje prediktorjev za resnost kapi pri pacientih, resnost kapi so razvrstili v tri ordinalne livele: 1-blažja kap, 2-velika kap, 3-huda kap. ker imamo 3 livele je mogoče narediti 2 presečne točke. 1 vs 2 in 3, 1 in 2 vs 3. ?

3.5 katero vrsto multivariatne analize uporabimo za nominalne spremenljivke

Nominalne spremenljivke so posebna kategorija spremenljivk, ki nima nobenih posebnih določil za računanje. Sama sprememba spremenljivke v numerično predstavlja le posebne kategorije nominalne spremenljivke in same številke nimajo nobene številske vrednosti (na primer 1= rak dojk, 2= rak debelega črevesja, 3= pljučni rak, ...), posledično iz teh številk ni smiselno računati povprečja in mediano. Zato za računanje nominalnih spremenljivk uporabimo multinominalno logistično regresijo, pri kateri moramo prevzeti eno referenčno kategorijo za računanje. Referenčna kategorija je lahko tista, ki je največja oziroma tista, na kateri želimo delati kontrast. Sama izbira referenčne kategorije ne vpliva na matematične odgovore ampak vpliva na izbiro načina poročanja. Model multinominalne logistične regresije primerja vsako kategorijo z referenčno kategorijo, pri tem primerjanju pa dobii razmerje tveganja. Število primerjav je za eno manjši kot je posameznih kategorij, na primer če imamo 4 kategorije (A, B, C in D), primerjamo naenkrat naslednje kategorije B proti A, C proti A in D proti A, ne primerjamo pa kategorije B proti C ali C proti D. Prav tako pa lahko multinominalno logistično regresijo 'razdelimo' na več enostavnih logističnih regresij. Ta model z razliko od multinominalnega na enkrat ocenjuje razmerje tveganja le med eno izmed kategorij in med referenčno skupino. Če se računanja razmerja lotimo na enostavnejši način, bo končni rezultat razmerij tveganja približno enak. Razlika je predvsem v tem da pri multinominalni regresiji dobimo rezultat odnosa neodvisnih spremenljivk na celotnem vzorcu, pri enostavni logistični regresiji pa rezultat odnosa temelji na posameznih vzorcih.

3.6 katero vrsto multivariatne analize uporabimo za časovno odvisne spremenljivke?

Časovno odvisne spremenljivke se nanašajo na študijo dogodkov, kot so smrt ali razvoj raka, ki se pojavijo v določenem časovnem obdobju. Za bivariatno analizo časovno odvisnih spremenljivk narišemo Kaplan-Meierjevo krivuljo za vsako skupino in jih med seboj primerjamo s pomočjo 'log-rank' statistike, katera pa ne more vključiti več neodvisnih spremenljivk.

Za izvedbo multivariantne analize časovno odvisnih spremenljivk se uporablja proporcionalna/sorazmernostna analiza tveganj. Ta za razliko od drugih multivariantnih tehnik omogoča vključitev predmetov oziroma spremenljivk z različno analizno dolžino spremljanja. Različne dolžine se po navadi pojavljajo pri longitudinalnih študijah različnih razlogov (smrt udeleženca, udeleženec ne želi več sodelovati, pojav stranskega učinka zato je potrebna izključitev iz poskusa, ...) Spopadanje s subjekti, ki predčasno 'izstopijo' iz študije je eden od pomembnejših pogojev, zaradi katerega se klinična raziskava loči od drugih študij. V laboratorijskih

raziskavah, je običajno mogoče pogoje nadzirati, tako da se ne izgubili noben opazovan pogoj. Večina družboslovnih raziskav (na katerih temeljijo klinične raziskave) se izvaja z uporabo presečnih modelov, kjer vedno obstaja tveganje da se udeleženci ne bodo odločili za sodelovanje, kar pomeni da v longitudinalnih študijah izgubimo podatke o samem udeležencu. Ena rešitev za take podatke je da jih preprosto izbrišemo, vendar pa izpustitev podatkov zmanjša moč študije in lahko v samo študijo uvaja pristranskost. Drug način spopadanja s takimi podatki pa je tehnika cenzuriranja, ki nam omogoča, da udeleženci prispevajo informacije toliko časa, dokler ne zapustijo same študije. Dejansko so vsi udeleženci/podatki v proporcionalni oz. sorazmernostni analizi tveganj enkrat cenzurirani, če ne med samo študijo pa takrat, ko se študija zaključi.

DEFINICIJA: Cenzuriranje se uporablja za vključitev tem/podatkov z različnimi dolžinami spremljanja zaradi različnih razlogov.

Osnovna predpostavka cenzuriranja je, da, če bi predmetu/udeležencu lahko 'sledili' po času ko smo ga cenzurirali, bi imel enako stopnjo izida kot tisti, ki niso cenzurirani v tistem času. Z drugimi besedami to pomeni, da do cenzuriranja pride naključno, neodvisno od rezultata.

3.12 Ali lahko spremenimo kodiranje “outcome” spremenljivke z uporabo različnih tipov multivariabilnih analiz.

Poudarek jena izbiri ustrezne vrste multivariabilne analize, ki temelji na vrsti (nature) naše “outcome” spremenljivke. Zdi se tako, kot , da ti vržem ukrivljeno žogo in čakam možnosti spremembe “outcome” spremenljivke z uporabo druge vrste analize. Včasih obstajajo tudi prednosti pri anlizah podatkov , če je le možno ugotviti ali nam da drugačen način podobne ali različne rezultate.

4. Neodvisne spremenljivkev multivariabill analizi

4.1Kako vključim neodvisne spremenljivke v multivariabilne analize

Ugotovitev, ki temelji na izidu spremenljivk, nam določa vrsto multivariabilne analize, ki jo bomo opravljali. Pomembno je, da razmislimo kako bomo vpisali oz vnesli naše neodvisne spremenljivke v multivariabilen model. Z dihotomnimi spremenljivkami pa lahko začnemo vse multivarabilne analize, brez posebnega preoblikovanja oz transformacije.

4.2 Kako vlkjučiti nominalne neodvisne spremenljivke v multivarabilne analize?

Nominalne neodvisne spremenljivke, kot so rasa ali vrsta raka, ni mogoče vnesti v multivariabilnie analize, če so bile spremenjene. Razlog je ta, da številčna kodiranja za spremenljivke nimajo pomena. Zato je vsaka multivariabilna ocena spremembe iz ene kategorije v drugo nesmiselna,.če želimo vključiti nominalno neodvisno spremenljivko v multivariabilni model jih moramo preoblikovati oz sprementi v več dihotomnih spremenljivk. Ta proces se običajno imenuje "dummying".(Vendar pa izraz "dummying" in "neprave spremenljivke", so sleng.) Raje se slkicujemo na ta proces kot na ustvarjanje več kategoričnih spremenljivk. Najpogostejša nominalna spremenljivka v kliničnih raziskavah je etičnost. Ko se etničnost uporablja kot neodvisne spremenljivke v multivariabilnih analizah, naj bi se predstavljala kot nekaj, kar se je zgodilo. Ko prikažemo nominalno spremenljivko, kot več dihotomnih spremenljivk v multivariabilni anlizi, rabimo eno spremenljivko manj kot je število kategorij. Zakaj? Če ustvarimo pet dihotomnih spremenljivke, ki so vse bodisi 1 (da) ali 0 (ne), bo računalnik to zaznal kot šest vzorcev.

Mi ne ustvarjajo spremenljivk bel / kavkaški ,ker jih zastopajo oz predstavlja pet drugih spremenljivk. V multivariabilni analizi, se to imenuje referenčna skupina. Etičnost je zanimiv primer nominalne spremenljivke, saj, kako se boste odločili za oznako bo odvisno od vaše študijske populacije. Na primer, v majhni klinični študiji - je na jugovzhodu Združenih držav Amerike, lahko zelo malo Indijancev ali Azijcev / Pacifiških otočan Če skupina predstavlja manj kot 5 odstotkov celotnega vzorca, ki ustvarja spremenljivko, za te skupine ne sme imeti večje statistično pomembne informacije. V tem primeru si lahko ustvarimo samo spremenljivko večjih etničnih skupin, nato pa je skupina, ki je "drugo"..Najboljši način, za združitev nominalne neodvisne spremenljivke, kot je narodnost, je odvisna od raziskovalnega vprašanje oz vprašanj, distribucije nominalne spremenljivke (koliko ljudi je v vsaki skupini), ter razmerja med različnimi kategorijami nominalne spremenljivke in izida, rezultata (outcome)

4.3 Kako vključiti intervalno neodvisno spremenljivko v multivariabilen model?

Vsaka enota spremenljivke intervala je enaka. Zato bo multivariabilen model, predvideval, da bo sprememba enote kjerkoli na lestvici spremenljivke intervala imela enak učinek na “ modeled outcome “.

Predpostavka, da bo sprememba intervale neodvisne spremenljivke imela enak učinek na “outcome” (izid),se imenuje predpostavka linearnosti, in jo najlažje ocenimo v primeru linearne regresije. Če predpostavka linearnosti drži , potem bi moral diagram dveh spremenljivk kazati linijo.

Možne so različna nelinearna razmerja .To so antilogaritmiranje, “ curvilinea “(U-oblika, na glavo obrnjena črka U, J-oblika) in mejna vrednost.

Da bi ocenili, ali se spremenljivka interval prilega linearnemu prevzemu logistične regresije , ki je sorazmerena analizi nevarnosti, in Poissonovi regresiji (tesno povezana tehnika negativne binomske regresije), ne moremo oceniti linearne domneve, samo preprosto z diagramom.To je zato, ker linearno razmerje ne obstaja na navadni aritmetični lestvici. Namesto tega kategorizira spremenljivko intervala v več dihotomnih spremenljivk enakih enot na lestvici spremenljivke. Na primer, če je spremenljivka testiranja starost in je starost oseb od 20 do 79 let , potem je starost 20-29 referenčna skupina. Vsaka sprmenljivka ima koeficient za referenčno skupino, ki po definiciji znaša 0. Graf nam bo prikazal razmerje med neodvisno in končno, “outcome” spremenljivko. Če je vpliv (učinek) linearen potem bodo koeficienti postopno naraščali (ali zmanjšali), ko gredo iz ene starostne skupine v drugo in dobimo linearno črto . Možno je tudi lahko še, da je graf videti kot eden od nelinearnih odnosov .

Obstaja pa še ena bivariatna metoda za ocenjevanje, ali interval neodvisne spremenljivke ima linearno povezanost z “outcome” (rezultatom, izidom), ki se lahko izvede pred logistično regresijo.

To zahteva združevanje intervalov neodvisne spremenljivke podatkov v kategorije, ki ohranjajo intervalno vrsto spremenljivke in so dovolj velika, da zagotovijo zadostno število izidov v vsaki kategoriji. Nato lahko prikažemo enostavno tabelo o neodvisnih spremenljivkah in “outcome”. Primerjalna tabela nam mora prikazati enakomerno naraščanje (ali zmanjšanje) deleža “outcome”, ko se povečajo (ali zmanjša), skupaj s stopnjami intervala neodvisne spremenljivk.

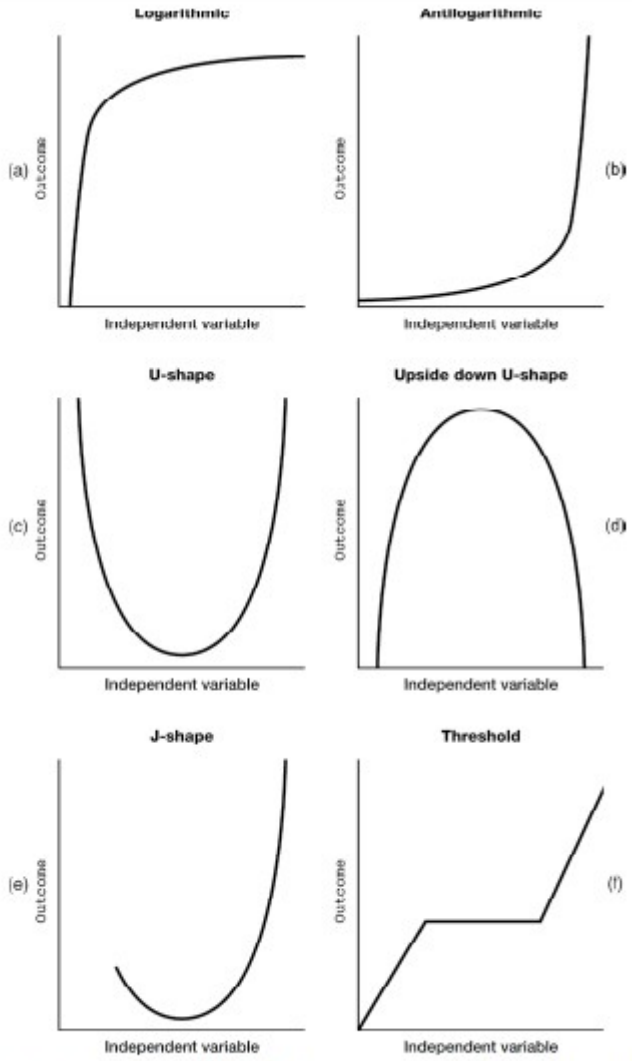


Figure 4.1 Variety of nonlinear relationships between an independent variable and an outcome.

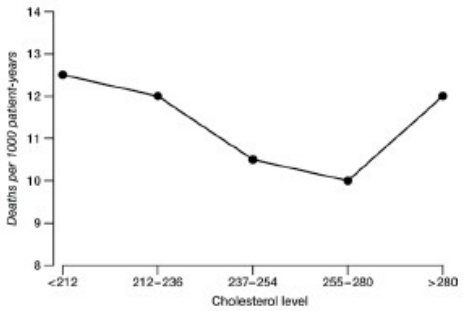


Figure 4.3 Relationship between cholesterol level and all-cause mortality among 1102 women. Adapted with permission from: Isles, C. G., *et al.* "Plasma cholesterol, coronary heart disease, and cancer in the Renfrew and Paisley survey." *Br. Med. J.* **298** (1989): 920–4. Copyright BMJ Publishing Group.

Table 4.3 Rate ratio for any fracture according to the base-line serum retinol level.

Retinol quintile	Multivariate RR* (95% CI)
1 (<1.95 µmol/liter)	0.93 (0.62–1.41)
2 (1.95–2.16 µmol/liter)	0.78 (0.50–1.23)
3 (2.17–2.36 µmol/liter)	1.00 (reference)
4 (2.37–2.64 µmol/liter)	0.91 (0.60–1.38)
5 (>2.64 µmol/liter)	1.64 (1.12–2.41)

* The analysis was adjusted for age, weight, height, and serum beta carotene, calcium, and albumin values (all continuous variables); smoking status (never smoked, former smoker, or current smoker); marital status (married or living with a partner vs. single); socioeconomic class (low, middle, or high); and physical activity at work, leisure physical activity, and alcohol consumption (all in three categories).

Data from: Michaelsson, K., *et al.* "Serum retinol levels and the risk of fracture." *N. Engl. J. Med.* **348** (2003): 287–94.

4.3 A) Matematična transformacija

Najpreprostejša metoda, ki vključuje intervalno neodvisno spremenljivko, ki nima linearne povezave z rezultatom, je da jo transformiraš tako, da izpolnjuje linearno predpostavko/domnevo. Na primer, če imajo spremembe v vrednosti na višjem koncu(točki) tvoje neodvisne spremenljivke manjši vpliv na rezultat spremenljivke, potem se spremembe v nižjih koncih (kar se kaže v postopnem padanju nagiba/naklona), z visoko točko neodvisne spremenljivke asimptotično približuje vodoravni ravni (kot na sliki 4.1(a)). Logaritemska transformacija neodvisne spremenljivke (logaritem spremenljivke) lahko linealizira trend (smernico). Naravni logaritem se uporablja pogosteje kot logaritem ki temelji na 10, čeprav lahko oba linealizirata učinek. Ne pozabimo, da mora biti pri obeh logaritmičnih transformacijah vrednost spremenljivke pozitivna(ne moreš vzeti logaritem od 0 ali negativnih števil). Če ima tvoja lestvica točno ničlo lahko še vedno uporabiš logaritmično spremembo z dodajanjem 1 vsem vrednostim.

Če imajo spremembe na spremenljivki na višjem koncu neodvisne spremenljivke večji vpliv na rezultat spremenljivke kot spremembe na nižjem koncu (kot je prikazano enakomerno naraščanje na sliki4.1.(b)) antilogaritmična sprememba (e □ or 10 □)neodvisne spremenljivke lahko linearizira trend/smernico. Logaritmične ali antilogaritmične spremembe se lahko naredi na odvisni ali neodvisni spremenljivki.

Občasno boste lahko ugotovili da je v U-obliki razmerje med vašim intervalom neodvisne spremenljivke in vašim rezultatom. Za primer slika 4.3 prikazuje U-oblikovano povezavo med nivojem holesterola in smrtnostjo na vzorcu 1102 žensk. Smrtnost je višja za ženske z nižjo in višjo vrednostjo holesterola. Med tem ko so raziskovalci obravnavali holesterol kot intervalsko spremenljivko tu ni bilo pomembnih vezi med nivojem holesterola in smrtnosti, ker dva trenda statistično razvrednotita drug drugega. Obravnavanje holesterola kot intervalne spremenljivke zgreši informacijo ključnega pomena vsebovano iz premosorazmernega razmerja. Višji nivoji holesterola je povezan z naraščajočo smrtnostjo zaradi srčno žilnih bolezni, medtem ko so nizke vrednosti holesterola povezane s povečano smrtnostjo zaradi raka in drugih primerov.

Ko opaziš U-obliko povezave razmisli o oblikovanju kvadratne oblike spremenljivke. Da ustvariš kvadratno obliko spremenljivke najprej odštejemo srednjo vrednost (povprečje) netransformirane spremenljivke (X) in nato kvadriramo rezultat (vrednost X - srednja vrednost X za vzorec). Nato vneseš obe kvadrirane oblike spremenljivke in netransformirano spremenljivko v model. Netransformirana spremenljivka mora biti v modelu zato, ker je kvadratni izraz primerljiv ekstremu srednje vrednosti netransformirane spremenljivke. Večje razlike srednje vrednosti v

katerokoli smer vodijo v statistično pomembnost. Če ima povezava U-obliko bosta oba termina (pojma,izraza) statistično pomembna v tvojem modelu. Kvadratna oblika spremenljivke bo tudi delovala pri J-oblikovani povezavi. Kot lahko opaziš na sliki, je J-oblikovana krivulja kot U-oblikovana krivulja z nekaj manjkajočimi podatki točk.

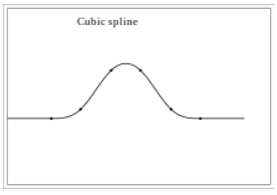
Omejitev uporabe matematičnih transformacij je to, da jih je težko interpretirati/razložiti. Na primer: kaj bi pomenilo kliničnemu zdravniku, če bi vedel, da je logaritem vrednosti holesterola povezan z naraščajočo smrtnostjo? Kvadratni termini so lahko predvsem težki za interpretiranje, ker združene spremembe v intervalskem neodvisnem faktorju vplivajo na rezultat preko dveh spremenljivk- vsaka z različno povezavo z rezultatom. Vendar pa je največja slabost matematičnih transformacij ta, da za veliko kombinacij ni preprostih matematičnih transformacij, ki bi izpolnjevale linearno prepostavko/domnevo.

4.3 C Multiple dihotomne spremenljivke

Tretja metoda za vključitev nelinearnih povezav med intervalom dejavnika tveganja in rezultatom v multivariabilnem/multivariantnem modelu je, da ustvariš dihotomno spremenljivko iz intervalne spremenljivke. To je enak proces, ki bi ga uporabil za vključitev nominalne spremenljivke v analizo (poglavje 4.2) ali za testiranje ali ima intervalna-neodvisna spremenljivka linearno povezavo z rezultatom (poglavje 9.4). Multipla dihotomna spremenljivka omogoči vsaki kategoriji, da je lahko lastna neodvisna spremenljivka in ima lastno povezavo z rezultatom.

Na primer: v študiji o *vrednosti retinolova v serumu v tveganju za zlom* je opisano oboje, avtorji so prikazali svoje rezultate z uporabo večih dihotomnih spremenljivk v povezavi z analizami z omejenenimi kubičnimi povezavami funkcij (cubic spline function). Rezultati so prikazani v tabeli 4.3.

Tveganje za zlomov je 1,64-krat večje med osebami v višjih kvantilnih vrednostih v primerjavi s temi v srednjih kvantilih. Opaziš lahko, da se predstavljeno relativno tveganje v povezavi z različnimi kvantili vrednosti retinola v serumu povejo isto zgodbo kot kubične povezave funkcij (cubic spline function), vendar veliko manj elegantno.



Eden glavnih izzivov uporabe multiplih dihotomnih spremenljivk je določitev zgornjih mej. Na splošno velja, da je najbolje da uporabite zgornje meje, ki odražajo naraven, klinično ustrezen standard. Na primer: smiseln zgornji prag za sistolični krvni pritisk bi bil < 90 mm, 90–140 mm, >140 mm.

Slabost izbire naravne zgornje meje je ta, da zgornja meja lahko razdeli vzorce v skupine z neenakimi velikosti vzorcev. Na primer: če razdelite vaš vzorec v desetletja starosti, imate lahko samo 2% svojih vzorcev z odgovorom »Da« na spremenljivki 80-89 let. Če je številka oseb, ki so v tej spremenljivki premajhna, spremenljivka ni smiselna v analizi. Za primerjavo, če izberete zgornjo mejo to pripelje do enakih velikosti, potem bo distribucija multiplih dihotomnih spremenljivk enaka. Za primer: predpostavimo, da deljenje vzorca na tercile starosti ustvari najmlajše predstavnike za referenčno skupino. Ena dihotomna spremenljivka bo imela vrednost »DA« za tretjino vzorca (odrasla doba) in vrednost »NE« za dve tretjini vzorca (najmlajši in starejše osebe), druga dihotomna spremenljivka bo imela prav tako vrednost »DA« za tretjino vzorca (starejši ljudje) in vrednost »NE« za dve tretjine vzorca (najmlajši in odrasli ljudje).

Ena metoda za zagotavljanje enako velikih skupin je izbiranje zgornje meje temeleče na enako distributiranih subjektih , kot so tercili, kvartili, … vrednosti na neodvisni spremenljivki, kot je bilo narejeno v študiji serotonina. Medtem ko se zagotavlja skupine enakih velikosti, izpuščanje naravnih enot neodvisne spremenljivke lahko privede do tega, da rezultati zvenijo manj prepričljivo. Katera od naslednjih zveni bolj prepričljivo? Osebe v višjih tercilih starosti imajo trikrat večjo možnost da umrejo, kot osebe v nižjih tercilih let ali osebe stare 70-89 let imajo trikrat večjo možnost da umrejo, kot osebe stare 30-49 let. Druga opcija se zdi prepričljivejša.

Kljub metodi izbire zgornjih mej je pomankljivost uporabe multiple dihotomne spremenljivke ki prikaže intervalno spremenljivko ta, da naraste število spremenljivk v tvojem modelu. To je lahko problem kadar nimamo dovolj velikih vzorcev.

4.4 Ob predpostavki da moja intervalna neodvisna spremenljivka ustreza linearni predpostavki, je tukaj kakšen razlog za grupiranje v intervalno kategorijo ali kreiranje multiplih dihotomnih spremenljivk?

Tudi ko interval neodvisne spremenljivke, kot so leta, krvni pritisk ali holesterol, ustreza linerani predpostavki, pogosto ni levo v njegovi originalni intervalni legi. Za to imamo različne razloge.

Kadar imaš majhen vzorec velikosti (100 oseb) je za eno stvar lahko težko oceniti, če intervalna spremenljivka (starost) dopolnjuje linearno predpostavko, dokler jo grupiraš v kategorije. Levo negrupiran model bo predpostavil, da je razlika v verjetnosti rezultatov med subjekti starosti 55 in subjekti starosti 57 enaka tisti med subjekti starosti 61 in subjekti starosti 63. Tukaj verjetno ne boš imel nikogar ali le eno ali dve osebi v tvojem celem vzorcu s temi leti. Tudi tvoje poslušatelje verjetneje bolj zanima učinek rezultatov v desetih letih, kot učinek rezultatov dobljenih v enem letu (kateri je najverjetneje manj zanesljiv za raziskovanje večine bolezni).

Ko grupiramo intervalno spremenljivko, obdržimo naravni interval lestvice (skupina starosti po desetletjih). To nam omogoča, da ohranimo prednosti intervalne lestvice (ker je razlika med starostmi 20-29, 30-39, in 40-49, enaka 10 let). Sedaj boste lahko bolje ocenili/ovrednotili ali spremenljivka ustreza predpostavki statističnega modela in boste lahko dobil in poročal več smiselnih/pomebnih rezultatov.

Včasih raziskovalci ustvarijo multiple dihotomne spremenljivke, čeprav mislijo da spremenljivka približa linearno povezavo. Razlog je, da je ustvarjanje multiplih dihotomnih spremenljivk bolj konzervativne strategije. Odkar linearna povezava ni predpostavljena, bralcu (ali tvojemu recenzentu/kritiku) ne rabiš dokazati, da je linearna predpostavka neizpolnjena/neizvršena. Kakorkoli, rezultati te strategije so v porastu v številu spremenljivk v tvojem modelu; to bi lahko bil problem če je tvoj vzorec premajhen (poglavje 6.5). Tudi statistično pomembno linearni trend med intervalnim faktorjem tveganja in rezultatom lahko ni dolgotrajno statistično pomemben, ko je intervalna spremenljivka pedstavljena kot multipla dihotomna spremenljivka, ker za nobeno od njih ni pomembna razlika referenčnih skupin.

4.5 Kako vključim ordinalne neodvisne spremenljivke v multivariantni model?

Ordinalne spremenljivke predstavljajo izziv podoben nominalnim spremenljivkam, ko se uporabi neodvisne spremenljivke v multivariantnih modelih. Ker tukaj ni enakih razdalj med vsako stopnjo ordinalne spremenljivke, imamo samo eno oceno učinka prehoda iz ene stopnje spremenljivke v drugo na odvisni spremenljivki. Po drugi strani pa lahko obstajajo spremenljivke, ki so lahko tehnično ordinalne v naravi, ampak jih je mogoče obravnavati kot da so intervalne, saj operirajo/delujejo kot ordinalne. Z drugimi besedami, ena ocena enake spremembe je kjerkoli na lestvici veljavna. To je pogost primer za intervalna merjenja, ki temeljijo na

psiholoških in sociooloških lestvicah, ki so pridobljene iz večih/multiplih vprašanj. (več v poglavju 6.5).

5.1 Ali je pomembno, če sta dve neodvisni spremenljivki povezani med sabo?

Če dve spremenljivki med sabo visoko korelirata na takšen način, če poznaš vrednost ene, poznaš tudi vrednost druge, potem multivariabilna analiza ne more ločeno oceniti vpliv obeh spremenljivk na rezultat. Ta problem imenujemo multikolinearnost. Primer tega je, da preučujemo dejavnike, ki so vplivale na dolžino bivanja v bolnišnici pri bolnikih s pljučnico. Za dejavnike si izberemo temperaturo in jo merimo v stopinjah Celzija in Fahrheita.tako dobimo eno spremenljivko, vendar dva različna rezultata. Če te rezultate vnesemo v Model ocenjevanja dolžine bivanja, bi bil model napačen in bi vseboval napako. Obstaja enostavna matematična pretvorba iz ene v drugo.

5.2 Ocenjevanje multikolinearnosti

Zelo težko dobimo dve spremenljivki, ki med sabo tako močno korelirajo, razen če se zmotimo in vključimo dve spremenljivki, kot so °C in °F. večja možnost je, da dobimo spremenljivke, ki niso dovolj drugačne, da bi jih razlikovali. Primer: Phibbs in socelavci so ugotovili, da sta porodna teža in gestacijska starost preveč tesno povezana, zato ju niso vključili oba v svojo študijo o neonatalni umrljivosti.
Če bosta v študijo vključene obe Neodvisne spremenljivke, ki med sabo visoko korelirata, bo v študija imela velike standardne napake. Vpliv posameznih spremenljivk na rezultate pa ne bo mogoče oceniti, vendar bo splošni model še vedno točen.

5.2 A Kako ocenim ali so moje spremenljivke multikolienarne?

Korelacijski koeficient imenovan tudi Pearsonov koeficient meri kako močno sta dve spremenljivki med seboj povezani. Koeficient korelacije prevzame, da je razmerje med spremenljivkama linearno. Vrednost Pearsonovega koeficienta korelacije se lahko nahaja med vrednostima -1 in 1. Tako vrednost -1 predstavlja popolno negativno povezanost spremenljivk, pri čemer je na grafu odvisnosti videti le ravno črto, ki z naraščajočo neodvisno spremenljivko potuje navzdol; obratno vrednost 1 pomeni popolno pozitivno povezanost in navzgor usmerjeno črto na grafu. Če dobljeni rezultat pomnožimo s 100, dobimo odstotek povezanosti.

5.3 Izogibanje multikolinearnosti

Če so nekatere spremenljivke zaskrbljujoče lahko naredimo dodatne analize in ugotoviti katere ostale spremenljivke so tesno povezane s problematično spremenljivko. To naredimo tako, da izvedemo regresijsko analizo, pri kateri uporabljamo ostale spremenljivke kot neodvisne spremenljivke in ocenimo problematičnost problematične spremenljivke.

5.3 A Kaj narediti z multikolinearno spremenljivko?to nam pokaže katere spremenljivke so tesno povezane in katere ne.

Če imamo spremenljivke, ki so zelo povezane lahko:

- Izpustimo spremenljivko

Izpustimo tisto spremenljivko, ki je teoretično manj pomembna, ima več pogrešanih podatkov, ima večjo napako merjenja ali je na kakšen drugačen način manj zadovoljiva.

- Uporabimo veznike in/ali
- Ustvarimo lestvico

Multivariabilna analiza

6.1 Katero neodvisno spremenljivko moram vključiti v svojo multivariabilno analizo?

To se na začetku zdi kot preprosto vprašanje. Moramo vključiti faktorje tveganja ter vse spremenljivke, ki lahko vplivajo na faktorje tveganja in na rezultate. Nekatere spremenljivke lahko vplivajo negativno na našo analizo, zato je te treba izključiti npr.:to so spremenljivke ki imajo veliko manjkajočih podatkov, so odveč,

6.2 Kako se odločimo katero skrito spremenljivko vključimo v svojo analizo?

Skrite spremenljivke, so spremenljivke, ko korelirajo z odvisno in neodvisno spremenljivko.
idealni raziskovalci bi v svojo raziskavao vključili vse spremenljivke, ki so označene kot skrite. Vendar lahko se pokaže, da je veliko skritih spremenljivk. Zato v svojo raziskavo vključimo samo tiste, ki ustrezajo empirični definiciji skritih spremenljivk. Vendar pa ne obstaja standard, ki bi nam povedal koliko mora skrita spremenljivka korelirati z drugima

dvema spremenljivkama, da jo vključimo v raziskavo kot skrito spremenljivko.

6.3 Katere neodvisne spremenljivke izključim iz svoje multivariabilne analize?

Izključiti moramo tuje spremenljivke, ki niso vzročno povezane z našim rezultatom, spremenljivke z manjkajočimi podatki, odvečne spremenljivke, multikolienarne spremenljivke, vmesne spremenljivke

6.4 How many subjects do I need to do multivariable analysis?

- manjši kot je učinek, večji vzorec potrebujemo, da bi pokazali, da obstaja statistično pomemben učinek

- več kot je variabilnosti v meritvah, večji vzorec potrebujemo, da dokažemo, da obstaja statistično pomemben učinek

Power calculation: (računanje moči); nam pove kako velik vzorec potrebujemo da zaznamo določen učinek.

-če ti izračun moči pokaže, da nimaš dovolj subjektov za dokaz učinka pri bivariatni analizi, jih tudi za dokaz efekta pri multivariabni analizi nimaš dovolj.

- najlažji način za izvedbo MA(multivariable analysis) je s pomočjo računalniških programov.

Programi ti pomagajo izracunati velikost vzroca za multiplo linearno regresijo in multiplo logistično regresijo (prog. Power and Precision). PASS (statistical and power analysis softwear program) izračuna velikost vzorca za multiplo linearno regresijo, multiplo logistično regresijo, analizo razmerja nevarnosti in Poissonovo regresijo.Za vse izračune na velikosti vzorca je potrebno napisati vrednost alfe(ponavadi 0.05) in bete(ponavadi 0.80).

Odvisno od tega kateri program izbereš za izvajanje analize, je potrebno vnesti različne podatke, ki so potrebni za nadaljevanje. Pri PASS moraš vnesti; naklon, ki predvideva nično hipotezo, naklon ki predvideva alternativno hipotezo, stand. Deviacijo neodvisnih spremenljivk in st.dev. odvisnih spremenljivk, relativno razmerje spremenljivk (npr št subjektov v vsaki skupini, npr; 1:2), mediano pogostosti učinka (isto za vse spremenljivke) in pogostost pojava na neki drugi točki, razen mediani neodvisne spremenljivke. Če vključiš dve neodvisni spremenljivki, moraš

vnesti tudi korelacijo med njima. Kaj je potrebno vnesti je odvisno od tega za katero analizo računaš velikost vzorca.

Če pogledaš čez palec in te zanima približna velikost vzorca za multiplo logistično regresijo in razmerje nevarnosti, potrebuješ za vsako neodvisno spremenljivko v svojem modelu, vsaj 10 različnih rezultatov. Za multiplo linearno regresijo pa 20 subjektov. Če je število manjše, to še ne pomeni, da je študija neveljavna, je pa potrebno bolj previdno interpretirati vrednosti koeficientov.

6.5 Kaj če imam preveč neodvisnih spremenljivk, glede na velikost vzorca?

V takem primeru moraš ali povečati velikost vzorca ali pa zmanjšati število neodvisnih spremenljivk.

1) Izključi spremenljivke, ki se ločijo od drugih. (imajo minimalni učinek na glavni efekt pri multivariabilni analizi, lahko uporabiš kakšen algoritem za selekcijo spremen.)
2) Izberi eno spremenljivko, ki predstavlja hkrati še dve podobni.
3) Združi spremenljivke v eno spremenljivko, oceno ali lestvico. (sprem. Lahko združiš s pomočjo izrazov in/ali) (alfa večja od 0.65 kaže na to, da je lestvica zanesljiva)
4) Analiza faktorjev združi več spremenljivk v nekaj faktorjev, ki so skupni vsem. Zna se pojaviti problem, saj s tem se izgubijo primarne spremenljivke.

6.6 Kaj storiti z manjkajočimi podatki pri neodvisnih spremenljivkah?

Manjkajoči podatki predstavljajo problem pri vseh tipih analize. Problemi, ki jih manjkajoči podatki predstavljajo v bivarjantni analizi, se urejujejo v murtivariabilni analizi, saj obstaja večja možnost, da bodo različni subjekti vsebovali manjkajoče podatke pri različnih spremenljivkah. Če pokažemo na primeru: študija vsebuje 300 udeležencev, z desetimi različnimi spremenljivkami. Vsaka spremenljivka vsebuje 10 manjkajočih podatkov. V bivariantni analizi bi bil vzorec velik n=290 (97%), v multivariabilni analizi pa bo verjetno več kot 10 manjkajočih podatkov, saj bodo vsi udeleženci, ki imajo vsaj en manjkajoči podatek, izbrisani. V takšnem primeru bi pri tej študiji izgubili 10 podatkov na spremenljivko, torej 100 podatkov. To je zelo velika količina manjkajočih podatkov, zato bo težje pridobiti veljavne ter

generalizirane rezultate. Ponavadi imajo osebki z manjkajočimi podatki na eni spremenljivki, tudi manjkajoče podatke na drugih spremenljivkah. S tem v analizi izgubimo ogromno podatkov.

Pri odločanju o tem kako bomo analizirali podatke, je dobro, da vemo koliko podatkov bo manjkajočih v multivariabilni analizi. Da lahko to določimo, ustvarimo spremenljivko, ki ima vrednost 1 če podatki na neodvisni spremenljivki manjkajo in vrednost 0, če podatki obstajajo. Frekvenca, ki jo iz tega dobimo nam pove, koliko podatkov bo manjkajočih v multivariabilni analizi, ki vključujejo vse spremenljivke.

Ko ugotovimo koliko manjkajočih podatkov je prisotnih, imamo na voljo pet metod za upravljanje s temi podakti.

1. Izbris primerov z manjkajočimi podatki:
   Je najbolj enostavna in pogosta metoda, ob kateri pa naletimo na problematičen pojav odstopanj zaradi izbrisa podatkov (če so manjkajoči podatki nastali naključno verjetno do odstopanj ne pride). Če se odločimo za izbris podatkov, lahko to storimo že na začetku analize, lahko pa počakamo do začetka multivariabilne analize. Če izbrišemo kasneje, moramo biti previdni pri predstavitvi rezultatov in bralcu povedati, kateri podatki so bili uporabljeni v določeni analizi (posebej v univariantni, bivariantni, in v multivariabilni).
2. Oblikovanje dihotomnih spremenljivk za predstavitev manjkajočih podatkov:
   Prednost te metode je, da so v analizo vključeni vsi subjekti, da ni vključenih močnih predpostavk o manjkajočih podatkih ter da lahko razumemo odstopanja, ki jih povzročijo ti manjkajoči podatki.
3. Dodatno prizadevanje za pridobitev pomanjkljivih podatkov:
   Problem te metode je, da te podatke težko pridobimo, če smo že v fazi analize.
4. Zmanjševanje števila neodvisnih spremenljivk v analizi:
   Ta metoda deluje samo ko imamo spremenljivke, ki jih lahko izbrišemo brez ogrožanja analize. Poznamo tudi več strategij zmanjševanja, vsekakor pa najprej izbrišemo spremenljivke, ki imajo veliko manjkajočih podatkov.
5. Ocenjevanje vrednosti manjkajočih podatkov:
   To je najbolj zadovoljiva a tudi nevarna metoda. Zadovoljiva je, ker ne izgubimo nobenih primerov in nevarna, ker lahko pride do odstopanj v rezultatih, ki jih je težko predviditi.

- Določanje povprečne vrednosti vzorca: s tem pokažemo na verjetnost, da so se manjkajoči podatki pojavili naključno in s tem povprečje poda najboljšo oceno. Prednost tega postopka je, da so v analizo vključeni vsi podatki, a je primeren le za subjekte, ki imajo samo dve neodvisni spremenljivki z manjkajočimi podatki.
- Določanje pogojne povprečne vrednosti podskupine: včasih raje ocenimo pogojno povprečno vrednost podskupin kot povprečje za celoten vzorec (pogojno glede na vrednosti ostalih spremenljivk).
- Modeliranje vrednosti manjkajočih podatkov z uporabo drugih neodvisnih spremenljivk v analizi: dovoljuje bolj precizno oceno manjkajočih podatkov kot pri ocenjevanju povprečja.
- Modeliranje vrednosti manjkajočih podatkov z uporabo drugih neodvisnih spremenljivk in vključitev naključnih komponent: ker lahko pri zgornji metodi precenimo napake, ki jih storimo z uporabo drugih spremenljivk, lahko za rešitev tega problema vključimo še naključne komponente.

Ker ima vsaka metoda svoje prednosti in pomanjkljivosti lahko velikokrat uporabimo več metod hkrati. Prednost tega je, da lahko pogledamo če naša izbira metode prikaže razliko v rezultatih.

<u>Osnovni vodnik pri problemih z manjkajočimi podatki je torej naslednji:</u>

1. Zbiranje podatkov, da minimiziramo manjkajoče podatke
2. Ocenite koliko manjkajočih podatkov vsebujejo posamezne neodvisne spremenljivke
3. Če imate eno ali dve nedovisni spremenljivki, ki imajo občutno več manjkajočih podatkov. Ne glede na pomembnost spremenljivke v vaši teoriji, če vsebuje veliko manjkajočih podatkov, bodo vaše informacije verjetno pomenile odstopanja.
4. Preverite vse ostale uporabljene spremenljivke in preglejte koliko manjkajočih vrednosti vsebujejo, preden uporabite multivariabilno analizo. Če pa imate malo primerov z manjkajočimi podatki jih izbrišite že takoj na začetku.
5. Če imate veliko primerov z manjkajočimi podatki, določite če se primeri z manjkajočimi razlikujejo od primerov brez manjkajočih podatkov.
6. Če se ne razlikujejo, uporabite določevanje povprečij ali pogojnih povprečij. Preden to storite se prepričajte, da imajo primeri ustrezne vrednosti za vsaj polovico neodvisnih spremenljivk v vaši analizi. Če imate primere z manjkajočimi podatki na večini neodvisnih

spremenljivk, jih izbrišite. S pomočjo biostatistike uporabite pristop multiplega pripisovanja.
7. Če se manjkajoči podatki razlikujejo od primerov brez manjkajočih podatkov, ste v težkem položaju. Poskusite uporabiti več različnih metod za reševanje manjkajočih podatkov.

6.7 Kaj storiti z manjkajočimi podatki pri preučevani spremenljivki *(outcome variable)*

Manjkajočih podatkov v preučevani spremenljivki ne moremo obravnavati kot tistih v ostalih spremenljivkah. Namen multivariabilne analize je ocena končnega izida preučevane spremenljivke na osnovi neodvisnih spremenljivk. Za preučevanje teh manjkajočih podatkov lahko uporabimo strategijo multiplega pripisovanja (*multiple imputation*). Multivariabilni modeli ocenjujejo izid glede na odnos med odvisnimi spremenljivkami in izidom. Ko enkrat ocenimo izdi, lahko ocenimo tudi izid primerov, kjer imamo samo informacije o neodvisnih spremenljivkah. Z ocenjevanjem preučevane spremenljivke v primeru manjkajočih podatkov, lahko ponovimo multivariabilno analizo z dodatnimi primeri in nato vidimo razlike med dobljenimi rezultati. Če razlik ni, to poveča veljavnost analize.

7.1 Katera števila naj določim za dihotomno ali ordinalno spremenljivko v moji analizi?

Vzemimo preprost primer dihotomne spremenljivke, ki temelji na vprašanju: Ali ste kdaj imeli sladkorno bolezen: da ali ne?

Običajno so enačbe rešile multivariabilne analize za katere smo potrebovali numerično predstavo da-jev in ne-jev. Od kar ima ta lestvica le dva vidika, numerična razdalja med vidikoma je lahko predstavljena s katerima koli dvema številoma med 0-1, 1-2, 0- (-1), itd. Oznaka koeficienta se lahko spremeni, odvisno od tega ali se odločimo za »da«, ki je lahko višje ali nizke vrednosti, vendar koeficient in stopnja pomena ostaneta enaka (glej 8.3). Dobili boste enako odgovor ali kodirate spremenljivke tako, da je med dvema številkama več kot ena točka. Primer: kodirajoče sheme kot so +1 in -1 nam dajo različne odgovore, ker je med njima več kot le ena točka.
Katerikoli dve števili, ki sta le eno število narazen, podata enak odgovor, smiseln odgovor, za neodvisno in odvisno spremenljivko uporabite 1 in 0. 1 predstavlja prisotnost in 0 predstavlja odsotnost stanja. To konvencijo si z lahkoto zapomnimo, ter poveča verjetnost, da nas zmede v smeri efekta. Ta kodirajoča shema ima še eno prednost: ko je spremenljivka kodirana v

tej smeri, pomeni, da spremenljivka predstavlja razširjenost stanja. Na primer, če imamo 100 subjektov in 10 jih je izkusilo rezultat, srednja vrednost spremenljivke (če kodiramo 0,1) bo ([0x90]+[1x10])/100=0,10. To je lahko priročno, ko želimo izvedeti prevalenco faktorja tveganja ali izid v točno določeni skupini pacientov.
Za nekatere spremenljivke, kot so spol, kjer ni prisotnosti ali odsotnosti stanja. Določimo vrednost 1, ki bo naredila rezultate najbolj smiselne. Na primer, če je naša hipoteza, da ženske z boleznijo koronarne arterije dobijo vročino. Postopek zaradi pristranskosti spola, bi bilo smiselno dodeliti ženskam 1 in moškim 0.

Lahko imate občutek, da sem že izpostavil vidik kodiranja, ampak, če niste previdni se z lahkoto zmedete o svojih rezultatih. Kot lahko vidite v tabeli 7.1 z samo eno neodvisno dihotomno spremenljivko in en dihotomen rezultat, saj so možna 4 različna kodiranja. Drugačno kodiranje nam bo dalo statistično enak rezultat, vendar interpretacija rezultata bo drugačna. Na primer, če se zmedemo pri kodiranju spremenljivk, lahko interpretiramo tako, da poročamo o naraščajočemu faktorju tveganja za rezultat, v bistvu pa faktor pada. Ta isti problem se lahko pojavi pri bivariantni analizi. Z več neodvisnimi spremenljivkami se lažje zmedemo.

Raziskave le redko poročajo o tem kako kodirajo njihove spremenljivke v rokopisu. Obstaja majhna verjetnost, da bodo zlorabljene spremenljivke odkrite med strokovnim pregledom. Od nas je odvisno ali bomo pravilno poročali o dobljenih rezultatih.

Poleg kodiranja spremenljivk in natančnega poročanja o našem delu, ne obstajajo nobene druge strategije za minimiziranje možnosti poročanja rezultatov v nasprotju s tem kakšni so bili dejanski rezultati. Najprej je potrebno poimenovati spremenljivke, tako specifično kot lahko v okviru statističnega paketa (omogočajo do 8 znakov). Na primer, bolje je poimenovati spremenljivko »femgend« kot »gender«.

Naslednja uporabna strategija je da uporabljamo etikete. Računalnik bo natisnil vrednosti, ki so dodeljene vsakokrat, ko uporabimo spremenljivko v analizi. Obstaja manjša možnost, da storimo napako, če vidimo na izpisu »1- zaposlen, 0-brezposelen« poleg spremenljivke. Vnos vrednosti etiket lahko vzame nekaj dodatnega časa preden začnemo z analizo, vendar se izplača.
Z ordinalnimi spremenljivkami, numerična reprezentacija različnih stopenj ne spremeni ničesar vse dokler se ne spremeni numerična vrednost med stopnjami. Na primer, ni pomembno ali je 4-stopenjska ordinalna spremenljivka (zelo zadovoljen, zadovoljen, nezadovoljen, zelo nezadovoljen) kodirana kot 0,1,2,3 ali 1,2,3,4 ali 9,10,11,12. Z

dihotomnimi spremenljivkami bo smer kodiranja vplivala na znak regresijskega koeficienta. Torej je pomembno za interpretacijo rezultatov ali je spremenljivka kodirana kot zelo zadovoljen=1, zadovoljen=2, nezadovoljen=3 in zelo nezadovoljen=4 ali zelo nezadovoljen=1, nezadovoljen=2, zadovoljen=3, zelo zadovoljen=4. Z dihotomno spremenljivko je ključ, da obdržimo vzorec po katerem kodiramo spremenljivke in ustrezno interpretiramo koeficient.

Tabela 7.1

Možno kodiranje 1	Možno kodiranje 2
Nezaposlen=0;zaposlen=1;brez zdravljenja=0;zdravljenje=1	Nezaposlen=1;zaposlen=0;brez zdravljenja=0;zdravljenje=1
Možno kodiranje 3	Možno kodiranje 4
Nezaposlen=0;zaposlen=1;brez zdravljenja=1;zdravljenje=0	Nezaposlen=1;zaposlen=0;brez zdravljenja=1;zdravljenje=0

7.2 Ali je pomembno, če izberem za mojo referenco kategorijo za multiple dihotomne spremenljivke?

V sekciji 4.2, je bilo razloženo, da spremenljivka, ki predstavlja referenčno skupino, ne bo bila vključena v analizo. Primerjana bo s to skupino. Če to vemo, je pomembno katero skupino bomo vzeli za referenčno skupino? Odgovor je, da tvoja izbira referenc naredi razliko v tem, kako sporočiš tvoje rezultate in manjšo razliko v rezultatih samih.

Tabela 7.2 prikaže implikacije variiranja referenčne skupine. Predvidevamo, da so podatki iz študije, ki je temeljila na asociacijah med etničnostjo in dostopom do zdravstvene oskrbe. V stolpcu 1, so referenčna skupina belci. 'odds radio' pokaže, da imajo Afro-Američani in Američani štirikrat manjši dostop do zdravniške oskrbe kot belci. Latino-Američani, Azijci in ostali nebelci pa imajo za polovico manjši dostop do zdravniške oskrbe kot nebelci.

V stolpcu 2 so našteti vsi isti podatki še enkrat, le da so tu referenčna skupina Afro-Američani. Spet vidimo, da imajo belci štirikrat večji dostop do zdravstvene oskrbe kot Afro- Američani, itd. čeprav sta si stolpca 1 in 2 matematično ekvivalentna, je analiza rezultatov različna. Če je bila hipoteza študije, da imajo nebelci manj dostopa do zdravstvene oskrbe kot belci, potem je smiselno, da so v prvem stolpcu belci referenčna skupina. To daje možnost sporočanju bralcem, kakšen je zdravstvena oskrba za nebelce v primerjavi z belci. Če pa so Afro-Američani izbrani za referenčno skupino v drugem stolpcu, primerjava med skupinami ni možna. Zaradi tega razloga, raziskovalci na splošno izbirajo referenčno skupino na

podlagi glavne hipoteze, ki se jo testira. Če ni glavne hipoteze, slepe spremenljivke predstavljajo intervalno spremenljivko (npr. starost), na splošno pa je lažje poročati rezultate, če jih primerjamo z empiričnimi dognanji. Na primer: če je starost povezana z naraščajočim (ali padajočim) odstotkom rezultatov, bi morali uporabiti ekstremno kategorijo (npr. najmlajše in najstarejše subjekte) kot referenčno skupino. To nam omogoča povzeti rezultate tako, da rečemo da so starejši ljudje bolj/manj dovzetni za nek izid rezultata kot mlajši ljudje. Nasprotno pa, če ima slepa spremenljivka U-obliko distribucije, bi bilo mogoče bolje vzeti srednjo skupino za referenčno skupino tako, da lahko demonstriramo povišano tveganje dveh ekstremov. Katerokoli skupino izbereš za referenčno skupino naredi majhno statistično razliko. Če izbereš večjo skupino za referenčno kategorijo, bodo bili standardi napak rahlo manjši in zaupni intervali bodo bili nekoliko omejeni, ker ima model veliko primerjalno skupino in lahko zato naredi bolj natančnejše ocene. Če hipoteze in empirični rezultati ne vodijo k izbiri določene kategorije za referenčno skupino, potem je potrebno izbrati tisto, ki je večja.

ble 7.2 Implications of changing the reference group fo
chotomous variables.

	Odds ratio	Odds ratio
hite/Caucasian	1.0 (reference)	4.0
rican-American	0.25	1.0 (reference)
tino	0.50	2.0
ian/Pacific Islander	0.50	2.0
tive American	0.25	1.0

8.1 Interpretacija rezultatov

Katere informacije pridobim z multivariabilno analizo?

Multivariabilna analiza proizvede 2 vrsti informacij: informacije o uspešnosti prileganja modela (vse neodvisne spremenljivke skupaj) na podatke in informacije o zvezi vsake neodvisne spremenljivke z izidom spremenljivke.

8.2 Več-linearna regresija

Ocena več-linearnega modela regresije se začne s testiranjem samostojnih spremenljivk, napovedujejo izid. Če poznamo vrednosti posameznih spremenljivk, nam to omogoča boljše predvidevanje kot naključnost. Zato bo F vrednost visoka. Visoka vrednost F za dano testno velikost in dano

število spremenljivk v modelu (ki določa stopnjo svobode), bo rezultirala v majhni vrednosti P-ja. To nam pove, da nična hipoteza ni asociirana med neodvisnimi spremenljivkami. Zato je lahko izid zavržen. Velika omejitev F testa je, da ne pove koliko in katere spremenljivke v tvojem modelu so pomembne. Naslednja pomanjkljivost je, če vemo, da spremenljivke kot skupina so bolj tesno povezane z izidom, kot je pričakovano po naključju. To nam ne pove kako dobro bodo posamezni rezultati spremenljivk šteli za izid. Če poznamo te omejitve, za kaj je potem ta test primeren? R2 test (tudi variacijski koeficient) je na splošno bolj uporaben kot F test, ker naredi kvantitivne meritve kako dobro lahko posamezne spremenljivke razložijo rezultat. Vrednost R2 gre od 0 proti 1. Ko je R2 pomnožen s 100 je lahko mišljen kot odstotek variance pri odvisnih spremenljivkah, ki so razložene s strani neodvisnih spremenljivk. Statistično prilagojen R2 izračuna ceno za vsako spremenljivko v modelu. Če dodamo spremenljivke, se prilagojen R2 poveča, zmanjša ali pa ostane enak.

Table 8.1 Methods for assessing how well a model accounts for the outcome.

	Multiple linear regression	Multiple (binary) logistic regression	Proportional odds regression	Multinomial logistic regression	Proportional hazards analysis	Poisson regression (negative binomial regression)
Independent variables are associated with outcome more than would be expected by chance	F test	Likelihood ratio test	Likelihood ratio test	Likelihood ratio test	Likelihood ratio test	Likelihood ratio test
Quantitative/ qualitative assessment of how well model accounts for outcome	R^{2*}	Pearson goodness-of-fit Deviance goodness-of-fit Comparison of estimated to observed value Hosmer–Lemeshow test* Sensitivity, specificity, accuracy (requires choosing a cut-off) c index*	c index Create separate binary models and use same statistics as with binary logistic regression.	Create separate binary models and use same statistics as with binary logistic regression.	Comparison of estimated to observed outcome.	Deviance goodness-of-fit test*

*Best tests/most commonly reported tests are shown with an asterisk.

8.3 Kaj mi povedo koeficienti o zvezi med vsako spremenljivko in izidom?

Beta koeficient pove kako se spremeni izid glede na spremembe pri neodvisnih spremenljivkah, medtem ko na model nastavljamo ostale neodvisne spremenljivke. Koeficienti so lahko pozitivni ali negativni, njihova interpretacija pa se zaradi različnih tehnik in izidov razlikuje.

8.3 A Koeficienti v multipli linearni regresiji

V tem primeru je srednja vrednost izida oblikovana. Za vsako povišanje/znižanje neodvisne spremenljivke, se srednja vrednost izida poviša/zniža za količino koeficienta. Pozitivni koeficient kaže, da se neodvisna spremenljivka in izid premikata skupaj (gor ali dol). Negativen koeficient pa kaže, da se neodvisna spremenljivka in izid premikata v nasprotne smeri.

Nagib pove zvezo med neodvisno spremenljivko in izidom in tako lahko narišemo premico, ki kaže najboljše ocenjene vrednosti za vse možne vrednosti odvisne spremenljivke.

Za vsak koeficient model multple linearne regresije izračuna tudi P vrednost, ki temelji na t testu. T je koeficient za standardno napako in ima intuitivni pomen. Če je T vrednost velika (večja od 2.0), potem je P majhen ($P<0,05$), je to statistično pomembno, saj zavrnemo ničelno hipotezo.

8.3.B Koeficient v multipli (binarni) logistični regresiji

Pomen tega koeficienta se razlikuje od prejšnjega primera, saj oblikuje logaritem za lihe izide – kar se imenuje *logit.* Koeficient pove kako sprememba ene enote v neodvisni spremenljivki spremeni logit. Pozitivni koeficient pomeni, da če se poveča spremenljivka, se poveča tudi *logit*, negativni koeficient pa ravno obratno.

Za interpretacijo pomena koeficienta moramo vedeti katere vrednosti izida *logit-a* se ocenjujejo. Programi večinoma privzeto določajo logit za nižjo numerično vrednost, vendar lahko sami določimo upoštevanje višje vrednosti. Rezultati so v obeh primerih enaki, vendar predznak koeficienta se spremeni.

Koeficienti imajo poseben pomen. Če vzamemo antilogaritem koeficienta, dobimo liho razmerje (to je konstanta e potencirana z vrednostjo koeficienta : liho razmerje = $e^{koeficient}$). Program to izračuna avtomatsko, če pa ga želimo izračunati sami, pa koeficient vstavimo v kalkulator (pazimo

na predznak) in pritisnemo tipko e. Liho razmerje nam pove verjetnost spremembe izida, če spremenimo eno enoto v neodvisni spremenljivki. Če je razmerje večje od 1 se bo povečal izid, če se poveča neodvisna spremenljivka ali če je prisotna dihotomna neodvisna spremenljivka. Če je razmerje manjše od 1 se izid zmanjša če se intervalna neodvisna spremenljivka poviša ali če je prisotna dihotomna neodvisna spremenljivka. Razmerje 1 pa nam pove, da ni sprememb v verjetnosti izida če se spremeni neodvisna spremenljivka. Za vrednotenje natančnosti lihega razmerja moramo izračunati 95% interval zaupanja. Program interval zaupanja avtomatsko izračuna, ročno pa zgornji interval zaupanja dobimo z uporabo raztegnjene formule za liho razmerje in predznakom za seštevanje, za spodnji interval pa z formulo in predznakom za odštevanje. Na računalniškem izpisu je poleg intervala zaupanja po navadi tudi standardna napaka.

95% interval zaupanja za liho razmerje = $e^{\text{koeficient} \pm 1.96\ (\text{standardna napaka})}$

Če je standardna napaka velika bomo seštevali oziroma odštevali večje število od koeficienta. To pomeni, da bo zgornja meja veliko večja od lihega razmerja in spodnja veliko manjša.

Wald test nam pove ali so koeficienti statistično pomembni, temelji pa na hi hvadratu ali na z porazdelitvi:

Hi kvadrat distribucija = $\{\text{koeficient/standardna napaka}\}^2$ oziroma

z distribucija = koeficient/standardna napaka

Wald test privzame velik vzorec, za statistično pomembnost velja $P < 0,05$.

Za določitev statistične pomembnosti določenega koeficienta lahko uporabimo tudi test razmerja verjetnosti (*the likelihood ratio test*) in točkovni test (*score test*). Test razmerja verjetnosti temelji na primerjavi verjetnosti če določena spremenljivka je v modelu z verjetnostjo če spremenljivke ni. Statistika sledi hi kvadrat porazdelitvi. Točkovni test temelji na izpeljanki verjetnostnega razmerja in prav tako sledi hi kvadrat porazdelitvi.

9. Preverjanje temeljne predpostavke analize

9.1. Kako naj preverim predpostavke multivariatnega modela?

Vsebina tega poglavja se pogosto nanaša na regresijsko diagnostiko (kot pri diagnosticiranju težav z regresijskim modelom). Eden izmed najuporabnejših modelov za oceno ali obstajajo težave z modelom je analiza ostankov (glej v naslednjem poglavju). V temu poglavju pa se avtor osredotoča kako lahko uporabiš residuals (= what remains or is a left over -> v slovenščini je prevodrezidual) da lahko podaš neko splošno mnenje o multivariatnih modelih. Prikazano bo tudi kako uporabiti ostanke in druge tehnike za prepoznavanje odstopanja od določenih predpostavk teh multivaritnih modelov.

9.2. Kaj so reziduali?? Kako jih uporabljamo za oceno uspešnosti modela?

Ostanki so razlike med opazovano in pričakovano vrednostjo. Lahko jih obravnavamo kot napake v oceni. Poleg »neobdelanih« rezidualov, obstaja še vrsto možnih transformacij ostankov za različne multivariatne postopke. Standardizirani reziduali so še posebej koristni pri tolmačenju linearne regresije modelov (npr. Cox-Snell, Martingale, deviantnost in Schoenfeld). Koristni so tudi za razlago sorazmerne analize nevarnosti. Različni reziduali so rezultat standardnih software programov.

Če želiš uporabiti reziduale za ocenitev splošnega prilegajanja večkratne linearne regresije ali binarne logistične regresije, je potrebno planirati surove reziduale na y osi in oceniti rezultat na x osi. V dobro prilegajočemu modelu, so reziduali blizu 0, kar pomeni, da so opažene in ocenjene vrednosti blizu drug drugega. Ko so opazovane vrednosti večje od pričakovanih, je vrednost rezidualov pozitivna in obratno; ko je opazovana vrednost manjša od pričakovane je rezidual negativen.

Definicija: reziduali so razlike med opazovano in pričakovano vrednostjo.

Ko so reziduali večji od nekaterih točk pričakovane vrednosti rezultata od ostalih točk (npr. velika pri ekstremnih vrednostih rezultata in majhna na vmesnih vrednostih), predlagamo kršitev v predpostavki modela kot ne-normalne porazdelitve, nelinearnost ali napačna specifikacija modela (npr. pomembno spremenljivko izpustimo). Če obstaja ena ali več točk z ekstremnimi reziduali, je tukaj možna rešitev »osamelec« (več o tem na 9.5.). Za logistično regresijo je lahko v pomoč pri ocenjevanju tudi

sprememba vrednosti Pearsonovega hi-kvadrata statistike; Y os v odvisnosti z ocenjeno verjetnostjo rezultata na x osi.

Za propocionalne kvote in multinomsko regresijo, je priporočljivo preveriti model prileganja z oblikovanjem dveh ali več (odvisno od tega ali imaš tri ali več kategorij izida spremenljivke) binarni logistični regresijski model ter oceniti vsakega od teh modelov. Za sorazmerno analizo »nevarnosti«, lahko uporabnost modela ocenjujemo s pomočjo Cox-Snell osntakov, kateri se lahko gibljejo od 0 in tja do 100. Kaplan-Meier -> preživetvena krivulja za rezultate, se navadno ustvari z uporabo Cox-Snell ostankov kot nekakšna časovna spremenljivka. Ta krivulja se nato primerja s funkcijo preživetja za izid, z enoto eksponentne porazdelitve. Če se sorazmeno tveganje ujema z modelom, bo krivulja zelo usklajena.

Pomembno tukaj je razumeti, da so reziduali bližje umetnosti kot sami znanosti. Med analizo lahko dobite zaskrbljajoče se vzorce rezidualov, čeprav so vaši podatki ustrezni glede na predpostavke modela. Na primer, z logistično regresijo lahko dobite ostanke, ki so navidez moteči – še posebej če imate dihotomne neodvisne spremenljivke (namesto intervalnih neodvisnih spremenljivk), četudi je vaš model primeren. Lahko se tudi zgodi, da so vaši ostanki videti v redu, a vendr ste kršili predpostavke modela. Majhni vzorci vam bodo še zlasti prinesli grde reziduale. To pomeni, da je bolj smiselno imeti večji vzorec, kajti tako vam bodo ostanki delali manj problemov.

9.3. Kako testiram normalno distribucijo (porazdelitev) in enako varianco predpostavke za multiplo linearen regresijski model?

Multipla linearna regresija predpostavlja normalno distribucijo in enako varianco okoli povprečja (glej poglavje 3.2.c). Če želiš preveriti ali ima tvoj rezultat normalno distribucijo in enako varianco okoli povprečja za katerokoli vrednost neodvisne spremenljivke, je potrebno načrtovati surovi ostanek za vsako neodvisno spremenljivko posebej in nato oceniti dobljeni rezultat to spremenljivko. Če so predpostavke pravilne, bi morali biti ostanki blizu vrednosti 0; ostale vrednosti pa nad ali pod 0; In te vrednosti ne smejo biti enakomerno razporejene.

Boljša metoda za odkrivanje nepravilnosti od normalne predpostavke, je da konstruiramo normalno predpostavko in načrtujemo standardizirane reziduale. Standardizirani reziduali, so torej SAMO ostanki deljeni s

standardnim odklonom od rezidualov. Normalna verjetnost načrtovanja je načrtovanje kumulativne pogostosti grafične lestvice. Če je predpostavka normalnosti pravilna, moramo dobiti ravno črto. Krivalja nakazuje na to, da je predpostavka normalnosti napačna. Točke daleč stran od naše linije pa so osamelci (o njih na 9.5.).

Nasvet: Uporabljaj normalno verjetnost izračuna za ugotovitev nepravilnosti o predpostavki normalnosti.

9.4. Kako preveriti predpostavko linearnosti multivariatnega modela?

Reziduali se lahko tudi uporabljajo za testiranje linearne predpostavke intervalnih neodvisnih spremenljivk, ki vstopajo v naš multivariatni model. Da testiraš, ali se ta intervalna neodvisna spremenljivka sklada z našim rezultatom v linearni ali binarni logistično regresijskim modelom, je potrebno izračunati surove vrednosti za vsako od neodvisnih spremenljivk posebej in nato oceniti rezultate teh spremenljivk. Če je razmerje linearno, bodo točke simetrične zgornji in spodnji ravni liniji, s približno enako mero širjenja vzdolž te črte. V primeru sorazmernega tveganja analize linearnosti intervalnih spremenljivk, je to mogoče oceniti s pomočjo Martingal-ovih rezidualov, ki so linearna transformacija Cox-Snell rezidualov, kateri so v razponu od negativne neskončnosti do 1. Matingal-ovi reziduali (y os) so izračunani nasprosti intervalnega prediktorja (x os). Oblika »zglajene krivulje« bo pokazala ali je razmerje linearno ali katero drugo. Ta informacija ti lahko pomaga ugotoviti ali je potrebno, da tvoj interval transformiraš da zadovoljiš linearni predpostavki. Druga metoda ocenjevanja linearne predpotavke je da ustvariš multiplo dihotomno spremenljivko na enakih intervalih tvoje spremenljike. Metoda je zelo prilagodljiva in se lahko uporablja z multiplimi logističnimi regresijami (npr. binarni, proporcionalne kvote,...)

9.5. Kaj so ostanki in kako jih zaznamo v multivariatnih modelih?

Reziduali so točke, katere ne sledijo nekemu vnaprej določenemu vzorcu drugih točk.

Kako jih zaznamo? Iz neobdelanih rezidualov lahko sicer dobimo nekaj vpogleda, a vendar jih je lažje zaznati z uporabo standardiziranih rezidualov, preko finančnih vzvodov in Cook razdalje.

Kaj so standarizirani reziduali? To so reziduali, katere deliš s standardnim odklonom. Surovi reziduali so odvisni od obsega odvisne spremenljivke in je ne moremo določiti brez da standardiziramo reziduale. S standardizacijo pa tako odpravimo enate in lahko ustvarimo neke smernice za velike vzrorce ostankov (se pravi so bolj primerni ko uporabljamo velike vzorce).

Osamelci na ekstremnih vrednostih od neodvisnih vrednosti, se izvajajo na večjih zvodih, kot tisti ki so bližje sredini. Da pa mi sploh imamo osamelce, lahko ima velik vpliv na sam rezultat.

Vpliv se pa nanaša na to kako odstraniti spremembe opazovanja ocenitve koeficienta modela.

12. Veljavnost modelov

12.1 Kako lahko potrdim veljavnost svoji modelov?

Veljaven oziroma potrjen model je tisti iz katerega lahko delamo pravilne sklepe. Veliko faktorjev lahko skazi veljavnost modela na primer nenatančnost, nepravilno merjenje, bias v sestavi študije in v vzorčenju ali napačna specifikacija samega modela. Ker je razvoj modela poveča možnosti za obdržanje vrednosti prvotnih podatkov, ki smo jih dobili kot izid, modeli na splošno ne bo delovali tako dobro z novimi podatki kot s prvotnimi podatki. To je posebno pomemben problem, ko izdelujemo modele, ki morajo predvidevati diagnozo ali prognozi in imeti visoko stopnjo zanesljivosti. Čeprav predikcije/predvidevanja, ki temeljijo na osnovi orginalnih podatkov ne bodo tako točne kot novi podatki je pomembno vprašanje kako velik je padec v zmogljivosti/preformansu. Če je padec majhen se za model reče, da je zanesljiv.
Metode za zagotavljanje veljavnosti so naslednje:

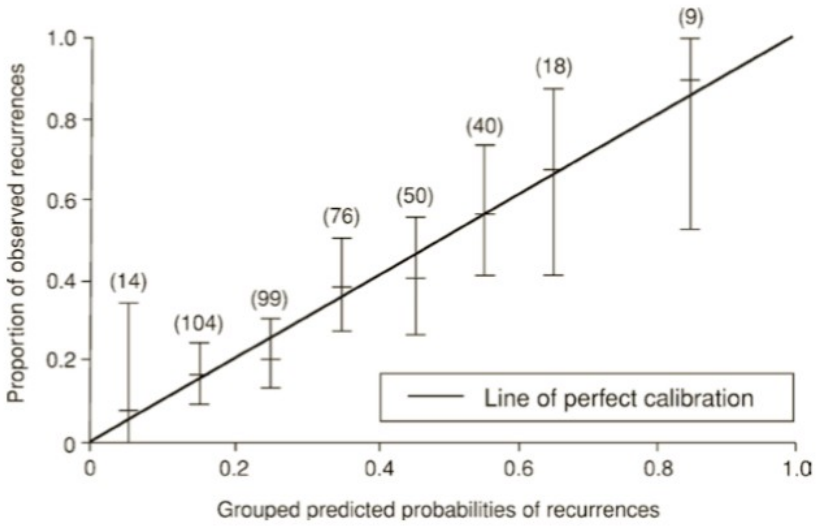
- Zbiranje novih podatkov
- Ločevanje svojega prvotnega data seta (zbirke podatkov)
- Jack-knife metoda
- Bootstrap brez vprašanj

Najboljša metoda za določanje veljavnosti empiričnih modelov je da se zbirajo podatki in da se preveri model z novimi podatki, torej modelu najprej damo za obdelati originalne podatke in nato nove, zato, da vidimo kako se le ta obnese. Podobno so ugotovili z delom prognostičnom

modelom za ocenjevanje verjetnosti za preživetje pacientov z primarno melanomo (opisano v sekciji 2.5). Raziskovalci so ugotovili, da so 4 faktorji pravilno klasificirali vitalni status (mrtev ali živ) pri 74% pacientov. Da bi naredili model veljaven, so preiskovali uspeh ta model štirih spremenljivk z gledanjem predikcij rezultatov med 142 pacienti, ki jim je bila diagnosticirana primarna melanoma v istem centru takrat in nato čez dve leti z istim številom ljudi, ki imajo prav tako isto diagnozo. Ko so čez dve leti gledali rezultate istega vzorca so ugotovili, da je model pravilno klasificiral 69% rezultatov, kar je relativno majhno odstopanje v preformansu v primerjavi z originalnim modelom. Čeprav testiranje modela z drugim setom pacientom ojača veljavnost modela ni tako močna kot veljavnost modela, ki je bil testiran v drugem centru. Razlog je, da se model ne bo tako dobro izkazal v drugačnih okoliščinah (prevalenca druge bolezni, drugi vzorec pacientov, druga klinična praksa, začasne spremembe…). V našem primeru, primeru primarne melanom je bila edina razlika v okoliščinah ta, da se je razlikovalo leto diagnoze, kar jo naredi manj rigorozno kot če bi raziskovalci spremenili institucijo v kateri poteka. Z določanjem veljavnosti z metodo razdelitve skupine tako naključno razdeliš svoje podatke na dva dela – v set za vajo in v veljavnostni set. Tadva dela sta lahko enaka oziroma enakovredna lahko pa razdeliš podatke tako, da je set za vajo večji kot set za veljavnostni set. Svoj model razviješ na setu za vajo (derivatni set – derivation set) in ga preveriš na setu za veljavnost (konformacijski set – confrimatory). Tak test veljavnosti (split group) razdelitve skupin je bil uporabljen za testiranje model, ki je bil ustvarjen za napovedovanje pogostosti napadov (seizures). Vzorec je vseboval 1013 ljudi, ki niso jemali tablete proti napadom že najmanj 2 leti. Raziskovalci so izkoristili obstoječe podatke za razvoj prognostičnega modela. Zbiranje dodatnih podatkov namreč ni bila opcija, zato so namesto tega razdelili vzorec na dva dela tako, da je bilo 60% ljudi v derivacijskem setu (setu za vajo) in 40% v konformacijskem testu (testu za veljavnost). Z uporabo derivacijkega sta do razvili proporcionalen hazardski model, ki je imel 8 prognostičnih faktorjev. Da bi vzpostavili veljavnost modela so ocenili verjetnost ponovitve napada za vsakega pacienta, ki je bil v konformacjskem setu (testu za določanje veljavnosti). V 8 različnih skupin so razvrstili paciente po dvigajoči se ocenjeni verjetnosti za napad.

V sliki 12.1 je prikazano, da so vsako izmed 8 skupin primerjali s proporcijo dejanskih ponovitev napada, ki jih je napovedal model. Stolpci, ki prikazujejo intervale zaupanja in »dash«, ki je blizu sredine kažejo mediano predvidene vrednosti. Če bi bila veljavnost popolna, bi vse »dashi« padli direktno na diagonalno črtno. Medtem a so »dashi« blizu diagonalne črte,

intervali zaupanja pa so široki, še posebej za tiste predvidene vrednosti kjer je število subjektov majhno.
Slika 12.1



Opazili boste, da slika 12.1 uporabi isto tehniko kot slika 8.1. Slika 8.1. prav tako temelji na primerjanju napovedanih opazovanih verjetnosti. Razlika je v tem, da je v sliki 8.1. napovedana verjetnost na istih subjektih. Še kot zadnje glede veljavnosti, ki uporabi ali drugi vzorec ali pa razdelitev vzorca je ta, da ko je bil model proglašen za veljavnega, bodo raziskovalci pogosto združili multiple vzorce ali pa združili razdeljen vzorec v končni model. Raziskovalci to naredijo za to, da dobijo večji model in tako bolj ozke intervale zaupanja. V primerih kjer ni praktično zbirati več podatkov ali razdeliti vzorec pa se lahko uporabi jack-knife metoda (večkrat jo kličejo cross-validation = veljavnost križanja). S to metodo lahko postopoma zbrišemo subjekte iz data seta, po enega vsakič in tako ponovno izračunaš tako, da vsakič manjka en subjekt. To nam omogoča, ocenitev dveh stvari. Prvič, da ocenimo pomembnost enega subjekta v rezultatih. Drugič pa lahko ocenimo veljavnost modela, v primeru, da se ta popolnoma spremeni le z odvzetjem enega subjekta.
Drugič, enkrat ko zavržemo primer lahko napovedujemo subjektove rezultat iz preostalih primerov. To je narejeno postopoma tako, da napovedujemo vrednosti vsakega subjekta z uporabo ostalih subjektov. V tem smislu je jack-knife metoda kot metoda razdelitve: razdelitev je cel vzorec minus tisti en primer (derivacijski set) proti drugemu primeru (konformacijski set). Ko imamo majhen vzorec je ta jack-knife metoda veliko bolj občutljiva kot metoda razdelitve vzorca. Jack-knife metoda je zelo lahka za izvedbo v multipli linearni regresiji. Bootstrap metoda nudi omejeno podporo veljavnosti modelu, ki ga preiskujemo. Pri bootsrap metodi vzamemo naključne vzorce subjektov ven iz data seta z zamenjavo

(to pomeni, da po tem ko je primer izbran postane ponovno zmožen, da je izbran). Tako lahko naključni vzorci vsebujejo iste subjekte več kot enkrat, medtem, ko pride do možnosti, da nekateri subjekti sploh nikoli niso vključeni. Ko je vzorec enkrat zastavljen, testiramo moč razmerij, ki smo jih našli v glavnem modelu naključnih vzorcev. Rezultati iz teh vzorcev so lahko uporabljeni za sestavo 95% interalov zaupanja z izključevanjem ekstremov 2,5% in 97,5% vrednosti. Če so intervali zaupanja relativno ozki, to pomeni, da so rezultat bolj zanesljivi. Na primer, Hamberg in kolegi so uporabili logistično regresijo da so naredili model, ki je temeljil na kliničnih in laboratorijskih podatkih, ki bi pravilno napovedovali cirozo med 303 moškimi, ki zlorabljajo alkohol. Če bi imel tak model klinično pomembnost bi lahko zmanjšal potrebo po biopsiji jeter. Njihov model je bil tako natančen in občutljiv, da je pravilno napovedal možnost ciroze v 88%. Poleg dajanja intervalov zaupanja, bootstrap metoda omogoča tudi dajanje mediane koeficientov in mediane standardne napake za naključne vzorce. Je pa ta metoda sicer manj zanesljiva kot metoda razdelitve in Jack-knife metoda. Še vedno pa veliko zdravnikov raje zaupa svoji intuiciji kot pa da bi zaupali modelom, ki niso veljavni.

13.3 Kaj so klasifikacijska in regresijska drevesa (CART) in kako naj jih uporabim?

Klasifikacijska in regresijska drevesa so tehnika za ločevanje subjektov v podskupine glede na rezultate. Najlažje je tej tehniki slediti vizualno. V sliki 13.1 lahko vidimo algoritem za ocenjevanje tveganja srčnega napada, ki je bil ustvarjen z uporabo CART. Tehnika CART poskuša razdeliti vzorec v podskupine z toliko pacienti, ki imajo enak rezultat v eno skupino (srčni kap) in čim manj pacientov z drugačnim rezultatom v drugo skupino (majhna verjetnost srčne kapi). Izbiranje iz variabilnosti kandidatov bo CART nadaljeval z razdelitvijo dokler ne bo prišel do točke kjer ni več možno razdeliti vzorca na pod skupine z različno verjetnostjo za srčno kap. Če pa CART razdeli vzorec v podskupine, ki ne zadoščajo dovolj natančno, lahko drevo še zmeraj »obrežemo«. Ta tehnika je zelo podobna logistični regresiji v tem, da ocenjuje dihotomne rezultate tako, da postopoma izbere najmočnejše faktorje tveganja za rezultate, ki jih potrebujemo. Tehnika CART je bolj primera za data sete v katerih so interakcije. Pomembna prednost te tehnike je, da se njihovi rezultati zelo približajo odločitvam, ki bi jih naredili pediatri/zdravniki. TA tehnika namreč pretehta vse negativne in pozitivne aspekte in nato naredi neko odločitev, pri čemer se pokaže razlika saj zdravniki velikokrat ne pretehtajo vseh možnosti. Ko so dali zdravnikom možnost zanašanja se na ta odločitvena drevesa so se na njih zanašali oziroma iz njih črpali podatke le v 46% primerov. Zanašanje na to metodo pa je bilo odvisno od izkušenj zdravnika, kljub temu, da se je tehnika odrezala bolje kot pa pediatri/zdravniki na univerzi.

Slika na drugi strani

15. POVZETEK:

1. korak: Glede na tip spremenljivke, ki jo imate, uporabite tabelo 3.1, da določite vrsto multivariabilnega modela za izvedbo (če imate ponavljajoče izide, glejte tabelo 11.2)

2. korak: opravite univariantno statistiko, zato da bi razumeli porazdeljenost vaše neodvisne in končne spremenljivke. Ocenite tudi katere vrednosti so neveljavne, pomembnost odstopanja od normalne porazdeljenosti in razlike v vrednostih.

3. korak: opravite bivariantno analizo, če so vaše neodvisne spremenljivke nasprotne vašim končnim.

4. korak: če imate nominalne neodvisne spremenljivke, jih preoblikujte v več dihotomnih spremenljivk.

5. korak: ocenite, ali so intervalne-neodvisne spremenljivke v linearni korelaciji s končnimi. Če niso, preoblikujte spremenljivko ali ustvarite več dihotomnih spremenljivk.

6. korak: zaženite korelacijsko matriko. Če ima katerikoli par neodvisnih spremenljivk korelacijski koeficient večji od 0.90 (multikolinearnost), se odločite kateri par boste obdržali in katero spremenljivko boste obdržali in katero izločili. Če kateri par spremenljivk korelira med 0.80 in 0.90, razmislite o tem, da enega izločite.

7. korak: ocenite koliko manjkajočih podatkov boste imeli v vaši multivariabilni analizi. Izberite strategijo ravnanja z manjkajočimi podatki iz tabele 6.4.

8. korak: izvedite analizo.

9. korak: ocenite kako dobro se vaš model ustreza podatkom.

10. korak: ocenite moč vaših individualnih spremenljivk pri ocenjevanju izida.

11. korak: uporabite regresivno diagnostiko za presojo temeljnih postavk vašega modela in določite strategije za izboljšanje prileganja vzorca.

12. korak: odločite se ali boste v vzorec vključili interakcijske izraze.

13. korak: premislite, ali bi bilo možno, da bi svoj model preverili.

14. korak: Objavite svoje rezultate v reviji