4. predavanje RVM – Kvantitativne metode
Borut Kodrič, Koper 4.6.2010

~~~~~~

1. del

Na podlagi česa ugotovimo kako sta dve spremenljivki med sabo povezani. Meri, ki nam kaže povezanost dveh spremenljivk, pravimo korelacijski koeficient (grška črka ro). Ta koeficient leži na intervalu med minus ena in ena.

$$-1 < \rho_{yx} < 1$$

Njegov predznak nam kaže smer povezanosti. Če je smer povezanosti negativna pomeni, da če se ena spremenljivka poveča, se druga manjša. Če je pozitivna pomeni, če se ena spremenljivka poveča, se tudi druga tudi poveča. **Absolutna vrednost korelacijskega koeficienta** nam kaže kako močno sta spremenljivki povezani (kako spreminjanje ene sledi spreminjanju druge). Bližje kot je vrednosti ena (1), močnejša je povezanost med dvema spremenljivkama. Bližje kot je po absolutni vrednosti nič (0) šibkejša je povezanost. Če bi y bila natančna funkcija x (deterministična), t.j. da bi natančno vsaki vrednosti x ustrezala natančno ena sama vrednost y. Spreminjanje y bi natančno sledilo spreminjanju x. V praksi takih odvisnosti ni.

$$y = f(x) + \varepsilon$$ povezanost med dvema spremenljivkama

V praksi nastopajo **odvisnosti**, ki so **v obliki stohastičnih odvisnosti** (povezanosti). Razlika je ta, da se tu zavedamo, da na spreminjanje y ne vpliva samo nek izbran in predpostavljeni dejavnik ampak da poleg tega dejavnika vplivajo tudi neznani ali slučajni dejavniki. Ko spremenimo x se y včasih lahko spremeni, lahko se pa tudi ne. Enkrat se spremeni v plus enkrat v minus. **V povprečju** y sledi spreminjanju x. Na nivoju vsake standardne enote pa to ni nujno.

Kaj je lahko vsebinsko ozadje povezanosti med dvema spremenljivkama? V grobem gledano je lahko dvoje. Imamo tip povezav, kjer imamo t.i. **VZROČNO POSLEDIČNO ZVEZO**. V takem primeru je smer vpliva vedno od spremenljivke, ki jo označimo z x (**neodvisna spremenljivka**), k spremenljivki, ki jo označimo z y (**odvisna spremenljivka**).

$$x \rightarrow y$$ vzročno posledična zveza

(Primer: če je y število točk na izpitu in x število ur učenja, je razumljivo, da glede na to koliko se učimo, lahko pričakujemo nek določen rezultat. Ne velja pa obratna zveza. Najprej se je potrebno učiti in nato pristopiti k izpitu).

REGRESIJSKA ANALIZA v osnovi temelji na proučevanju takih odnosov. **Izhaja iz proučevanja vzročno-posledičnih odnosov**.

Spoznali naj bi najpreprostejši model bivariantne regresijske analize, kjer je ena odvisna in ena neodvisna spremenljivka. Sledi še **MODEL MULTIVARIANTNE REGRESIJSKE ANALIZE**, kjer dopuščamo, da na nek proučevani pojav vpliva večje število dejavnikov (Primer: število točk na izpitu je odvisno od števila ur učenja, od nivoja predznanja, od prisotnosti na predavanjih, itn.)

Ločimo dva termina:

POVEZANOST – KORELACIJA **x1 ←→ x2** (obojesmerna povezava; ni mogoče opredeliti kaj je vzrok kaj je posledica)

ODVISNOST – REGRESIJA **x → y**
~~~~~~

Proučevanje odvisnosti - 1

- Zanima nas variiranje posameznega pojava ob sočasnem variiranju enega ali več drugih pojavov.
- Značilnosti želimo izraziti kvantitativno, da bi lahko ugotovili stanje in razmere v preteklosti oziroma napovedali razvoj pojava na osnovi poznavanja enega ali več drugih pojavov.

METODA FAKTORSKE ANALIZE je osnovana na analizi medsebojnih korelacij. Tu potrebujemo neko vsebinsko poznavanje oz. neko teorijo, da pojav x vpliva na y.

Pri analizi korelacij ni nujno, da gre za neko vsebinsko pojasnitev (kaj je vzrok, kaj je posledica). Kljub temu ugotavljamo, da sta dve spremenljivki med sabo povezani. Največkrat je vzrok temu, da na neko množico spremenljivk vpliva nek zunanji dejavnik na enak način. (Primer: x1 je pridelek pšenice, ki ga opazujemo po slovenskih občinah, x2 je pridelek koruze. Če bi zbrali podatke po slovenskih občinah, za letino, bi ugotovili, da sta ti dve spremenljivki med sabo močno povezani. V tistih občinah, kjer bi ugotovili velik pridelek pšenice bi zelo verjetno ugotovili tudi velik pridelek koruze. Tam kjer bi ugotovili manjši pridelek pšenice pa verjetno tudi manjši pridelek koruze. Tako na en kot na drug pridelek na enak način vpliva rodovitnost zemlje, intenzivnost obdelovanja, vremenski pojavi. Zaradi teh zunanjih dejavnikov se odraža vpliv na teh spremenljivkah, ki sicer lahko nimata nikakršne vsebinske zveze.

Pri faktorski analizi tem dejavnikom pravimo **SKUPNI FAKTORJI**.

F1(skupni faktorji) → x1 ←→ x2

REGRESIJSKA ANALIZA bo namenjena **analizi vzročno posledičnih zvez** (smer vpliva gre vedno od neodvisne spremenljivke k odvisni). Pri regresijski analizi naj bi vse **spremenljivke**, ki vstopajo v regresijski model, bile **prave numerične spremenljivke** (**nanašajo se na pojave, ki jih je mogoče izmeriti**). Gre za vrednosti, ki imajo svojo mersko enoto.

FAKTORSKA ANALIZA bo namenjena **analizi medsebojnih, obojesmernih korelacij**, katerih prisotnost pripišemo obstoju nekih zunanjih dejavnikov. Faktorska analiza je namenjena **analizi pojavov, ki jih je praktično nemogoče neposredno izmeriti**. (Primeri takih pojavov na področju družboslovja: zadovoljstvo, organizacijska klima v podjetju,..). V teh primerih skušamo poiskati neke posredne kazalnike za tak pojav (anketna vprašanja).

Proučevanje odvisnosti - 2

- Odvisna spremenljivka (y): posledica
- Neodvisna spremenljivka (x): vzrok

- *Cilj:* postaviti napoved za spreminjanje odvisnega pojava (y) v odvisnosti od drugega pojava (x).

REGRESIJSKI MODEL

x → y

Kot najpreprostejši primer imamo eno neodvisno in eno odvisno spremenljivko. Naša naloga bo opisati odnos med tema dvema spremenljivkama z nekim kvantitativnim modelom. Če poznamo kaj vse vpliva na naš proučevani pojav, potem tak pojav veliko lažje napovemo. (Primer: če na eni strani vemo, da se brezposelnost spreminja zaradi nivoja gospodarske aktivnosti, demografske strukture prebivalstva, izobrazbene strukture, lahko s predvidevanjem oz. gibanjem dejavnikov predvidimo gibanje brezposelnosti. Osnovni motiv je spoznati kako se spreminja naš pojav glede na neke dejavnike in to z namenom, da se na zgodovinskih izkušnjah skušamo nekaj naučiti in ta spoznanja uporabiti za napovedovanja kaj bi se zgodilo, če…)

Deterministična odvisnost

- Spremenljivka y je funkcija spremenljivke x, če za vsako dano vrednost spremenljivke x zavzame odvisna spremenljivka y eno samo, točno določeno vrednost

$$y=f(x)$$

Odvisnost bi se lahko pojavljala v obliki **DETERMINISTIČNE ODVISNOSTI**. Gre samo za teoretično izhodišče. V praksi se zavedamo, da na naš proučevani pojav nikoli ne vpliva samo en dejavnik.

Stohastična (korelacijska) odvisnost - 1

- Odvisna spremenljivka y ni odvisna samo od proučevane spremenljivke x, temveč tudi od drugih dejavnikov, ki so v danem primeru **neznani dejavniki**.

  Primer: Preučujemo odvisnost uspešnosti študentov pri predmetu Raziskovalna metodologija v družboslovju od števila ur namenjenih za pripravo na izpit. Vemo, da poleg obsega učenja, na uspešnost pri izpitu vplivajo še drugi dejavniki (predznanje, naklonjenost predmetu...)

Tudi če poznamo množico dejavnikov se lahko zgodi nek slučajni dogodek. Zato ni mogoče vedno natančno predvideti kaj se bo z našim pojavom zgodilo. V tem primeru pravimo, da gre za **STOHASTIČNO ODVISNOST**. Naš odvisni pojav je v neki meri rezultat vpliva neodvisne spremenljivke.

Stohastična (korelacijska) odvisnost - 2

$Y=y'(x)+\varepsilon$

ε: vpliv neznanih (med njimi vsaj slučajnih) dejavnikov

- *Enosmerna korelacija:* odvisna spr. Y odvisna od pojasnjevalne spr. X, obratno pa ne drži
- *Dvosmerna korelacija:* spremenljivki sta odvisni druga od druge

Poleg neodvisne spremenljivke je vedno prisoten vpliv neznanih dejavnikov, ki povzročajo, da se lahko ob isti spremembi x, y različno spreminja. V povprečju pa y le sledi x.

Linearni bivariatni regresijski model-1

- Odvisno spremenljivko (y) proučujemo v odvisnosti le od ene neodvisne spremenljivke (x).
- Predpostavljamo linearno odvisnost (regresijska funkcija je premica).
- Linearni bivariatni regresijski model:

$$y_i=\alpha+\beta x_i+\varepsilon_i$$

Predpostavimo da je spremenljivka y odvisna od enega samega dejavnika. To je zelo velika poenostavitev. V praksi je pojavov, odvisnih od enega samega dejavnika, relativno malo. V veliki večini primerov je vsak pojav odvisen od večjega števila dejavnikov. Zato **ima ta model zelo omejeno uporabno vrednost**.

Linearni bivariatni regresijski model-2

- Slučajnostni odkloni ε_i so porazdeljeni normalno:

$$\varepsilon_i = N(0,\sigma_{\varepsilon_i}^2)$$

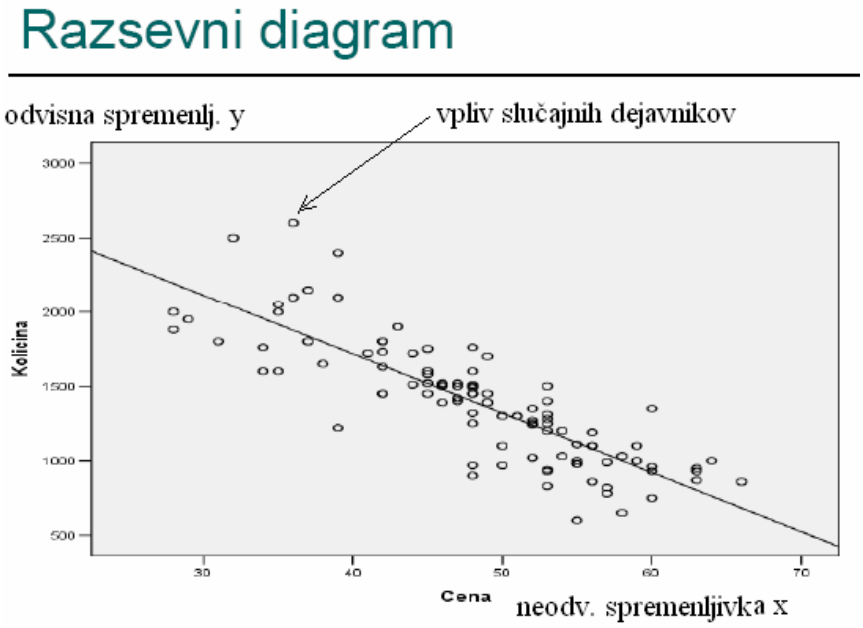
- Slučajnostni odkloni so med seboj neodvisni:

$$Cov(\varepsilon_i,\varepsilon_j)=0; i\neq j$$

- Varianca za y_i in za ε_i ni odvisna od vrednosti za x_i in je konstanta (homoskedastičnost):

$$\sigma_y^2=\sigma_\varepsilon^2=konst.$$

Predpostavke, ki so v ozadju tega modela, so podobne predpostavkam multipla regresijskega modela. Slučajni odkloni so normalno porazdeljeni z vrednostjo nič in nekim standardnim odklonom. Ta vrednost nič pomeni, da so slučajnostni odkloni v povprečju realno enaki nič. Toliko kot je pozitivnih vplivov slučajnostnih odklonov, toliko mora biti tudi negativnih.



Preprost primer analize povezanosti, zgolj dveh spremenljivk, omogoča tudi grafično analizo te odvisnosti. Pripomoček, ki nam to omogoča je **razsevni diagram** (scatter diagram). Na abscisni osi nanašamo vrednosti neodvisne spremenljivke x, na ordinatno os nanašamo vrednosti odvisne spremenljivke y. Opazujemo kako se količina prodanega blaga spreminja glede na ceno. Točke pomenijo posamezne prodajalne. Opazujemo neko množico prodajaln. Za vsako prodajalno narišemo točko. Taka slika nam omogoča kar precejšen vpogled v analizo povezanosti. Lahko hitro vidimo kakšna je smer povezanosti med dvema pojavoma. V tem primer je smer negativna. Če se cena povečuje se količina **v povprečju** zmanjšuje.

$$y = f(x) + \varepsilon$$

(y predstavlja KOLIČINO, x CENO, ε predstavlja dejavnike, ki vplivajo na prodajno količino)

Če bi bila zveza taka, da se količina spreminja natančno ob vsaki spremembi cene, potem bi vse točke morale ležati natančno na premici. Na prodano količino pa ne vpliva samo cena. Vpliva lahko tudi prijaznost prodajalcev, lokacija prodajalne, poreklo blaga, itd. Ti dejavniki so zajeti v ε. Odnos zato ni funkcijski, ampak ga moramo vedno opisovati z besedico **V POVPREČJU**.

- **Predpostavka slučajnostnih odklonov** pomeni, da je nekaj točk pod premico (v teh primerih so slučajnostni odkloni negativni), nekaj pa nad premico (slučajnostni odkloni so pozitivni). **Slučajnostni odkloni so normalno porazdeljeni, kar pomeni, da velika večina točk leži v bližini regresijske funkcije. Povprečje slučajnostnih odklonov je enako nič. Pomeni da funkcija ki jo ocenjujemo, leži vedno točno na sredini vseh točk. Če bi sešteli vse epsilone (ε), vsota negativnih ε mora biti vedno točno enaka vsoti pozitivnih ε. V povprečju naj bi vse točke ležale na premici**.
- Druga predpostavka v tem modelu je, **da so slučajnostni odkloni med sabo nepovezani**, neodvisni (glej točke na izročku Linearni bivariantni regresijski model-2 zgoraj!). To da ena enota proda več druga manj nima nobene medsebojne zveze.

- **PREDPOSTAVKA HOMOSKEDASTIČNOSTI**. Pravi da naj bi bila varianca slučajnostnih odklonov konstanta. Na sliki (razsevni diagram) bi lahko narisali en pas, ki bi zajemal točke. Bil naj bi ves čas enako širok. To pomeni da je **varianca okrog regresijske premice ves čas konstanta**.

Na naši stopnji se bomo s temi predpostavkami samo seznanili, ne bomo pa preverjali ali so v nekem regresijskem modelu izpolnjene ali ne. Naša naloga bo omejena samo na to, da bomo skušali na podlagi podatkov oceniti regresijski model. Skušali bomo priti do enačbe te premice, ki kaže kako se odvisna spremenljivka spreminja glede na neodvisno.

Analiza razsevnega diagrama

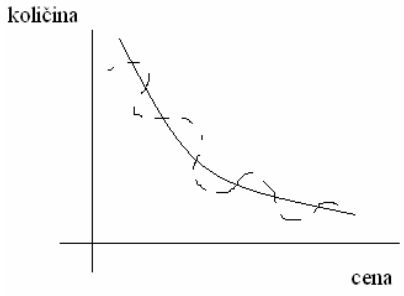
- Pozitivna ali negativna korelacijska odvisnost
- Linearna ali nelinearna odvisnost
- Visoka ali nizka odvisnost

Na podlagi grafičnega prikaza, kot je razsevni diagram, lahko enostavno razberemo tri ključne elemente te analize.

- pri analizi odvisnosti nas zanima ali gre za negativno ali pozitivno zvezo oz. odvisnost
- zanima nas ali je odvisnost linearna ali ni linearna
- zanima nas njena jakost (ali sta dve spremenljivki močno povezani (y močno odvisen od x ali ne))

Za vse to obstajajo tudi kazalniki, s čimer numerično ovrednotimo ali je odvisnost pozitivna ali negativna, linearna ali nelinearna, šibka ali močna. Že na podlagi analize razsevnega diagrama lahko kar precej natančno napovemo vrednost kazalnikov.

V našem primeru (razsevni diagram) **je negativna**. **Če se x povečuje se y v povprečju zmanjša**. Na podlagi oblike lahko razsodimo ali je linearna ali nelinearna. Točke so bolj ali manj orientirane v nekem pasu tako da lahko rečemo, da **gre za linearno povezanost**.



črtkano predstavlja porazdelitev točk. V tem primeru ne moremo govoriti o linearni povezanosti.

Šibka ali močna odvisnost: (tega ne presojamo na podlagi naklona!) te premice ampak na podlagi tega ali so točke bolj skoncentrirane ob sami premici ali so nekoliko bolj oddaljene od premice. Če so koncentrirane na sami premici je odvisnost močnejša, bolj kot so oddaljene, šibkejša je odvisnost. V tem primeru je večji vpliv slučajnih dejavnikov.

Mere linearne korelacije in regresije-1

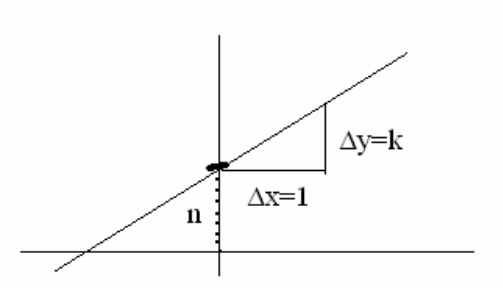
- Enačba regresijske premice:

$$y' = \alpha + \beta x$$

- α: regresijska konstanta: odsek regresijske premice y' na ordinatni osi v razsevnem diagramu
- β: regresijki koeficient: pokaže povprečno spremembo vrednosti odvisne spremenljivke y, če se vrednost pojasnjevalne spremenljivke x poveča za eno enoto

Če imamo v mislih najpreprostejši primer analize odvisnosti, kjer gre za analizo odvisnosti zgolj med dvema pojavoma. Končni cilj bi bila ocena enačbe, konkretno parameter α in parameter β.

enačba premice: **y = kx + n**



n pomeni vrednost y kadar je x enak nič. Pomeni odsek na ordinatni osi. **k** pomeni koliko se spremeni y kadar se x poveča za eno enoto. Če je (delta x) ena je (delta y) enak k. Pri naši enačbi (**y' = α + βx**)je n označen z α (je tista vrednost odvisne spremenljivke kadar je neodvisna enaka nič). Tako kot je k (kot smerni koeficient) kazal naklon premice, je v statistični simboliki označeno z β.

y = kx + n predstavlja funkcijski odnos. Zato se ob vsaki spremembi x za eno enoto, y vedno natančno spremeni za k.

V našem primeru vse točke ne ležijo na sami regresijski premici, zato je treba ta odnos razlagati **v povprečju**. Če se (na razsevnem diagramu) cena poveča za 5 denarnih enot se bo količina zmanjšala v povprečju npr. za 400 enot.

Mere linearne korelacije in regresije-2

- **Determinacijski koeficient:** delež variabilnosti odvisne spremenljivke pojasnjen z linearno odvisnostjo od neodvisne spremenljivke

$$0 \leq \rho_{yx}^2 \leq 1$$

- **Korelacijski koeficient:** meri stopnjo in smer linearne korelacije med spremenljivkama x in y

$$-1 \leq \rho_{yx} \leq 1$$

KORELACIJSKI KOEFICIENT (**kaže** nam **jakost in smer povezanosti**) Glede na predznak **kaže smer povezanosti**, ki je bila v našem primeru negativna. (Če vzamemo primer y, ki je ocena na izpitu, x so pa ure učenja bi recimo korelacijski koeficient pokazal verjetno plus vrednost. Bolj kot se učite, višjo oceno v povprečju dobite). Po absolutni vrednosti nam korelacijski koeficient pokaže jakost odvisnosti. Bližje kot je ena (1), po absolutni vrednosti, močnejša je odvisnost in bližje proti nič, šibkejša je odvisnost.

DETERMINACIJSKI KOEFICIENT skuša razčleniti **varianco odvisne spremenljivke**. V našem primeru se količina spreminja od prodajalne, kjer je bilo nekaj več kot 200 enot pa vse do prodajalne kjer je bila prodana količina 2600 enot. Na tem intervalu količina od prodajalne do prodajalne variira. Determinacijski koeficient skuša oceniti kolikšen del te variabilnosti, odvisne spremenljivke, lahko pojasnimo s spreminjanjem neodvisne. V večji meri kot se variabilnost odvisne pojasni z variabilnostjo neodvisne, večji je determinacijski koeficient manjši je delež nepojasnjenih dejavnikov. Naša težnja je, da bi poznali čim večji delež variiranja ker to pomeni da poznate ključen dejavnik, pomeni da svoj pojav dobro poznate. Tudi vse ocene, ki bi se izdelovale na podlagi takega modela, bi imele zelo majhno napako.

V praksi lahko pozabimo na populacijske podatke in se je treba, v večini primerov, zadovoljiti z vzorčnimi podatki. Kaj se spremeni kadar regresijsko analizo postavimo na vzorčne podatke. Stvar se spremeni z vidika označb minimalno. Po pomenu pa je to velika razlika. Ko imamo opravka z vzorci nam vsak vzorec da nekoliko drugačno oceno parametra. Ne spoznamo pa nikoli vseh vzorcev, da bi dobili vse ocene parametrov, na podlagi njih izračunali povprečje in na podlagi tega dobili pravo vrednost.

		korelacijski	determinacijski	
populacijski podatki	$y' = \alpha + \beta x$	ρ_{yx}	ρ^2_{yx}	, ε
vzorčni podatki	$y'' = a + bx$;	r	r^2	, e

Analiza odvisnosti za vzorčne podatke - 1

- Kazalci odvisnosti, ki jih izračunamo so le ocene pravih vrednosti za te kazalce.
- Kakovost ocene regresijske enačbe (y'') za napovedovanje vrednosti:
  - e=(y-y'')
  - vzorčna variabilnost ocen parametrov regresijske premice a in b

Zavedamo se da poznamo enega samega iz vseh možnih vzorcev. Poznamo tudi eno samo izmed možnih ocen. V rokah moramo imeti nek mehanizem, da bi ugotovili koliko je ta ocena verjetno pravilna ali verjetno ni pravilna. Z gotovostjo nikoli ne moremo presoditi ali je ocena, ki smo jo dobili pravilna ali ne.

Analiza odvisnosti za vzorčne podatke - 3

- Ocena variance slučajnostnih odklonov:

$$s_e^2$$

- Ocena standardne napake ocene:

$$s_e = \sqrt{s_e^2}$$

- Ocene standardnih napak ocen regresijskih koeficientov:
  - se(a)
  - se(b)

Z ocenami (preizkušanje domnev) z neko verjetnostjo ocenimo ali je naša ocena parametra z neko verjetnostjo pravilna ali ne. Ključni preizkus, ki se uporablja v okviru regresijske analoze, je **PREIZKUS NEODVISNOSTI**.

Preizkušanje domnev o regresijskih koeficientih - 1

- Preizkus neodvisnosti:

$$H_0 : \beta = 0$$

$$t = \frac{b}{se(b)} \qquad t_{(\alpha, m=n-2)}$$

- Če zavrnemo $H_0 : \beta = 0$, pri dani stopnji tveganja sprejmemo sklep, da je pojav y odvisen od pojava x.

To je preizkus, ki se nanaša na vrednost regresijskega koeficienta β.

Ničelna domneva: Ho : $\beta = 0$

Alternativna domneva: H1 : $\beta \neq 0$

V ozadju je tudi t-preizkus.

Zakaj je ta preizkus imenovan preizkus neodvisnosti?

V ničelno domnevo smo zapisali da je beta enak nič. Če bi ta vrednost res bila enaka nič, pomeni, da kakorkoli spreminjamo x bo to vedno pomnoženo z nič in se v y nikoli nič ne bo zgodilo. Y je vedno konstanta. To je stanje ko odvisnosti med dvema pojavoma praktično ni.

Čim pa uspemo z našo oceno zavrniti ničelno domnevo (to je naša težnja), smo s tem dokazali, da pojav x vpliva na pojav y. Na podlagi ocene smo dokazali odnos za populacijo.

Preizkušanje domnev o regresijskih koeficientih - 2

$$H_0 : \beta = \beta_0$$

$$t = \frac{b - \beta_0}{se(b)} \qquad t_{(\alpha, m=n-2)}$$

$$H_0 : \alpha = 0$$

$$t = \frac{a}{se(a)} \qquad t_{(\alpha, m=n-2)}$$

Imamo neko oceno na podlagi vzorca in s pomočjo preizkušanja domnev skušamo ugotoviti ali je ta ocena tudi značilni pokazatelj situacije v populaciji.

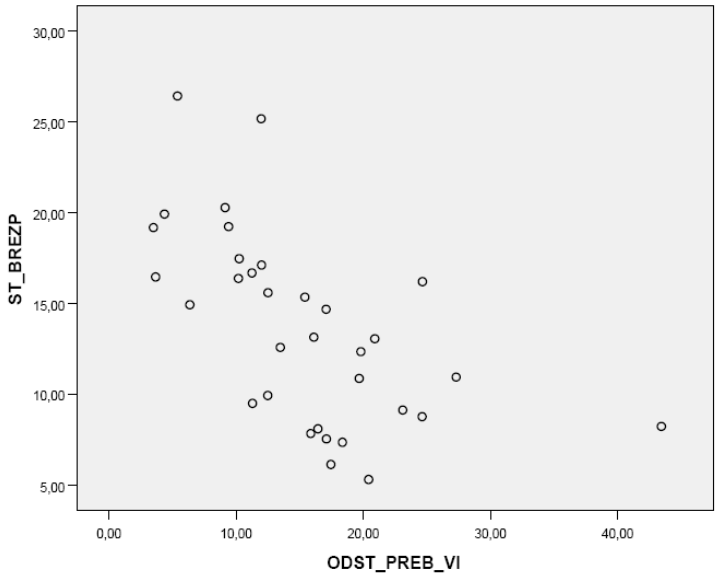
Ocena linearnega regresijskega modela - primer

- Analizirajte vpliv visoke izobrazbe na pojav brezposelnosti, na primeru slovenskih občin. (*Podatki v priloženi datoteki*):
  - Analizirajte razsevni diagram.
  - Napišite enačbo regresijske premice in pojasnite regresijske koeficiente.
  - Kolikšen delež variabilnosti stopnje brezposelnosti je pojasnjen z linearno odvisnostjo od % aktivnega prebivalstva z visoko izobrazbo?

Izobraženost prebivalstva ima neke ugodne vplive na nižjo brezposelnost. **Bolj izobraženo prebivalstvo naj bi bilo na trgu bolj fleksibilno**. V primeru izgube delovnega mesta je potrebna neka prekvalifikacija.

Pojav, ko imamo regresijsko analizo zgolj med dvema pojavoma, lahko zelo preprosto analiziramo tudi z excel-om (namesti SPSS-a). Imamo funkcije INTERCEPT (x;y) in dobimo oceno za a oz. α, SLOPE (x;y) in dobimo podatek b oz. β in COREL (x;y) in ocena korelacijski faktor.

Odvisnost najprej pogledamo na podlagi razsevnega diagrama. Ali sploh odvisnost obstaja, v kakšni meri se pojavlja ali je močna ali šibka. Potem skušamo oceniti enačbo regresijske premice.



Na y os nanašamo stopnjo brezposelnosti po posameznih občinah, ker pričakujemo da je ta odvisna. Na x osi pa prikazujemo odstotek prebivalstva z visokošolsko izobrazbo. **Odnos na diagramu je negativen**. Pomeni če se odstotek prebivalstva z visokošolsko izobrazbo

povečuje je stopnja brezposelnosti v povprečju nižja. **Zveza ni linearna**. Točke so razsejane v obliki neke potenčne ali eksponentne krivulje. *Zaradi poenostavitve rečemo kot da ta zveza je linearna*. Če bi na te točke nasilno želeli postaviti premico, jo bodo vlekle oddaljenejše točke navzgor. To je tudi ena od slabosti regresijske analize. Občutljiva je namreč na ekstremne vrednosti. **Povezanost na diagramu je šibka**. Točke so kar razsejane. Ni pričakovati tako močne odvisnosti.

$y' = \alpha + \beta x$

y" = a + bx

ST _ BREZP" = a + b*ODST _ PREB _VI

(da je bralcu oz. tistemu, ki je raziskava namenjena bolj jasno razvidno, se namesto y in x piše kar imena spremenljivk)

Izpis, ki ga dobimo iz SPSS-a je sestavljen iz štirih tabel. Prva tabela je informativne narave. Izpisano je samo katere spremenljivke smo vključili v regresijski model.

Potem sledijo trije izpisi. Včasih so potrebni vsi trije. V našem konkretnem primeru, ko imamo zgolj bivariantni regresijski model, sta zanimivi le **Model Summary** in **Coefficients** (prva in tretja). Tabela ANOVA bo zanimiva v primeru ko bomo prišli do multivariantnega regresijskega modela. SPPS jo vsakič izpiše ker ima univerzalni modul za oceno regresijskega modela.

Tabela Coefficients po kateri sta razvidni obe oceni regresijskih parametrov

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	19,707	1,639		12,021	,000
	ODST_PREB_VI	-,393	,095	-,595	-4,127	,000

a. Dependent Variable: ST_BREZP

a

b

Najprej imamo vrstico v kateri se nahaja ocena konstante. Njena vrednost je 19,707. To je naš **a**. Nato je zapisana spremenljivka. Ta ki je bila vključena kot neodvisna (odstotek prebivalcev z visokošolsko izobrazbo). Zraven nje je koeficient, ki je dejansko vrednost **b** in je - 0,393.

a=19,707 b= -0,393

Na podlagi teh dveh lahko zapišemo enačbo.

ST _ BREZP" = a + b*ODST _ PREB _VI

ST _ BREZP" = 19,707 - 0,393* ODST _ PREB _VI

Z matematičnega vidika nam konstanta predstavlja vrednost odvisne spremenljivke, če bi neodvisna spremenljivka zavzela vrednost nič (0). V tem primeru bi y bil enak samo oceni parametra. V dejanskih primerih v družboslovju je velikokrat nerealno pričakovati, da spremenljivka, ki jo analiziramo, zavzame vrednost nič. (v našem primeru bi to pomenilo, da bi imeli neko občino v kateri ne bi bilo niti enega prebivalca z visokošolsko izobrazbo. Če pogledamo vse podatke po naših občinah, takšne ni).

Naslednji primer bi bil ko smo imeli odnos količina/cena.

količina = a + b*cena nerealno je pričakovati da bi neka prodajalna prodajala blago po ceni nič.

Ocene za **a** skoraj nikoli ne interpretiramo (razen v izjemnih primerih).

V vsakem primeru pa interpretiramo oceno za **b**, ker je to osrednje kar nas zanima.

V splošnem velja da β pomeni delta y (povprečno spremembo y), če se x poveča za eno enoto.

$$\beta \text{ pomeni } \overline{\Delta y} \quad \text{če je } \Delta x = 1$$

Odstotek prebivalcev z visokošolsko izobrazbo smo merili z spremenljivko, ki je bila v osnovi izražena v odstotkih. Tudi stopnjo brezposelnosti smo merili s spremenljivko izraženo v odstotkih. Kadar je neka spremenljivka izražena v odstotkih se njene absolutne spremembe izražajo v odstotnih točkah. (če imamo recimo občino KP (x=10%) in občino PO (x=11%). Pomeni da je v občini Postojna odstotek prebivalcev z visokošolsko izobrazbo za eno odstotno točko večji kot v Kopru. Ali pa rečemo da je v občini Postojna odstotek prebivalcev z visokošolsko izobrazbo 10 odstotkov večji kot v Kopru. Pozor! Odstotek je vedno relativna mera.)

PO/KP = 11/10 = 1,1*100 = 110 v Postojni je odstotek za 10% višji kot v Kopru

PO-KP = 11% - 10% = 1% ena odstotna točka

Pri tem regresijskem modelu je to pomembno. Ker govorimo o linearnem regresijskem modelu so vse spremembe tako x kot y absolutne. Če imamo neko spremenljivko, ki je izražena v odstotkih, se njena sprememba interpretira v odstotnih točkah.

To bi pomenilo: ***Na podlagi ocene regresijskega koeficienta, ki znaša – 0,393 ocenjujemo, da se stopnja brezposelnosti v povprečju zmanjša za približno 0,4 odstotne točke, če se odstotek prebivalcev, z visokošolsko izobrazbo, poveča za eno odstotno točko.***

Sledi tabela **Model Summary** kjer sta pomembna dva parametra. Prvi je R (korelacijski koeficient), drugi pa R kvadrat.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,595[a]	,355	,334	4,32383

a. Predictors: (Constant), ODST_PREB_VI

$r = 0,595$ korelacijski koeficient

$r^2 = 0,355$

Specifika!

Pripisati je treba vrednosti minus (-) predznak. Modul za izračun je univerzalen, ki ga SPSS uporablja ne glede na to ali imamo regresijsko analizo bivariantnega tipa (samo y in x, torej samo dve spremenljivki), ali pa multivariantno regresijsko analizo, ko je poleg enega x še vrsto ostalih x-ov.

Če imamo regresijski model tipa: y'' = a + b1x1 + b2x2 + b3x3

Neka spremenljivka lahko vpliva plus, nekatera minus, nekatera spet plus. V teh primerih je nesmiselno, če gledamo globalno, govoriti o tem kakšna je smer. Za to SPSS vedno ta R izračuna samo absolutno. **Ko imamo bivariantno regresijsko analizo moramo sami uskladiti predznak**. Če pa je spremenljivk več, je to že kar pravilna vrednost.

približna lestvica korelacij:

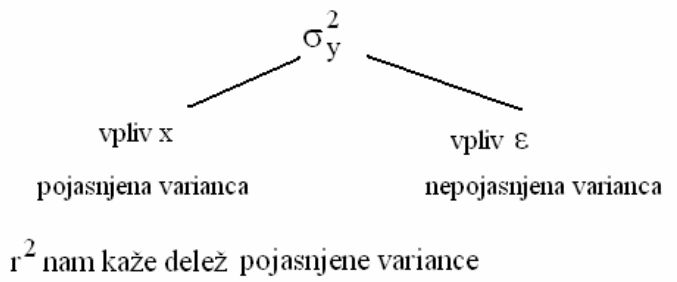
- nekje do 0,6 šibke korelacije
- 0,6 – 0,75 srednje močne
- 0,75 – 0,9 močne
- nad 0,9 zelo močne

Različni avtorji to različno interpretirajo. **To je odvisno tudi od tega na katerem področju regresijsko analizo uporabljamo**. Na področju družboslovja, kjer so pojavi nekoliko bolj ohlapno povezani, smo tudi manj kritični. Že pri 0,75 bi morda rekli, da gre že za zelo močno povezanost. Obratno pa na področju naravoslovja. Tu naj bi bili pojavi bolj eksaktno povezani med sabo. Smo nekoliko bolj kritični. Tam bi morda celo zahtevali namesto 0,9 vrednost 0,95 in rekli, da gre za zelo močno povezanost.

V našem primeru kazalnik pokaže da je to šibka povezanost.

INTERPRETACIJA – ***Na podlagi ocene korelacijskega koeficienta, ki znaša – 0,595 ocenjujemo, da je odvisnost stopnje brezposelnosti, od odstotka prebivalstva z visokošolsko izobrazbo, negativna in šibka***.

Drugi **parameter** je r^2, ki **kaže delež pojasnjene variance**.



r^2 nam kaže delež pojasnjene variance

Razlaga: Imamo variabilnost spremenljivke y. V našem primeru je to variabilnost stopenj brezposelnosti med posameznimi občinami. Skušamo ugotoviti kolikšen delež te variabilnosti povzroča vpliv x in kolikšen del vpliv slučajnih dejavnikov. Pravimo da celotno varianco razstavimo na pojasnjen in nepojasnjen del. Tisti del, ki jo povzroča x spremenljivka je pojasnjena varianca, tisti, ki jo povzročajo slučajnostni odkloni je pa nepojasnjena varianca. Tisti model v katerem je delež pojasnjene variance večji je tudi boljši model (bolj natančne so ocene). V tistem modelu kjer prevladuje del nepojasnjene variance je pa slabši model, ker je večji del spreminjanja y nepoznan zakaj se spreminja, ker ga povzročajo neznani dejavniki.

INTERPRETACIJA – ***Na podlagi ocene determinacijskega koeficienta, ki znaša 0,355 ocenjujemo, da lahko približno 35% variabilnosti stopenj brezposelnosti, med posameznimi občinami, pojasnimo z linearnim vplivom odstotka prebivalcev z visokošolsko izobrazbo. Preostalih približno 65% variabilnosti povzročajo neznani oz. slučajni dejavniki***.

Samo 1/3 razloga, zakaj je v občini manjša ali višja brezposelnost, lahko pripišemo odstotku prebivalstva z visokošolsko izobrazbo.

Kar 2/3 pa z vidika tega modela ne znamo pojasniti.

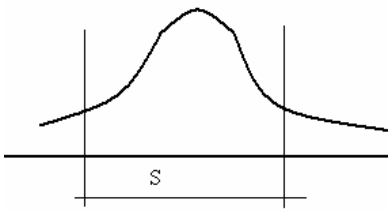
ničelna domneva Ho: $\beta \geq 0$

alternativna domneva H_1: $\beta < 0$

Preizkus neodvisnosti se v osnovi glasi (dvostranski preizkus)

Ho: $\beta = 0$

H_1: $\beta \neq 0$



Ničelno domnevo zavrnemo ali kadar bo ocena močno pozitivna ali kadar bo ocena močno negativna. V tem trenutku je pomembno samo da se razlikuje.

Če želimo konkretno ugotoviti (ne samo ali izobraženost prebivalstva vpliva na stopnjo brezposelnosti ali ne) ali višja izobraženost ugodno vpliva na nižjo brezposelnost, potem je naše vprašanje ali lahko trdimo da je beta negativen ($\beta < 0$). Na podlagi ekonomske teorije pričakujemo, da naj bi pri višji stopnji izobrazbe morala biti stopnja izobrazbe nižja. Že vnaprej pričakujemo, da naj bi višja izobrazba vplivala na nižjo brezposelnost. Pričakujemo da bo beta negativen. To si zadamo kot cilj. Z našo analizo bomo skušali potrditi ali res ta ekonomska teorija, na primeru slovenskih občin, drži. Dobili smo oceno. Sama ocena po sebi tega še ne more kazati ali je to res ali ne. Zato je vsakič potreben nek preizkus. V tem primeru je ustrezen t- preizkus.

t je b deljeno s standardna napaka za b

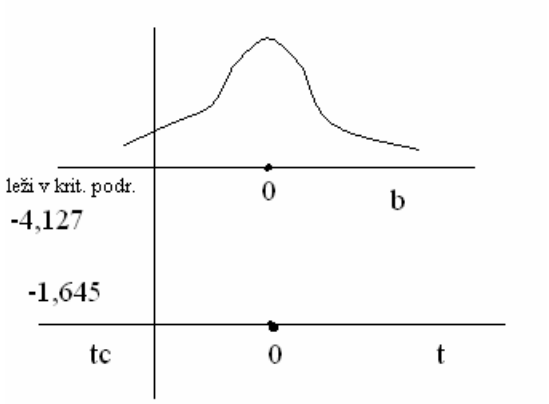
$$t = \frac{b}{se(b)} = \frac{-0{,}393}{0{,}095} = -4{,}127$$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	19,707	1,639		12,021	,000
	ODST_PREB_VI	-,393	,095	-,595	-4,127	,000

a. Dependent Variable: ST_BREZP

Kadarkoli delamo preizkuse domnev je priporočljiva skica vzorčne porazdelitve.



α=0,05 m = n – 2 = 33 – 2.= 31

Porazdelitev koeficienta **b** je približno kot na skici. Na sredini je vrednost 0. Vrednost v ničelni domnevi. Ker smo v alternativno zapisali tako, da skušamo dokazati, da je beta negativna, je ločnica na levi strani. Če bo naša ocena dovolj negativna bo to dokaz, da tudi v populaciji velja taka ocena. Spodaj si vzporedno lahko predstavljamo t porazdelitev. Vrednost nič, na levi kritična vrednost tc. V regresijskem modelu ta tc dobimo tako, da pogledamo pri stopnji tveganja α=0,05, stopinje prostosti so pa nekoliko drugačne (m = n – 2). V našem primeru imamo 33 občin, stopinje prostosti so (m = 33 – 2 = 31). Če bi rešitev s temi podatki pogledali v tablicah (meje prostosti so ∞) je ta kritična vrednost -1,645. V tablicah so samo pozitivne vrednosti, ker je t-statistika simetrična. Mi moramo pa, glede na lego kritičnega področja, določiti predznak. Če je kritično področje na levi imamo minus predznak, če je na desni, pišemo plus predznak. V našem primeru je kritični t enak -1,645 izračunani t je pa enak -4,127 kar pomeni, da leži v kritičnem področju. To pomeni da ničelno domnevo lahko zavrnemo in sprejmemo alternativno. Potrdimo tisto kar so bila naša pričakovanja.

Na podlagi vzorca slovenskih občin lahko pri stopnji tveganja α=0,05 zavrnemo ničelno domnevo in sprejmemo sklep, da je regresijski koeficient β negativen, kar pomeni, da višji odstotek prebivalcev, z visokošolsko izobrazbo, ugodno vpliva na nižjo stopnjo brezposelnosti.

To je primer preproste regresijske analize na dveh spremenljivkah. Naša naloga je bila:

- oceniti enačbo regresijske premice
- komentirati
- oceniti korelacijski koeficient, ki nam kaže jakost odvisnosti in smer
- oceniti r^2, ki nam kaže delež pojasnjene variance
- in kot zadnje preizkusiti (tabela coefficients) ali na podlagi te ocene res lahko potrdimo naša pričakovanja. Zaradi ekonomske teorije smo pričakovali, da naj bi beta bila negativna, zato smo postavili kritično področje na negativno območje. Ugotovili koliko znaša kritična vrednost, koliko znaša izračunana, videli smo, da je ocena v kritičnem področju, kar je pomenilo zavrnitev ničelne domneve in sprejem alternativne. S tem smo potrdili naša pričakovanja na podlagi ekonomske teorije.

ODGOVOR NA VPRAŠANJE B.F.:

Z regresijskim koeficientom ugotovimo, strogo gledano, samo ali ni nikakršne odvisnosti (je nič) ali neka odvisnost je. To še ni pokazatelj, da je to tudi močna odvisnost. Ko pa ugotovimo odvisnost, je ta lahko šibka, srednja ali zelo močna. Mi smo ugotovili, da odvisnost je ampak razmeroma šibka.

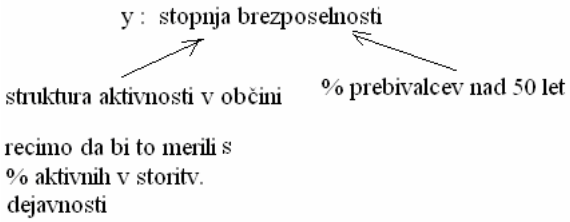
2. del

Multipla odvisnost - 1

- *Bivariatni regresijski model:* predpostavimo, da na spreminjanje vrednosti odvisne spremenljivke vpliva ena sama neodvisna spremenljivka
- V dejanskih razmerah je za proučevani pojav običajno odločilen vpliv več dejavnikov -> z eno samo neodvisno spremenljivko lahko pojasnimo le majhen delež skupne variabilnosti odvisne spremenljivke (nizka vrednost R^2)

$r^2 = 0{,}355$ Da je model zadovoljiv naj bi bil vsaj vrednosti 0,6.

Kako bi model lahko izboljšali? Kaj vse še lahko vpliva na stopnjo brezposelnosti?



Multipla odvisnost - 2

- Z vključitvijo dodatnih spremenljivk bi lahko povečali delež pojasnjene variance (R^2).

  Primer: Stopnja brezposelnosti ni odvisna zgolj od izobraženosti prebivalstva. Nanjo vplivajo še drugi dejavniki: splošne gospodarske razmere, struktura dejavnosti...

Model skušamo razširiti tako da bi rekli: y ni več samo a + bx ampak:

y'' = a + b1x1 + b2x2 + b3x3 + ·····

(x2 odstotek zaposlenih v storitveni dejavnosti, x3 je % prebivalcev nad 50 let, …itd.)

Pridemo do modela **MULTIPLE REGRESIJSKE ANALIZE**, katerega cilj je povečati r^2. Z eno samo spremenljivko ne zmoremo dovolj dobro pojasniti našega proučevanega pojava. V model skušamo vključiti čim več dejavnikov, ki bi lahko vplivali na naš odvisni pojav. Če smo "zadeli" pomembne dejavnike se bo s tem moral r kvadrat povečevati. Če bi model napisali z simboliko, vezano na populacijske podatke. Uporabljali bi α in β

Model linearne multiple regresije

$$y = \alpha + \beta_1 x_1 + ... + \beta_k x_k + \varepsilon$$

kjer pomeni:

- α: regresijska konstanta
- β_j: parcialni regresijski koeficient
- ε: slučajnostni odkloni

- Matrični zapis:

$$y = x\beta + \varepsilon$$

Zakaj pa, če vse dejavnike, ki jih poznam, vključim v model in še vedno obstaja nek ε. **Najmanj kar se v tem epsilonu skriva so slučajni dejavniki**, ki jih ne moremo nikoli predvideti. Teh, ki jih ne moremo predvideti tudi nikoli ne moremo vključiti.

V praksi se skušamo omejevati, da modeli nimajo deset ali več neodvisnih spremenljivk ampak je v praksi velikokrat število neodvisnih spremenljivk 5,7 ali manj.

Kar se spremeni je pri interpretaciji regresijskega koeficienta. Regresijski koeficient smo razlagali kot povprečno spremembo y, če se x poveča za eno enoto. Izraz parcialni uporabljamo, ker se običajno koncentriramo na enega, ostale pa smatramo "ceteris paribus" (kot da so nespremenjeni).

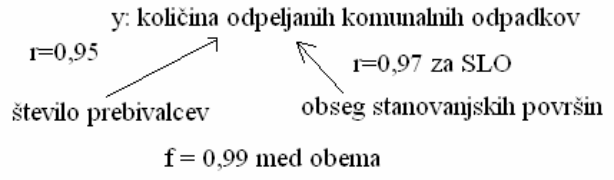
Predpostavke modela - 1

- Neodvisne spremenljivke x_j so fiksne, merjene so brez merskih napak.
- Med dvema poljubnima neodvisnima spremenljivkama lahko obstaja značilna linearna odvisnost, ki pa ne sme biti funkcijska (multikolinearnost).
- Odvisna spremenljivka je slučajnostna, normalno porazdeljena spremenljivka. Njena pričakovana vrednost je

$$E(y/x_1, x_2, ...x_k) = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k = y'$$

Med nobenim parom ne sme biti funkcijske zveze.

Primer:



Količina je zelo močno povezana z številom prebivalcev v neki občini. Povezana je tudi z obsegom stanovanjskih površin. Večji kot je ta obseg, več odpadkov se ustvari En dejavnik zelo močno vpliva na celotno količino, drugi dejavnik zelo močno vpliva. Ko pa damo oba hkrati ugotovimo da je razlika enaka nič. Če gremo preveriti je med njima f=0,99. Pomeni, da, če vemo koliko ima občina stanovalcev je to isti podatek kot kakšna je kvadratura stanovanj. Če damo dva ista podatka hkrati v model nismo naredili čisto nič novega.

Preizkušanje domnev o regresijskih koeficientih

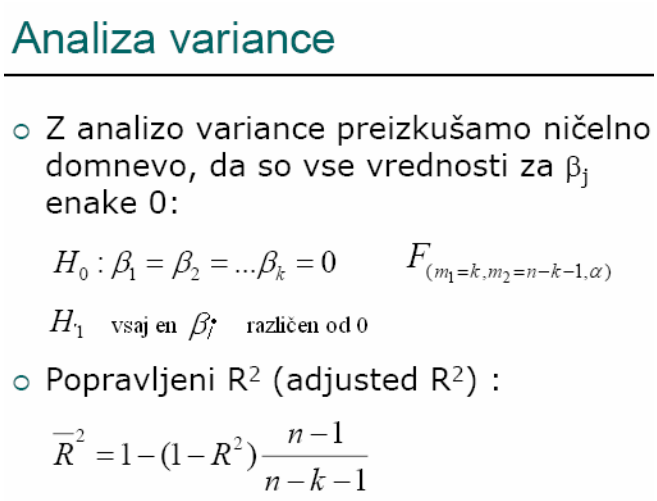
1) $H_0: \beta_j = 0$ ali $H_0: \beta_j = \beta_{j0}$

t-preizkus: $t = \frac{b_j - \beta_{j0}}{se(b_j)}; t_{\alpha, m=n-k-1}$

2) $H_0: \beta_i = \beta_j$

t-preizkus: $t = \frac{b_i - b_j}{se(b_i) - se(b_j)}; t_{\alpha, m=n-k-1}$

Izvedli bomo hkrati tudi preizkus, ki mu tudi tu pravimo **Analiza variance** (podobno kot pri povprečjih). Ta **preizkus preizkuša, da so vsi hkrati enaki nič**. Kaj bi to pomenilo? Če bi res imeli situacijo v kateri so vsi koeficienti hkrati enaki nič to pomeni, da imamo povsem nesmiseln regresijski model. Katerokoli spremenljivko x1 ali x2 ali x3 spremenimo so vse te spremembe pomnožene z nič. To bi bil skrajno slab model.



Pri izračunavanju Determinacijskega koeficienta bomo morali upoštevati popravek. Izračun r kvadrata je sam po sebi inflacijski. V model lahko vključujete spremenljivke do onemoglosti, tudi take, ki so povsem nesmiselne. r kvadrat se le malo poveča. Na ta način bi umetno napihnili r kvadrat. Da ta r kvadrat malo očistimo te napihnjenosti, se ga korigira in zato v primeru multiple regresijske analize **na izpisu ne upoštevamo r kvadrata ampak popravljeni r kvadrat**.

1. ker imamo več neodvisnih spremenljivk pomeni, da bomo njihove odzive razlagali parcialno, sicer bi bila zadeva preveč kompleksna
2. ker imamo več spremenljivk moramo biti pozorni na to koliko so te med seboj povezane da ne pride do multikolinearnosti
3. ker imamo več spremenljivk moramo izvesti preizkus za vse spremenljivke hkrati potem pa še vsako posebej
4. ker več spremenljivk napihuje r kvadrat ga je potrebno očistiti inflacijskega vpliva in za to uporabljamo za interpretacijo popravljeni r kvadrat

Multipla regresija - primer

- V eni izmed revij so se odločili za raziskavo trga nepremičnin na območju ljubljanske mestne občine. Za 30 stanovanj, ki so naprodaj so zbrali podatke o njihovi ceni (CENA; v 10^6 SIT), površini (POVM2; v kvadratnih metrih), starosti (STAR;v letih) in oddaljenosti od centra mesta (ODD;v kilometrih) (*podatki v priloženi datoteki*).
  - Izračunajte korelacijsko matriko in preverite domnevo o neodvisnosti med površino stanovanja in oddaljenostjo od centra mesta.
  - Ocenite regresijsko funkcijo, ki opisuje spreminjanje cene stanovanja v odvisnost od navedenih dejavnikov (površina, starost...). Pojasnite postopek ocenjevanja.
  - Izračunajte mere korelacije in jih obrazložite.
  - Preverite ničelno domnevo, da so vsi parcialni regresijski koeficienti hkrati enaki nič.

Imamo rabljena stanovanja. Pri vsakem stanovanju imamo znane naslednje podatke:

- CENA - ceno po kateri naj bi se stanovanje prodalo
- POVM2 - površino v kvadratnih metrih
- STAR - starost v letih
- ODD - lokacijo v km v smislu koliko je stanovanje oddaljeno od centra mesta
- 30 stanovanj

Najprej pogledamo **korelacijsko matriko** in preverimo domnevo o neodvisnosti med površino in oddaljenostjo od centra mesta. To je v bistvu matrika bivariantnih korelacijskih koeficientov. Za vse spremenljivke, ki jih vključujete v model, naj bo to regresijski model, faktorski model nas zanimajo odnosi med spremenljivkami. Za vsako dvojico spremenljivk imamo ugotovljeno korelacijo. Taka matrika je vedno kvadratna (toliko kot je stolpcev toliko je vrstic).

a) Korelacijska matrika

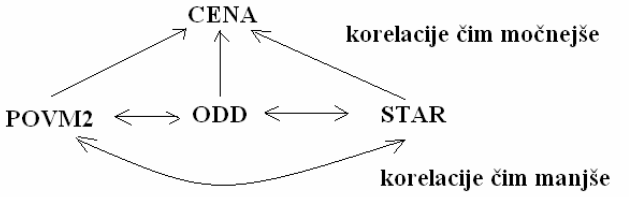
diagonala

Correlations

		CENA	POVM2	STAR	ODD
Pearson Correlation	CENA	1,000	,925	,196	-,286
	POVM2	,925	1,000	,308	-,077
	STAR	,196	,308	1,000	-,006
	ODD	-,286	-,077	-,006	1,000
Sig. (1-tailed)	CENA	.	,000	,150	,063
	POVM2	,000	.	,049	,343
	STAR	,150	,049	.	,487
	ODD	,063	,343	,487	.
N	CENA	30	30	30	30
	POVM2	30	30	30	30
	STAR	30	30	30	30
	ODD	30	30	30	30

Na **diagonali** so vrednosti ena (1) ker je vsaka matrika sama s sabo funkcijsko povezana. Ker je korelacija med CENO in POVM2 enaka korelaciji med POVM2 in CENO, je ta matrika vedno simetrična. Vse vrednosti nad diagonalo so simetrične tistim pod diagonalo. Zadostuje da se osredotočimo samo na zgornji ali spodnji trikotnik matrike.

CENA je odvisna spremenljivka.



Če naj bi POVM2, ODD in STAR bili pomembni dejavniki cene stanovanja, želimo da so korelacije v prvi vrstici čim močnejše. Sledijo ostale neodvisne dvojice, za katere želimo, da so korelacije čim manjše, da ne bo problema multikolinearnosti. Na podlagi te korelacijske matrike preverimo domnevo, da sta POVM2 in ODD od centra mesta neodvisni.

Ho : $\rho_{POVM2, ODD} = 0$

H1 : $\rho_{POVM2, ODD} \neq 0$

Če naj bi POVM2 in ODD bili neodvisni bi moralo veljati, da je korelacijski koeficient med to dvojico spremenljivk enak nič (0).

Čim ničelno domnevo zavrnemo in sprejmemo alternativno, je to dokaz, da sta ti dve spremenljivki med sabo povezani. To še ne pomeni, da bo prišlo do problema multikolinearnosti. Če bo ta ocena r po absolutni vrednosti do 0,3 ali 0,4, še ne bo večjih težav. Čim pa naraste nekje nad 0,6 ali 0,7 (odvisno od primera do primera), postane lahko problem. Ali POVM2 in ODD sta ali nista povezana? Odgovor je v korelacijski matriki.

V njej imamo oceno: **$r_{POVM2,ODD} = -0,077$** → je skorajda enak 0

Stopnja značilnosti : **sig = 0,343**

pravilo: sig ≤ 0,05 → Ho zavrnemo

sig > 0,05 → Ho ne moremo zavrniti

Če Ho ne moremo zavrniti je to z našega vidika zelo dobro. Če ne moremo negirati, da je korelacija med njima nič, kar pomeni da je zelo velika verjetnost da je enaka nič. Ker sta to dve neodvisni spremenljivki, je kvečjemu zaželeno, da sta čim manj povezani. To je ugodni izid.

Za vse dvojice neodvisnih spremenljivk je zaželeno, da so stopnje značilnosti čim višje.

Če pogledamo poleg teh dveh spremenljivk še za ostale. Za POVM2 in ODD ni mogoče potrditi odvisnosti, STAR in ODD tudi, ker je stopnja značilnosti precej visoka (0,487), torej sta tudi neodvisni. Edina stopnja značilnosti, ki je na meji kaže, da sta STAR in POVM2 rahlo povezani.

Starejša ko so stanovanja, večja so, mlajša stanovanja, manjša so. Odraz nekega trenda na tržišču.

Izkustveno gledano pribl. 0,3 naj ne bi predstavljalo težav. To je bil pred-korak preden se sploh lotimo regresijske analize. Vedno si predstavljamo korelacijsko matriko. Pogledamo kakšni so odnosi med vsemi spremenljivkami. Med odvisno in neodvisnimi je zaželeno, da so korelacije čim močnejše. Med neodvisnimi želimo, da so korelacije čim nižje.

Ocena modela:

y'' = a + b1x1 + b2x2 +b3x3

cena '' = a + b1 * povm2 + b2 *star + b3 *odd

Pogledamo najprej tabelo Coefficients

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3,207	1,078		2,976	,006
	POVM2	,205	,014	,937	15,044	,000
	STAR	-,049	,033	-,094	-1,512	,143
	ODD	-,164	,045	-,214	-3,616	,001

a. Dependent Variable: CENA

Po enaki logiki imamo vrstico Constant, potem pa za vsako spremenljivko posebej oceno njenega parcialnega regresijskega koeficienta.

cena '' = 3,207 + 0,205 * povm2 -0,49 *star – 0,164 *odd

Preveriti moramo kakšne imamo enote (površina v kvadratnih metrih, ceno v mio SIT).

Na podlagi ocen parcialnega regresijskega koeficienta b1, ki znaša 0,205 ocenjujemo, da se cena stanovanja v povprečju poveča za 205.000, če je površina stanovanja za 1 kvadr. meter, večja, ob nespremenjeni starosti in oddaljenost stanovanja od centra.

//Konstanta 3,207 bi pomenila ceno stanovanja, ki nima nič kvadratnih metrov, ki nima starosti niti eno leto in ki je locirano v centru. (stanovanje brez kvadratnih metrov si je težko predstavljati, da bi bilo staro nič let in locirano v centru je še razumljivo)//.

Pogledamo še r in r^2

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,954[a]	,909	,899	1,75830

a. Predictors: (Constant), ODD, STAR, POVM2

Ocena **r = 0,954**, **popravek na kvadrat = 0,899**

To je v bistvu zelo dobro. Korelacija je zelo močna 0,95. V 90% primerov nam dejavniki znajo pojasniti ceno. Razlika 10% so res slučajni dejavniki.

Izbrani dejavniki so očitno res trije pomembni dejavniki cene stanovanj.

Na podlagi ocene korelacijskega koeficienta (multiple korelacije), ki znaša 0,954 ocenjujemo, da je odvisnost cene stanovanja od njegove površine, starosti in oddaljenosti od centra zelo močno povezana.

Kadar imamo koeficient multiple korelacije je vedno samo pozitivna vrednost, ker ne more biti drugače. Rekli smo samo močna povezanost ne pa pozitivna. Površina je povezana pozitivno, starost negativno, oddaljenost negativno. Vsaka spremenljivka ima drugačno smer vpliva. Zato je nesmiselno, z vidika modela (globalno), govoriti o tem kakšna je smer. **O smeri pri korelacijskem koeficientu govorimo samo takrat kadar je korelacijski koeficient med samo dvema spremenljivkama**. Takrat ugotavljamo kakšna je smer. Čim je korelacijski koeficient med večjim številom spremenljivk potem smer ni več relevantna.

Determinacijski koeficient

Rečemo: ***Na podlagi ocene popravljenega determinacijskega koeficienta, ki znaša 0,899 ocenjujemo, da je lahko približno 90% razlik v cenah med stanovanji, mogoče pojasniti z linearnim vplivom površine, starosti in oddaljenosti. Preostalih 10% variabilnosti povzročajo neznani oz. slučajni dejavniki.***

Preizkus modela kot celote (postopku pravimo **Analiza varianc**)

Ho : $\beta 1 = \beta 2 = \beta 3 = 0$ F-test

H_1 : vsaj_en_ $\beta 1 \neq 0$

To metodo uporabimo za preizkus domneve regresijskega modela. Gre za to, da so koeficienti enaki 0. Preizkušamo na podlagi F-testa, ki je v tabeli ANOVA (Analysis of Variance).

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	805,587	3	268,529	86,857	,000[a]
	Residual	80,382	26	3,092		
	Total	885,969	29			

a. Predictors: (Constant), ODD, STAR, POVM2

b. Dependent Variable: CENA

(F α, m1, m2, m1 = k, m2 = n – k - 1) m1=3, m2=30-3-1=26; če bi pogledali v tablicah pri α=0,05 in m1=3 in m2=26, dobimo Fc (kritična vrednost).

F = izračunani

če je $F \geq Fc \rightarrow$ Ho zavrnemo

$F < Fc \rightarrow$ Ho ne zavrnemo

Tablice so zelo nerodne (na osmih straneh). Da se izognemo nepotrebnemu iskanju po tablicah, zadostuje stopnja značilnosti:

sig > 0,05 → Ho ne moremo zavrniti

sig ≤ 0,05 → Ho zavrnemo

V našem primeru ne samo da je manjša od 0,05 ampak je praktično tako nizka da na treh decimalkah sploh ni vrednosti. (Tega nikoli ne enačimo z 0; $0{,}000 \neq 0$!!; pomeni da je lahko nekaj malega na kasnejši decimalki; vrednost samo interpretiramo kot zanemarljivo nizko).

V našem primeru vidimo, da na podlagi preizkusa vrednosti regresijskih koeficientov pri zanemarljivi stopnji tveganja zavrnemo ničelno domnevo in sprejmemo sklep:

da je vsaj en parcialni regresijski koeficient različen od 0 (vsaj ena izmed predpostavljenih dejavnikov ima vpliv na ceno stanovanja. Pomeni da smo na dobri poti. Kateri dejavnik bomo pogledali sedaj).

Ta rezultat je bil pričakovan. (če imamo 90% pojasnjene variance, mora biti nek vpliv, zato je ta sklep pričakovan).

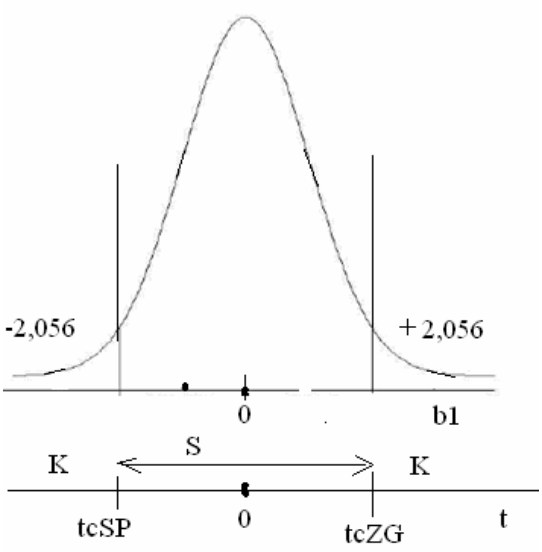
Ugotovili smo, da je vsaj ena spremenljivka pomembna. Pogledamo spet v **tabelo Coefficients**, kjer imamo za vsako od neodvisnih spremenljivk t preizkus in stopnjo značilnosti. V ozadju so dejansko preizkusi tipa βi=0.

Ho : $\beta 1 = 0$

$H_1 : \beta 1 \neq 0$

V korelacijski matriki je na nekaterih mestih pisalo sig. (1-tailed). Zadnjič, ko smo imeli t-preizkus za primerjavo dveh aritmetičnih sredin je nekje pisalo sig (2-tailed). V tej tabeli ne piše nič ampak predpostavlja da velja standard. Standard pa je, da se ta preizkus vedno izvaja kot dvostranski preizkus. Kljub temu, da v tej tabeli ni nobenega komentarja, so stopnje značilnosti vedno dvostranskega preizkusa.

Skica:



Ker imamo dvostranski preizkus je tudi kritično področje na levi in desni strani. Vmes je območje sprejema S. Po enakem principu velja za katerikoli koeficient. Vzporedno lahko naredimo skalo za t z zgornjo in spodnjo mejo.

tc moramo ugotoviti pri : α/2=0,025 (ker imamo dvostranski preizkus), stopinje prostosti: m=n – k -1 = 30 – 3 – 1 = 26

tc = ± 2,056 (odčitamo v tablicah; ker imamo na obeh straneh je tc plus minus)

Če je izračunani t večji od tega ali manjši od tega v negativno smer, bo znan, da pade v kritično področje in bomo ničelno domnevo zavrnili.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3,207	1,078		2,976	,006
	POVM2	,205	,014	,937	15,044	,000
	STAR	-,049	,033	-,094	-1,512	,143
	ODD	-,164	,045	-,214	-3,616	,001

a. Dependent Variable: CENA

Za prvo spremenljivko POVM2 je izračunani t=15, kar je daleč v kritičnem področju. Za to spremenljivko lahko ničelno domnevo zavrnemo, sprejmemo alternativno, kar pomeni da **ima površina vpliv na ceno**. Za starost STAR je izračunani t=-1,5. Pade v območje sprejema (označeno s piko na sliki, levo od ničle). Ničelne domneve ni mogoče zavrniti. Pomeni da **starost nima vpliva na ceno**. Oddaljenost ODD ima izračunani t=-3,6. Pade v kritično področje. Ničelno domnevo lahko zavrnemo in sprejmemo alternativno, kar pomeni, da ima **oddaljenost tudi vpliv na ceno**.

Še lažji pristop, ko sploh ne uporabimo tablic. Pogledamo stopnjo značilnosti sig.

Če izvajamo dvostranski preizkus in **nas zanima samo ali spremenljivka ima vpliv ali nima,** so te stopnje značilnosti prave. Ni se potrebno ukvarjati na katero stran pade. Tam kjer je stopnja značilnosti sig=0,05 ali manj, pomeni da ničelne domneve zavrnemo in sprejmemo alternativne. Tam kjer je stopnja značilnosti sig nad 0,05 pomeni, da taka spremenljivka vpliva nima.

Taki primeri, ko ugotovite da neka spremenljivka nima vpliva, nadaljujete z oceno regresijskega modela tako, da tako spremenljivko skušate izločiti. Nesmiselno je da imamo v regresijskem modelu spremenljivke, ki nimajo vpliva.

(Primer: če bi ta model dali nekemu posredniku nepremičnin, da bo lahko izračunal cene stanovanj, bi od njega zahtevali, da mora za vsako stanovanje postaviti tri podatke, čeprav je en podatek povsem nesmiseln).

Če ponovimo **oceno regresijskega modela tokrat brez spremenljivke STAR** (starost).

Model Summary zmanjšanje iz 0,899 na 0,894, neznatno

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,949[a]	,901	,894	1,79966

a. Predictors: (Constant), ODD, POVM2

korelacijski prej 0,95 sedaj 0,94 minimalna sprememba

F statistika še bistveno večja

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	798,522	2	399,261	123,276	,000[a]
	Residual	87,447	27	3,239		
	Total	885,969	29			

a. Predictors: (Constant), ODD, POVM2

b. Dependent Variable: CENA

Z lahkoto trdimo, da je model kot celota v redu. Vsaj en koeficient je različen od 0.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2,796	1,067		2,620	,014
	POVM2	,199	,013	,908	14,973	,000
	ODD	-,165	,046	-,216	-3,561	,001

a. Dependent Variable: CENA

V tabeli Coefficients je najprej treba preveriti, če so ostale vse ocene značilne in lahko vidimo da so. Površina ima še vedno velik t, kar pomeni da vpliva. Oddaljenost ima tudi manjši t od

kritičnega, kar pomeni da vpliva. To da je večje stanovanje dražje je razumljivo. Razumljivo je tudi, da bolj kot je stanovanje oddaljeno od centra, nižjo ima ceno.

S tem ko smo dali spremenljivko ven, se je model poenostavil. Če hočemo napovedati ceno, potrebujemo samo dva vhodna podatka in ne tri. Z njima pojasnimo ravno toliko kot prej s tremi.

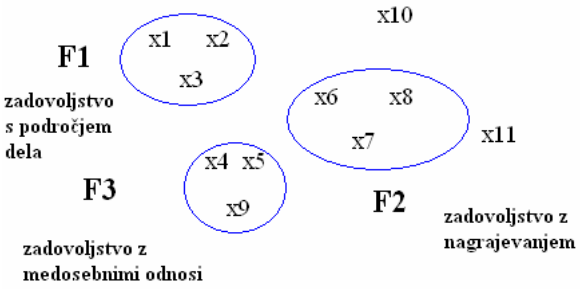
FAKTORSKA ANALIZA

Faktorska analiza je **analiza obojesmernih korelacij**. Namenjena je analizi nemerljivih pojavov. Število spremenljivk je običajno večje.

Regresijska analiza, **temelji na vzročno – posledični zvezi**. Gre za analizo merljivih pojavov.

Pri faktorski analizi je število spremenljivk bistveno večje kot pri regresijski. Dvajset, trideset spremenljivk ni nič kaj posebnega. Pri regresijski je pa to že zelo kompleksen model.

Imamo neko množico spremenljivk med katerimi se težko znajdemo. Faktorska analiza pogleda kako so bile spremenljivke med sabo korelirane. (dobimo korelirane trojice, ni pa nujno da so trojice). Ostanejo pa kakšne spremenljivke, ki niso korelirane z nobenim (slika).



Logika je : če so te tri spremenljivke med sabo korelirane je očitno, da imajo nekaj skupnega. To skupno poimenujemo **SKUPNI FAKTOR** (F1). Če so naslednje tri spremenljivke povezane med sabo je očitno, da imajo tudi nekaj skupnega. To je spet nek skupni faktor, ki vpliva na vse tri na enak način (F2), itn. do faktorja F3. Recimo x10 in x11 nista z nobeno spremenljivko dovolj močno povezana. Nanju torej ne vpliva noben skupni faktor. Prevladuje vpliv specifičnih dejavnikov. Z vidika faktorske analize vse takšne spremenljivke izločimo.

Realni kontekst: x-si so po svoji vsebini indikatorji ker se faktorska analiza uporablja za analizo nemerljivih podatkov. To so pojavi, ki jih ne moremo meriti z nekim metrom (zadovoljstvo, organizacijska klima, odnos do nečesa, itd.). Zato jih največkrat merimo posredno s pomočjo anket. Vprašanja postavimo na tak način, da primer osvetlimo z različnih zornih kotov. V našem primeru bi lahko faktor F1 pomenil zadovoljstvo s področjem dela, F2 recimo zadovoljstvo z nagrajevanjem, F3 pa recimo zadovoljstvo z medosebnimi odnosi.

Pri faktorski analizi se ta poimenovanja faktorjev izvedejo šele na koncu na podlagi rezultata. (Poglej si postopek ocene faktorskega modela v e-učilnici). Gre za primer blagovnih znamk. Na primeru je šolsko prikazano kakšen je postopek analize in kako je na koncu vse poimenovanje lahko tudi subjektivno. Na podlagi nekega rezultata se namreč skušamo vprašati kaj je skupnega tem spremenljivkam in na podlagi tega dobimo ime.

Pri regresijski npr. točno vemo da je y stopnja brezposelnosti, ki jo meri statistični urad in točno vedo kako se to računa.

Pri faktorski teh imen ne vemo vse do konca te analize. Dostikrat so poimenovani subjektivno.

IZPIT: Vprašanja vezana na faktorski model so v smislu, da poznamo kaj posamezni parametri pomenijo. Vedeti moramo kaj pomeni **faktorska utež**, kaj pomeni **komunaliteta** (ocena deleža variance proučevanih spremenljivk pojasnjenega s skupnimi faktorji). Torej da tehnično poznamo model. Dobro si je potrebno pogledati prosojnice in rešitve v e-učilnici. Datum izpita: 29. junij (uradni rok za magistrske študente v vseh treh centrih). Od septembra naprej so roki samo v Kopru. (4 računske naloge; mag. študentje imajo eno nalogo intervalne ocene, (nalogo za dokončati t-preizkus en vzorec ali pa dva vzorca??; /**se slabo razume!**/), eno t-preizkus za dva vzorca, ena regresijska analiza, ena naloga kombinatorika iz verjetnostnega računa (Strašek). Tisto je največkrat trši oreh kot to. Konzultacije **v sredo 23. junij ob 16.30**. Konkretno naloge, ki bodo lahko na izpitu.