



ПРЕДМЕТ
< МЕДИЦИНСКА СТАТИСТИКА >

Област број 4
< РЕГРЕСИЈА И КОРЕЛАЦИЈА И МЕТОДЕ БАЗИРАНЕ НА
ПОРЕТКУ РАНГА >

Област	Наставна јединица	Тематске јединице	Резултат – знања или вештине које студент треба да добије
4	Регресија и корелација. Методе базиране на поретку ранга.	Статистичке технике за истраживање веза између променљивих. Дијаграми растурања. Регресија. Метода најмањих квадрата. Стандардна грешка коефицијента регресије. Корелација. Значај теста и интервал поверења за r . Коришћење коефицијента корелације. Вишеструка регресија. Не-параметарске методе. Mann-Whitney U тест. Wilcoxon-ов тест еквивалентних парова. Spearman-ов коефицијент корелације ранга. Kendall-ov коефицијент корелације ранга. Исправке континуитета. Параметарске или не-параметарске методе?	Упознавање са регресијом и корелацијом, као и са методама базиранима на поретку ранга.

САДРЖАЈ

Регресија и корелација и методе базиране на поретку ранга	2
Статистичке технике за истраживање веза између променљивих	2
Преглед основних начела	3
Чиниоци које треба узети у обзир приликом тумачења коефицијента корелације	3
Претпоставке	4
8. Регресија и корелација	5
8.1 Дијаграми растурања	5
8.2 Регресија	6
8.3 Метода најмањих квадрата	7
8.4 Стандардна грешка коефицијента регресије	10
8.5 Корелација	10
8.6 Значај теста и интервал поверења за r	12
8.7 Коришћење коефицијента корелације	14
8.8 Вишеструка регресија	15
9. Методе базиране на поретку ранга	21
9.1 Не-параметарске методе	21
9.2 Mann-Whitney U тест	21
9.3 Wilcoxon-ов тест упарених (еквивалентних) парова	28
9.4 Spearman-ов коефицијент корелације ранга, ρ	30
9.5 Kendall-ов коефицијент корелације ранга, τ	33
9.6 Исправке континуитета	35
9.7 Параметарске или не-параметарске методе?	36

Област бр. 4

< РЕГРЕСИЈА И КОРЕЛАЦИЈА И МЕТОДЕ БАЗИРАНЕ НА ПОРЕТКУ РАНГА >

Регресија и корелација и методе базиране на поретку ранга

Статистичке технике за истраживање веза између променљивих

У овом делу, размотрићемо неке СПСС-ове технике за истраживање веза између променљивих. Усредсредимо се на откривање и описивање тих веза. Све овде објашњене технике засноване су на корелацији. Корелационе технике често користе истраживачи ангажовани у неексперименталним пројектима. За разлику од експерименталних пројеката, овде се променљиве намерно не модификују нити контролишу, већ се описују у свом природном стању. Тим техникама се може:

- истражити веза између парова променљивих (корелација)
- предвидети вредности једне променљиве на основу вредности друге (биваријантна регресија)
- предвидети вредности зависне променљиве на основу вредности више независних променљивих (вишеструка регресија) и
- идентификовати структура групе повезаних променљивих (факторска анализа).

Овом породицом техника тестирамо моделе и теорије, предвиђамо исходе, и оцењујемо поузданост и ваљаност скала. СПСС има цео низ техника за истраживање веза. Оне се разликују по врсти истраживачког питања на које треба одговорити и врсти доступних података. Овде су обрађене само оне које се најчешће употребљавају.

Корелација се употребљава за описивање јачине и смера везе између две (обично непрекидне) променљиве. Може се употребити и када је једна од тих променљивих дихотомна, тј. може имати само две вредности (нпр. пол: мушкарци/жене). Статистички показатељ који се добија назван је Пирсонов коефицијент линеарне корелације (r). Израчунава се и статистичка значајност показатеља (везе између две променљиве) t .

Парцијална или делимична корелација служи за истраживање везе између две променљиве уз статистичку контролу утицаја треће. То је погодно за ситуације када сумњате да на везу две променљиве можда утиче трећа. Делимична корелација статистички уклања утицај треће променљиве, чиме добијате јаснију слику везе двеју променљивих од интереса.

Вишеструка регресија служи за предвиђање вредности једне зависне непрекидне променљиве помоћу групе независних променљивих. Њоме се може испитати предиктивна моћ скупа променљивих и оценити релативан допринос сваке променљиве појединачно.

Логистичка регресија (користи се уместо вишеструке регресије када је зависна променљива категоријска. Њоме се може испитати предиктивна моћ скупа променљивих и оценити релативан допринос сваке променљиве појединачно.

Факторска анализа служи за истраживање структуре великог броја повезаних променљивих (нпр. ставки од којих се састоји скала). Користи се за свођење великог броја повезаних променљивих на мањи број димензија или компонената, с којим је лакше радити.

Преглед основних начела

Коефицијент корелације (нпр. Пирсонове линеарне корелације) показује смер и јачину линеарне везе између две променљиве. Пирсонов коефицијент корелације (r) има вредност у опсегу од -1 до $+1$. Предзнак показује да ли је корелација позитивна (обе променљиве заједно и опадају и расту) или негативна (једна променљива опада када друга расте и обрнуто). Апсолутна вредност тог коефицијента (када занемаримо његов предзнак) показује јачину везе. Савршена корелација, тј. коефицијент 1 или -1 показује да се вредност једне променљиве може тачно утврдити када знамо вредност друге. С друге стране, корелација 0 показује да између те две променљиве не постоји никаква веза. Познавање вредности једне променљиве нимало не помаже у предвиђању вредности друге.

Везу две променљиве треба испитати и визуелно, цртањем дијаграма растурања. На њему су сви парови резултантних вредности променљивих добијени од субјеката у узорку. Износи прве променљиве се цртају дуж X (хоризонталне) осе, а одговарајући резултати друге дуж Y (вертикалне) осе. Прегледом дијаграма растурања види се и смер везе (позитиван или негативан) и њена јачина. Међутим, када је $r=0$, дијаграм растурања изгледа као облак тачака које не чине никакав геометријски облик.

Чиниоци које треба узети у обзир приликом тумачења коефицијента корелације

Када тумачите резултате корелационе анализе или других техника заснованих на корелацији, морате водити рачуна о неколико ствари.

Нелинеарна веза

Коефицијент корелације (нпр. Пирсонов r) показује линеарну (праволинијску) везу између променљивих. Када су променљиве повезане нелинеарно (нпр. криволинијски), Пирсонов r ће показивати да је веза много слабија него што јесте. Зато увек погледајте дијаграм растурања, нарочито када добијете малу вредност r .

Нетипичне тачке

Нетипичне тачке (чије су вредности знатно мање или веће од вредности осталих тачака у скупу података) много кваре коефицијент корелације, поготово у малим узорцима. У неким околностима, нетипичне тачке чине вредност r много већом него што би она требало да буде, а у другим подбацују у оцени јачине везе. Нетипичне тачке се лако уочавају на дијаграму растурања; то су усамљене тачке, изузеци. Можда су последица грешке приликом уношења података (уписано 11 уместо 1), нетачног одговора испитаника или је у питању стварни одговор необичне особе! Када пронађете нетипичну тачку, проверите да ли се ради о грешци и исправите је ако треба. Иначе, размислите о уклањању или решифровању те необичне вредности да би се смањио њен утицај на коефицијент r .

Ограничен опсег резултата

Коефицијенте корелације морате веома пажљиво тумачити када потичу од малог подопсега стварно могућег распона резултата (нпр. када коефицијент интелигенције, IQ, проучавате на узорку универзитетских студената). Коефицијенти корелације добијени проучавањем ограниченог подопсега резултата често се разликују од оних када је узорком обухваћен пун опсег резултата. Да би се добио тачан и поуздан показатељ јачине везе између две променљиве, свака од њих би требало да има најшири могући опсег резултата. Уколико проучавате екстремне групе (нпр. клијенте с високим нивоом анксиозности), не уопштавајте корелацију на случајеве изван опсега података употребљених у узорку.

Корелација у односу на каузалност

Корелација је показатељ да постоји веза између две променљиве; међутим, она не показује да једна променљива проузрокује ону другу. Корелација између две променљиве (А и Б) може бити последица чињенице да А проузрокује Б, да Б проузрокује А, или да (како би ствари биле још компликованије) трећа променљива (Ц) проузрокује и А и Б. Могућност да трећа променљива проузрокује обе опсервиране променљиве никад не треба губити из вида.

То илуструје чувена прича о јакој корелацији коју је неки истраживач открио између потрошње сладоледа и броја пријављених убистава у Њујорку. Да ли конзумирање сладоледа проузрокује насилно понашање у Њујорку? Не. На обе променљиве (потрошњу сладоледа и стопу криминала) утичу временске прилике. Током летњих врућина расту и потрошња сладоледа и стопа криминала. Упркос добијеној позитивној корелацији, тиме није доказано да лизање сладоледа проузрокује убилачко понашање. Што је одлично, пошто би произвођачи сладоледа иначе брзо затворили своје фабрике!

Усредсређе је јасно - пазите се могућег утицаја треће, реметилачке променљиве када пројектујете своје истраживање. Ако сумњате да би неке друге променљиве могле утицати на резултат, погледајте можете ли да их измерите у исто време. Делимичном корелацијом може се статистички контролисати утицај додатних променљивих и тако стећи јаснији и мање контаминиран показатељ везе две променљиве од интереса.

Статистичка значајност у односу на практичну

Немојте се превише узбуђивати када добијете статистички значајне коефицијенте корелације. На великим узорцима, статистичку значајност могу досећи чак и сасвим мали коефицијенти корелације. Практичан значај корелације 0,2 веома је ограничен, макар њен статистички значај био доказан.

Усредсредите се на стварну величину Пирсоновог коефицијента r и износ заједничке варијансе две променљиве. **Коефицијент детерминације** показује колики је део варијансе две променљиве заједнички; још се каже "колики је део варијансе једне променљиве објашњен (или проузрокован) варијансом друге". Израчунава се једноставно; само квадрирајте вредност Пирсоновог коефицијента r (помножите r са самим собом). Да би сте то претворили у "проценат варијансе", добијени производ помножите са 100.

Приликом тумачења јачине корелације морате узети у обзир друга истраживања у истој области. Ако су други истраживачи у тој области успели да предвиде само 9 процената варијансе (јер су добили коефицијент корелације 0,3) одређеног исхода (нпр. анксиозности), онда би ваша студија која објашњава 25 процената у поређењу с тим била импресивна. У некој другој области, 25 процената објашњене варијансе може изгледати као мали и неважан резултат.

Претпоставке

Све технике обрађене у овом делу имају неколико заједничких претпоставки, које ћемо сада размотрити.

Ниво мерења

Скала за мерење променљивих у већини техника требало би да буде интервална (непрекидна). Изузетак од тога би биле једна дихотомна независна променљива (са само две вредности: нпр. пол) и једна непрекидна зависна променљива. Међутим, у свакој категорији дихотомне променљиве требало би да имате приближно једнак број особа или ана лизираних случајева.

Спирманов коефицијент ρ , што је коефицијент корелације прикладан за ординалне или рангиране податке, обрађује се заједно са својом параметарском алтернативом, Пирсоновим коефицијентом корелације r . ρ се често употребљава у здравственој и медицинској литератури, а све више и у психолошким истраживањима, зато што су истраживачи постали свеснији могућих проблема које уме да проузрокује претпоставка да су нумеричке удаљености између рангова приближно једнаке или сразмерне разликама у интензитету посматраних обележја, што је својство интервалних скала.

Повезани парови

Сваки субјект мора дати своју оцену и променљиве X и променљиве Y (повезани парови). Оба податка морају потицати од истог субјекта.

Независност опсервација

Опсервације од којих се подаци састоје морају бити узајамно независне, тј. на било коју опсервацију или мерење не сме утицати ниједна друга опсервација или мерење.

Нормалност

Резултати добијени за све променљиве требало би да су нормално расподељени.

Линеарност

Веза између две променљиве требало би да је линеарна. То значи да би на дијаграму растурања требало да видите (приближно) праву линију, а не криву.

Хомогеност варијансе

Променљивост резултата добијених за променљиву X требало би да је слична за све вредности променљиве Y. Проверите да ли је тако на дијаграму растурања. Требало би да по целој дужини изгледа као приближно једнако дебела цигара.

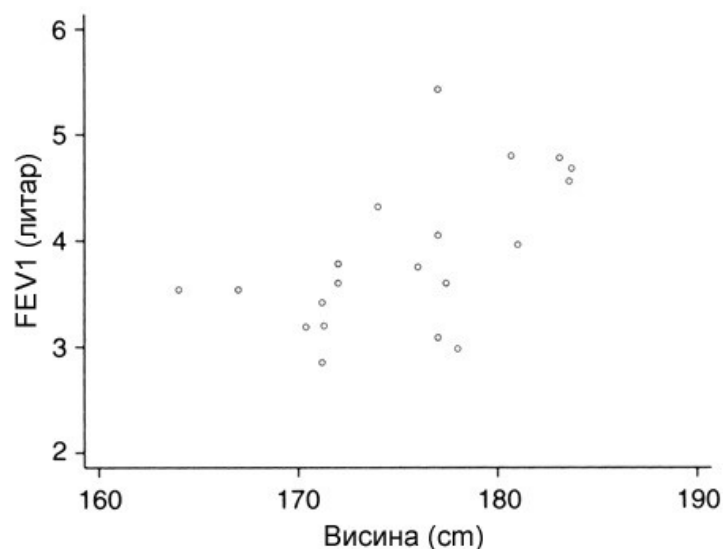
8. Регресија и корелација

8.1 Дијаграми растурања

У овом делу ћемо погледати методе анализирања односа између две квантитативне променљиве. Размотрите табелу 8.1, која приказује податке прикупљене од стране групе студената медицине на часу физиологије. Преглед података указује на то да је могуће да постоји неки однос између FEV1 и висине. Пре покушаја да се квантитативно одреди овај однос, можемо урадити дијаграм за податке и добити слику о природи односа. Уобичајени први дијаграм је дијаграм растурања, део 2.6. Коју променљиву бирамо за коју осу зависи од наших представа о основном односу између њих, као што ће бити објашњено у наставку.

Табела 8.1 FEV1 и висина за 20 мушких студената медицине

Висина (цм)	FEV1 (литри)	Висина (цм)	FEV1 (литри)	Висина (см)	FEV1 (литри)
164.0	3.54	172.0	3.78	178.0	2.98
167.0	3.54	174.0	4.32	180.7	4.80
170.4	3.19	176.0	3.75	181.0	3.96
171.2	2.85	177.0	3.09	183.1	4.78
171.2	3.42	177.0	4.05	183.6	4.56
171.3	3.20	177.0	5.43	183.7	4.68
172.0	3.60	177.4	3.60		



виши од своје деце, ниски родитељи имају тенденцију да буду нижи од своје деце. Galton је назвао овај феномен “регресија према осредњости”, што значи “иде назад према просеку”. Сада се зове **регресија ка средини (regression towards the mean)** (део 8.4). Метод који се користио за испитивање звао се регресиона анализа и име се задржало. Међутим, у Galton-овој терминологији није било “регресије” ако је однос између променљивих био такав да је једна променљива предвиђала тачно другу променљиву; у модерној терминологији не постоји регресија ако се променљиве не односе на све.

Код проблема регресије ми смо заинтересовани за то колико се добро једна променљива може искористити за предвиђање друге променљиве. У случају FEV1 и висине, на пример, ми се бавимо предвиђањем средине FEV1 за дату висину, пре него предвиђањем средине висине за дато FEV1. Имамо две врсте променљивих: променљиву исхода коју покушавамо да предвидимо, у овом случају FEV1, и предсказивач (*predictor*) или променљиву која објашњава, у овом случају висину. Предсказивач променљива се често назива независна променљива и променљива исхода се зове зависна променљива. Међутим, ови појмови имају друга значења у вероватноћи (део 3.2), тако да их нећемо користити. Ако означимо предсказивач променљиву са X , а променљиву исхода са Y , однос између њих може бити написан као

$$Y = a + bX + E$$

где су a и b константе, а E је случајна променљива са средином 0, звана **грешка (error)**, која представља онај део варијабилности од Y који није објашњен односом са X . Да средина од E није била нула, могли бисмо то учинити променом a . Претпостављамо да је E независно од X .

8.3 Метода најмањих квадрата

Када би све тачке лежале дуж линије и када не би било случајне варијације, било би лако повући линију на дијаграму растурања. На слици 8.1 то није случај. Постоји много могућих вредности a и b које би могле да представљају податке и потребан нам је критеријум за одабир најбоље линије. Слика 8.3 приказује одступање тачке од линије, растојање од тачке до линије у Y правцу. Линија ће се добро уклопити у податке ако су одступања од ње мала, а лоше ако су одступања велика. Ова одступања представљају грешку E онај део променљиве Y који X не објашњава. Једно решење проблема проналажења најбоље линије је да изаберете ону која оставља минималну количину варијабилности Y необјашњеном, правећи варијансу E минималном. Ово ће се постићи правећи збир квадрата одступања око линије минималним. То се зове метода најмањих квадрата и пронађена линија је линија најмањих квадрата.

Метода најмањих квадрата је најбољи метод ако су одступања од линије Нормално расподељена са униформном варијансом дуж линије. Ово ће вероватно бити случај, пошто регресија има тенденцију да уклони из Y варијабилност између субјеката и остави грешку мерења, која ће вероватно бити Нормална.

Многи корисници статистике су збуњени минимизирањем варијације само у једном смеру. Обично су обе променљиве мерене са неким грешкама па опет изгледа да игноришемо грешку у X . Зашто не би смањили нормална одстојања до линије пре него вертикална одстојања? Постоје два разлога за то. Прво, ми налазимо ударно предвиђање Y из посматраних вредности X , а не из “правих” вредности X . Грешка мерења у обе променљиве је један од узрока одступања од линије, и налази се у овим одступањима измереним у Y правцу. Друго, линија пронађена на овај начин зависи од јединица у којима се променљиве мере. За податке из табеле 8.1 линија пронађена овим методом је

$$\text{FEV1 (литар)} = -9.33 + 0.075 \times \text{висина (цм)}$$

Ако меримо висину у метрима уместо у сантиметрима, добијамо

$$\text{FEV1 (литар)} = -34.70 + 22.0 \times \text{висина (м)}$$

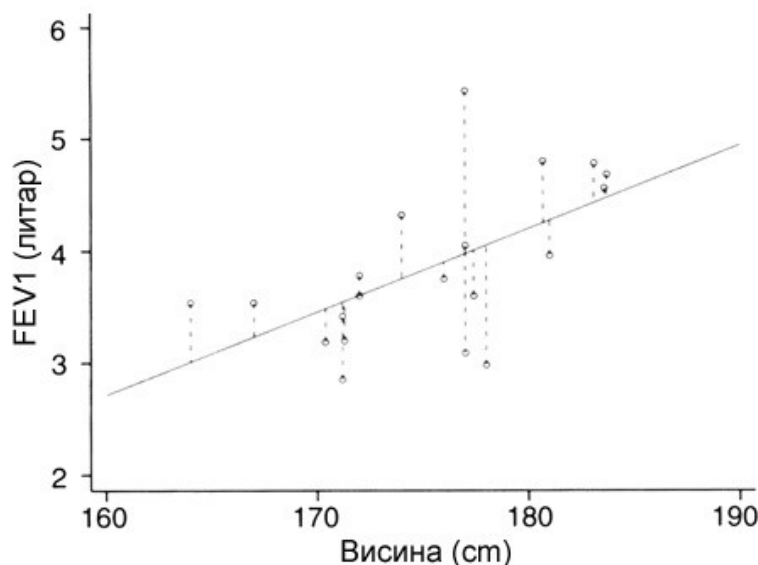
Тако по овом методу предвиђени FEV1 за студента висине 170 цм је 3.42 литара, али за студента висине 1.70 м је 2.70 литара. Ово је очигледно незадовољавајуће и нећемо даље узимати у обзир овакав приступ.

Вративши се на слику 8.3, једначина линије која минимизира збир квадрата одступања од линије у променљивој исхода, пронађена је прилично лако. Решење је:

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} = \frac{\text{сума производа око средине } X \text{ и } Y}{\text{сума квадрата око средине } X}$$

Затим проналазимо одсечак a преко

$$a = \bar{y} - b \bar{x}$$



Слика 8.3 Одступање од линије у правцу "у"

Приметите да линија мора да прође кроз тачку средине, (\bar{x}, \bar{y}) . Бројилац, **сума производа око средине (sum of products about the mean)**, је сличан суми квадрата око средине као што је употребљено у израчунавању варијансе. Рећи ћемо нешто више о својствима **суме производа (sum of products)**, како се обично назива, када будемо расправљали о корелацији. Уклапање праве линије помоћу овог метода зове се **једноставна линеарна регресија (simple linear regression)**.

Једначина $Y = a + bX$ се зове **регресиона једначина Y на X (regression equation of Y on X)**, где је Y променљива исхода и X предсказивач. Градијент b се такође зове **коэффициент регресије (regression coefficient)**. Ми ћемо га израчунати за податке из табеле 8.1. Имамо да је

$$\begin{array}{lll} \sum x_i = 3507.6 & \sum x_i^2 = 615739.24 & n = 20 \\ \sum y_i = 77.12 & \sum y_i^2 = 306.8134 & \sum x_i y_i = 13568.18 \\ \bar{x} = 3507.6/20 = 175.38 & & \bar{y} = 77.12/20 = 3.856 \end{array}$$

$$\text{сума квадрата за } X = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = 615739.24 - \frac{(3507.6)^2}{20} = 576.352$$

$$\text{сума квадрата за } Y = \sum y_i^2 - \frac{(\sum y_i)^2}{n} = 306.8134 - \frac{(77.12)^2}{20} = 9.43868$$

$$\text{сума производа око средине} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} = 13568.18 - \frac{3507.6 \times 77.12}{20} = 42.8744$$

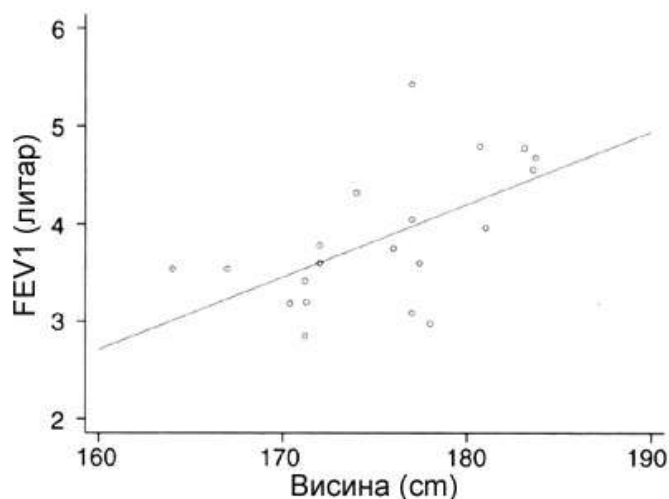
Не треба нам сума квадрата за Y још увек, али биће нам потребна касније.

$$b = \frac{42.8744}{576.352} = 0.074389 \text{ литара/цм}$$

$$a = \bar{y} - b\bar{x} = 3.856 - 0.074389 \times 175.38 = -9.19 \text{ литра}$$

Отуда је једначина регресије FEV1 за висину

$$\text{FEV (литар)} = -9.19 + 0.0744 \times \text{висина (цм)}$$

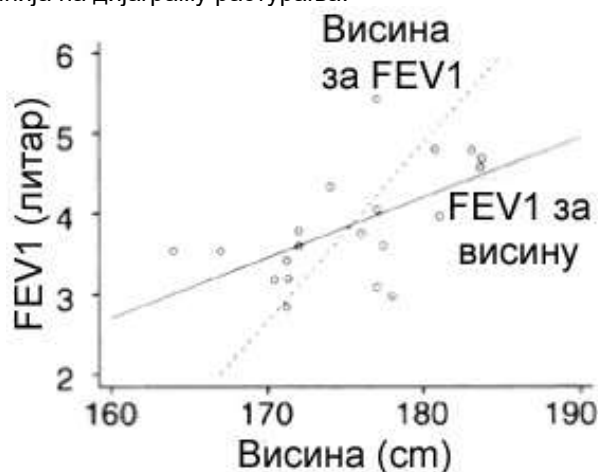


Слика 8.4 Регресија FEV1 за висину

Слика 8.4 приказује линију нацртану на дијаграму растурања. Коефицијенти a и b имају димензије, у зависности од димензија X и Y . Ако променимо јединице у којима су X и Y измерени такође мењамо a и b , али не мењамо линију. На пример, ако се висина мери у метрима, поделимо x_i са 100 и нађемо да је b помножено са 100 да би дало $b = 7.4389$ литара/м. Линија је

$$\text{FEV1 (литар)} = -9.19 + 7.44 \times \text{висина (m)}$$

Ово је потпуно иста линија на дијаграму растурања.



Слика 8.5 Линија регресије за податке из табеле 8.1

Шта се дешава ако променимо наш избор променљивих исхода и предиктора? Регресиона једначина висине на FEV1 је
 $\text{висина} = 158 + 4.54 \times \text{FEV1}$

Ово није исто као линија регресије FEV1 на висину. Ако преуредимо ову једначину дељењем сваке стране са 4.54 добијамо
 $\text{FEV1} = -34.8 + 0.220 \times \text{висина}$

Нагиб регресије висине на FEV1 је већи него FEV1 на висину (слика 8.5). У принципу, нагиб регресије X на Y је већи него Y на X , када је X хоризонтална оса. Само ако све тачке леже тачно на правој линији две једначине су исте.

8.4 Стандардна грешка коефицијента регресије

У сваком поступку предвиђања, желимо да знамо колико су поуздане наше предвиђања. То радимо проналазећи њихове стандардне грешке и тиме интервале поверења. Можемо такође тестирати хипотезе о коефицијентима, на пример, нулту хипотезу да у популацији нагиб је нула и да не постоји линеарна зависност. Прво налазимо суму квадрата одступања од линије, тј. разлику између посматраних y_i и вредности предвиђених регресионом линијом. Ово је

$$\sum (y_i - \bar{y})^2 - b^2 \sum (x_i - \bar{x})^2$$

$\sum (y_i - \bar{y})^2$ је наравно, укупан збир квадрата око средине y_i . Појам $b^2 \sum (x_i - \bar{x})^2$ се зове **збир квадрата услед регресије на X (sum of squares due to regression on X)**. Разлика између њих је **резидуални збир квадрата (residual sum of squares)**, или **збир квадрата око регресије (sum of squares about regression)**. Збир квадрата услед регресије подељен са укупним збиром квадрата зове се **пропорција варијабилности објашњена регресијом (proportion of variability explained by the regression)**.

У циљу предвиђања варијансе потребни су нам степени слободе са којима делимо збир квадрата. Проценили смо не један параметар на основу података, као за суму квадрата око средине (део 1.6), већ два параметра, a и b . Губимо два степена слободе, остављајући нас са $n - 2$. Стога је варијанса Y око линије, звана **резидуална варијанса (residual variance)**

$$s^2 = \frac{1}{n-2} \left(\sum (y_i - \bar{y})^2 - b^2 \sum (x_i - \bar{x})^2 \right)$$

Да бисмо предвидели варијацију око линије, морамо претпоставити да је иста све до краја линије, односно да је варијанса униформна. Ово је исто као и за t метод два узорка (део 7.3) и анализу варијансе. За FEV1 податке збир квадрата услед регресије је $(0.074389)^2 \times 576.352 = 3.18937$, а збир квадрата око регресије је $9.43868 - 3.18937 = 6.24931$.

Има $20 - 2 = 18$ степени слободе, тако да је варијанса око регресије $s^2 = 6.2493/18 = 0.34718$. Стандардна грешка од b је дата помоћу

$$SE(b) = \frac{s^2}{\sum (x_i - \bar{x})^2} = \sqrt{\frac{0.34718}{576.352}} = 0.02454 \text{ литара/цм}$$

Већ смо претпоставили да је грешка E нормално расподељена, тако да такође и b мора бити нормално расподељено. Стандардна грешка је заснована на једном збиру квадрата, тако да је $b/SE(b)$ посматрање из t расподеле са $n - 2$ степена слободе (део 7.1). Можемо пронаћи 95% интервал поверења за b узимајући t стандардне грешке на обе стране предвиђања. На пример, имамо 18 степени слободе. Из табеле 7.1, 5% тачка t расподеле је 2.10, тако да 95% интервал поверења за b је $0.074389 - 2.10 \times 0.02454$ до $0.074389 + 2.10 \times 0.02454$ или 0.02 до 0.13 литара/цм. Можемо видети да су FEV1 и висина повезани, иако нагиб није добро процењен.

Можемо такође тестирати нулту хипотезу да је у популацији, нагиб $= 0$ у односу на алтернативу да нагиб није једнак 0, односно у било ком смеру. Тест статистика је $b/SE(b)$ и ако је нулта хипотеза тачна ово ће бити из t расподеле са $n - 2$ степени слободе. На пример,

$$t = \frac{b}{SE(b)} = \frac{0.074389}{0.02454} = 3.03$$

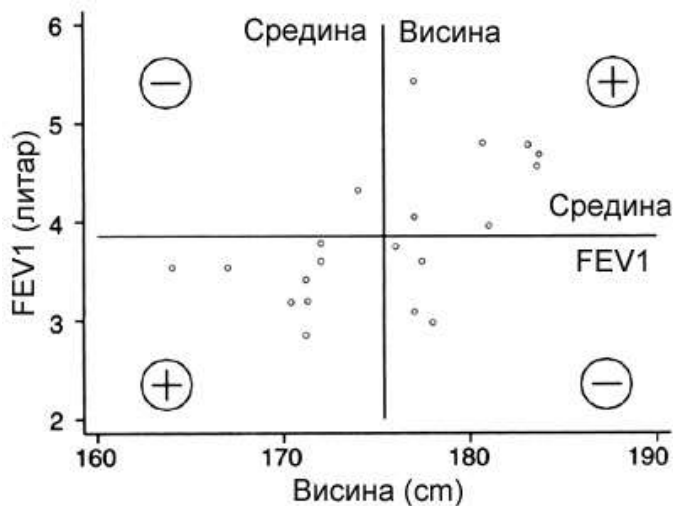
Из табеле 7.1 ово има двострану вероватноћу мању од 0.01. Рачунар нам говори да је вероватноћа око 0.007. Стога подаци су у супротности са нултом хипотезом и подаци пружају прилично добар доказ да постоји веза. Да је узорак био много већи, могли бисмо да га изделимо t расподелом и употребимо Стандардизовану Нормалну расподелу.

8.5 Корелација

Регресиони метод нам говори нешто о природи односа између две променљиве, како се једна мења са другом, али нам не каже колико је тај однос веран. Да бисмо то учинили потребан нам је другачији коефицијент, коефицијент корелације. Коефицијент корелације је заснован на збиру

производа око средине две променљиве, тако да ћемо почети разматрањем особина збира производа и зашто је то добар показатељ блискости односа.

Слика 8.6 приказује дијаграм растурања са слике 8.1 са две нове осе нацртане кроз тачку средине. Растојања тачака од ових оса представљају одступања од средине. У горњем десном делу слике 8.6, одступања од средине обе променљиве, FEV1 и висине, су позитивна. Стога ће њихови производи ће бити позитивни. У доњем левом делу, одступања обе променљиве од средине су негативна. Поново, њихов производ ће бити позитиван. У горњем левом делу слике 8.6, одступања FEV1 од своје средине биће позитивна, а одступања висине од своје средине биће негативна. Њихов производ биће негативан. У доњем десном делу, производ ће опет бити негативан. Дакле, на слици 8.6 скоро сви ови производи ће бити позитивни, и њихов збир биће позитиван. Кажемо да постоји **позитивна корелација (positive correlation)** између две променљиве; док се једна повећава, то чини и друга. Да се једна променљива смањује док се друга повећава, имали бисмо дијаграм растурања где би већина тачака лежала у горњем левом и доњем десном делу. У овом случају би збир производа био негативан и постојала би **негативна корелација (negative correlation)** између променљивих. Када две променљиве нису повезане, имамо дијаграм растурања са приближно истим бројем тачака у сваком од делова. У овом случају, има онолико позитивних колико и негативних производа, и сума је нула. Постоји **нула корелација (zero correlation)** или **нема корелације (no correlation)**. За променљиве се каже да **нису у корелацији (uncorrelated)**.



Слика 8.6. Дијаграм растурања са осама кроз средину

Вредност збира производа зависи од јединица у којима су две променљиве измерене. Можемо наћи коефицијент без димензија ако поделимо збир производа са квадратним кореном збира квадрата од X и Y . Ово нам даје **производ момента коефицијента корелације (product moment correlation coefficient)**, или укратко **коефицијент корелације (correlation coefficient)**, обично означен са r . Коефицијент Пирсонове линеарне корелације r погодан је за интервалне (непрекидне) променљиве.

Ако су n парова посматрања обележени са (x_i, y_i) , онда се r добија помоћу

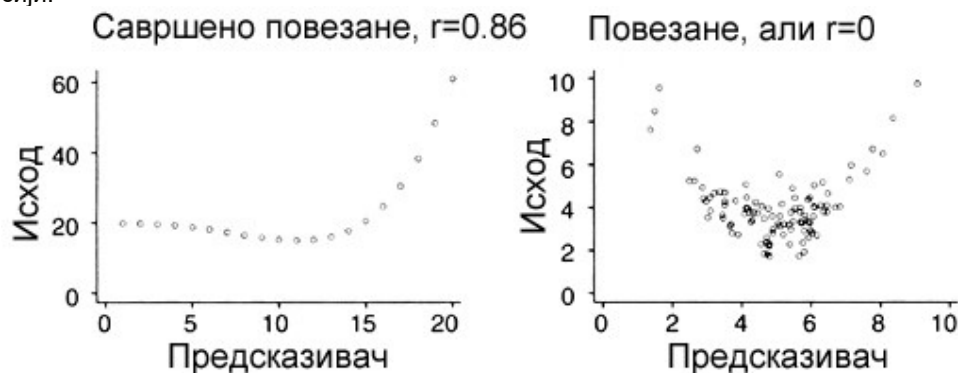
$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\sum (x_i - \bar{x})^2\right)\left(\sum (y_i - \bar{y})^2\right)}} = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sqrt{\left(\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right)\left(\sum y_i^2 - \frac{(\sum y_i)^2}{n}\right)}} =$$

$$= \frac{\text{suma proizvoda oko sredine } X \text{ i } Y}{\sqrt{\text{suma kvadrata oko sredine } X \times \text{suma kvadrata oko sredine } Y}}$$

За FEV1 и висину имамо

$$r = \frac{42.8744}{\sqrt{576.352 \times 9.43668}} = 0.58$$

Ефекат дељења збира производа са квадратним кореном збира квадрата одступања сваке променљиве је прављење да коефицијент корелације лежи између -1.0 и +1.0. Када све тачке леже тачно на правој линији таквој да Y расте док X расте, тада је $r = 1$. Ово може бити приказано стављањем $a + b x_i$ на место y_i у једначини за r ; све се поништава остављајући $r = 1$. Када се све тачке леже тачно на правој линији са негативним нагибом, $r = -1$. Када уопште не постоји никаква веза, $r = 0$, јер је збир производа нула. Коефицијент корелације описује блискост линеарног односа између две променљиве. Није битно коју променљиву узимамо да буде Y , а коју да буде X . Не постоји избор предсказивача и променљиве исхода, као што постоји у регресији.



Слика 8.7 Подаци где коефицијент корелације може да доведе у забуну

Коефицијент корелације мери колико су тачке близу правој линији. Чак и ако постоји савршен математички однос између X и Y , коефицијент корелације неће бити тачно 1, осим ако је у облику $y = a + bx$. На пример, слика 8.7 приказује две променљиве које су савршено повезане, али имају $r = 0.86$. Слика 8.7 такође приказује две променљиве који су јасно повезане, али имају нула корелацију, јер однос није линеаран. Ово поново показује значај прављења дијаграма података, не ослањајући се на преглед статистике, као што је само коефицијент корелације. У пракси, односи као они на сликама 8.7 су ретки у медицинским подацима, иако могућност за овакве податке увек постоји. Чешће, постоји толико случајних варијација да није лако препознати било какав однос уопште.

Коефицијент корелације r је у вези са коефицијентом регресије b на једноставан начин. Ако је $Y = a + bX$ регресија Y на X , и ако је $X = a' + b'Y$ регресија X на Y , онда је $r^2 = bb'$. Ово произилази из формула за r и b . За FEV1 податке, $b = 0.074389$ и $b' = 4.5424$, тако да је $bb' = 0.07439 \times 4.5424 = 0.33790$, па је квадратни корен од тога једнак 0.58129, и то је вредност коефицијента корелације. Такође имамо

$$r^2 = \frac{(\text{сума производа око средине})^2}{\text{сума квадрата од } X \times \text{сума квадрата од } Y} = \frac{(\text{сума производа око средине})^2}{(\text{сума квадрата од } X)^2} \times \frac{\text{сума квадрата од } X}{\text{сума квадрата од } Y} = b^2 \times \frac{\text{сума квадрата од } X}{\text{сума квадрата од } Y}$$

То је објаснила пропорција варијабилности, описана у делу 8.4.

8.6 Значај теста и интервал поверења за r

Тестирање нулте хипотезе да је $r = 0$ у популацији, односно да не постоји линеарна зависност, је једноставно. Тест је нумерички еквивалентан тестирању нулте хипотезе да је $b = 0$, и тест важи под условом да је најмање једна од променљивих из Нормалне расподеле. Овај услов је ефективно исти као и онај за тестирање b , где остаци у Y правцу морају да буду Нормални. Ако је $b = 0$, остаци у Y правцу су једноставно одступања од средине, и они ће бити нормално расподељени само ако је и Y нормално расподељено. Ако услов није испуњен, можемо користити трансформацију, или један од методе корелације ранга (део 9.4 и 9.5).

Зато што коефицијент корелације не зависи од средина или варијанси посматрања, расподелу коефицијента корелације узорка када је коефицијент корелације популације нула је лако саставити у виду табеле. Табела 8.2 приказује коефицијент корелације на 5% и 1% нивоа значајности. На пример, имамо да је $r = 0.58$ из 20 посматрања. Тачка од 1% за 20 посматрања је 0.56, тако да имамо да је $P < 0.01$, а мало је вероватно да би се корелације појавиле да није било линеарног односа у популацији. Обратите пажњу да су вредности r које могу да настану случајно са малим узорцима прилично високе. Са 10 тачака r би требало да буде веће од 0.63 да би било значајно. Са друге стране, са 1 000 тачака, врло мале вредности r , мале као и 0.06, биће значајне.

Проналажење интервала поверења за коефицијент корелације је теже. Чак и када су X и Y Нормално расподељени, r само не приступа Нормалној расподели док величина узорка није у хиљадама. Штавише, његова расподела је веома осетљива на одступања од Нормале у X и Y правцу.

Међутим, ако су обе променљиве из Нормалне расподеле, Фишера z трансформација даје Нормално расподељену променљиву чија средина и варијанса су познате у појмовима коефицијента корелације популације који желимо да проценимо. Из овога може бити пронађен интервал поверења. **Фишера z трансформација (Fisher's z transformation)** је

$$z = \frac{1}{2} \log_e \left(\frac{1+r}{1-r} \right)$$

која следи Нормалну расподелу са средином

$$z_\rho = \frac{1}{2} \log_e \left(\frac{1+\rho}{1-\rho} \right) + \frac{\rho}{2(n-1)}$$

и варијансом $1/(n-3)$ приближно, где је ρ коефицијент корелације популације и n је величина узорка. 95% интервал поверења за z ће бити приближно $z \pm 1.96 \sqrt{1/(n-3)}$. За FEV1 податке, $r = 0.58$, а $n = 20$.

$$z = \frac{1}{2} \log_e \left(\frac{1+0.58}{1-0.58} \right) = 0.6625$$

95% интервал поверења ће бити $0.6625 \pm 1.96 \sqrt{1/17}$, дајући 0.1871 до 1.1379. Трансформација назад од z скале до скале коефицијента корелације је

$$r = \frac{\exp(2z) - 1}{\exp(2z) + 1}$$

тако да за доњу границу имамо

$$\frac{\exp(2 \times 0.1871) - 1}{\exp(2 \times 0.1871) + 1} = 0.18$$

а за горњу границу

$$\frac{\exp(2 \times 1.1379) - 1}{\exp(2 \times 1.1379) + 1} = 0.81$$

и 95% интервал поверења је 0.18 до 0.81. Ово је веома широко, одражавајући варијацију узорковања коју коефицијент корелације има за мале узорке. Коефицијенти корелације се морају третирати са дозом опреза када су изведени из малих узорака.

Олакшање теста значајности у односу на релативну сложеност израчунања помоћу интервала поверења је значило да је у прошлости тест значајности обично био дат за коефицијент корелације. Повећана доступност рачунара са добро написаним статистичким пакетима би требало да доведе до појављивања коефицијената корелације са интервалима поверења у будућности.

Табела 8.2 Двостране 5% и 1% тачке расподеле коефицијента корелације r , по нултој хипотези

n	5%	1%	n	5%	1%	n	5%	1%
3	1.00	1.00	16	0.50	0.62	29	0.37	0.47
4	0.95	0.99	17	0.48	0.61	30	0.36	0.46
5	0.88	0.96	18	0.47	0.59	40	0.31	0.40
6	0.81	0.92	19	0.46	0.58	50	0.28	0.36
7	0.75	0.87	20	0.44	0.56	60	0.25	0.33
8	0.71	0.83	21	0.43	0.55	70	0.24	0.31
9	0.67	0.80	22	0.42	0.54	80	0.22	0.29
10	0.63	0.77	23	0.41	0.53	90	0.21	0.27
11	0.60	0.74	24	0.40	0.52	100	0.20	0.25
12	0.58	0.71	25	0.40	0.51	200	0.14	0.18
13	0.55	0.68	26	0.39	0.50	500	0.09	0.12
14	0.53	0.66	27	0.38	0.49	1000	0.06	0.08
15	0.51	0.64	28	0.37	0.48			

 n = број посматрања

8.7 Коришћење коефицијента корелације

Коефицијент корелације има неколико примена. Коришћењем табеле 8.2, он обезбеђује једноставно тестирање нулте хипотезе да променљиве нису линеарно повезане, са мање израчунавања у односу на метод регресије. Такође је користан као статистика сумирања за снагу односа између две променљиве. Ово је од велике вредности када разматрамо међусобне односе између великог броја променљивих. Можемо поставити квадратни низ корелација за сваки пар променљивих, који се зове **матрица корелације (correlation matrix)**. Испитивање матрице корелације може бити веома поучно, али морамо имати на уму могућност нелинеарних односа. Не постоји замена за прављење дијаграма података. Матрица корелације такође обезбеђује почетну тачку за одређени број метода за решавање великог броја променљивих истовремено.

8.8 Вишеструка регресија

Вишеструка регресиона анализа може се посматрати као проширење корелационе анализе. Њен резултат је једначина која представља најбоље предвиђање зависне променљиве на основу више независних променљивих. Регресиона анализа се користи када је познато да је зависна променљива корелисана са независним променљивим. Независне променљиве могу бити категоријске или непрекидне. Међутим, у случају када су независне променљиве категоријске променљиве, оне се морају прекодирати у вештачке променљиве да би могле да се укључе у регресиону анализу. На другој страни, зависна променљива мора бити мерена на непрекидној скали. Ако зависна променљива није непрекидна, треба размислити о евентуалној примени дискриминационе анализе.

Постоје три основна регресиона модела или приступа регресионој анализи: стандардни или истовремени (симултани), хијерархијски или секвенцијални и постепени, тј. "у корацима" (*stepwise*). Наведени модели се разликују по два основа: по третману дела варијабилитета који се преклапа између независних променљивих, јер је обично присутна корелација независних променљивих, и по редоследу укључења независних променљивих у регресиони модел.

У случају примене стандардног регресионог модела, све независне променљиве се укључују у регресиони модел заједно, јер је циљ анализе да се испита веза између скупа предиктора (све независне променљиве) и зависне променљиве.

У случају примене хијерархијске вишеструке регресије, сам истраживач одређује на основу претходног теоријског знања, редослед укључења независних променљивих у модел.

У случају примене модела постепене регресије, или регресије "у корацима", један број независних променљивих се укључује у модел; редослед њиховог укључења зависи искључиво од статистичких критеријума, који су уграђени у *stepwise* процедуру у СПСС програму. Конкретна метода укључења независних променљивих у регресиони модел може бити унапред - *forward*, уназад - *backward* или комбинација претходне две методе - *stepwise*. У случају примене форвард методе, укључује се једна по једна независна променљива у модел. Поредак укључења и њихов опстанак у моделу одређени су на бази статистичког критеријума. Као статистички критеријум за улазак променљиве у модел користи се вредност F-тест статистике, која треба да буде већа од одређене критичне вредности (FIN), а потребно је да је постигнута и критична вредност α нивоа (PIN). Код *backward* методе полази се од тога да су све независне променљиве укључене у регресиони модел, а затим се једна по једна променљива евентуално искључују из модела на бази критеријума да ли је парцијална F-вредност мања од критичне вредности (FOUT). Мора бити задовољен и критеријум за α ниво (POUT). У СПСС програму постављене су одређене вредности за статистичке критеријуме: FIN PIN, FOUT и POUT (default values), које се, по потреби могу мењати. *Stepwise* метода је комбинација претходне две методе. За њу је карактеристично наизменично проверава оправданост и укључења и евентуалног искључења променљиве из модела, да се може десити да се из модела искључи независна променљива која је у неком претходном задовољила критеријум за укључење у модел.

Избор методе у великој мери зависи од самог циља истраживања.

Тестирање претпоставки

Вишеструки регресиони модел се заснива на низу претпоставки:

1. **Количник броја опсервација и независних променљивих** - Број опсервација (случајева) који потребан за спровођење регресионе анализе зависи од типа регресионог модела који ће бити примењен. *За стандардни и хијерархијски регресиони модел идеално је да имате 20 пута више опсервација него независних променљивих*, тј. предиктора, док је за постепени регресиони модел потребно још више опсервација. *Минимални захтев је да имате бар пет пута више опсервација него независних променљивих*.
2. **Нетипичне вредности (outliers)** - Екстремне вредности имају значајан утицај на регресиони модел, тако да их треба или искључити из анализе или их модификовати да би био смањен њихов утицај. Униваријационе нетипичне вредности, нетипичне вредности за сваку променљиву појединачно, могу се открити у поступку скрининга података. Са друге стране, мултиваријационе нетипичне вредности се могу открити уз примену статистичких метода, као што је Махаланобисово одстојање и графичке методе, као што је дијаграм распршености резидуала. Одлука о искључењу нетипичних вредности из скупа података мора бити пажљиво донета, јер брисање једних

нетипичних вредности обично има за последицу генерисање других нетипичних података.

3. **Мултиколинеарност и сингуларност** - Мултиколинеарност се односи на присуство високе корелације између независних променљивих ($r=0,7$ или више), а појам сингуларност на постојање перфектне корелације између независних променљивих (када је једна "независна" променљива заправо комбинација других независних променљивих). Ови проблеми се одражавају на начин интерпретације јачине везе између независних променљивих и зависне променљиве. Они се могу открити посматрањем корелационе матрице (свих независних променљивих међу собом и сваке независне променљиве са зависном променљивом), квадрата вишеструке корелације и показатеља мултиколинеарности - нивоа толеранције за сваку променљиву. Већина рачунарских програма има уграђене критичне вредности за мултиколинеарност и не дозвољава укључење независних променљивих које би могле да створе проблем.
4. **Нормалност расподеле, линеарност, хомоскедастичност и независност резидуала** – На основу дијаграма распршености резидуала, може се проверити да ли су наведене претпоставке испуњене. Претпоставља се да разлике између оригиналних вредности зависне променљиве и моделом предвиђене вредности зависне променљиве (резидуали) имају нормалну расподелу. Штавише, претпоставља се да резидуали имају линеарну везу са предвиђеним вредностима зависне променљиве и да је варијанса резидуала иста за све предвиђене вредности. Благо одступање од линеарности није велики проблем, али умерено и екстремно одступање може довести до озбиљног потцењивања зависности у моделу.

Претпоставка 1 се односи на сам план истраживања, а претпоставке 2, 3 и 4 се проверавају у регресионе анализе.

Полази се од претпоставке да што је више независних променљивих у моделу, све је мањи утицај латентне променљиве (стандардне грешке) ε_i , $i = 1, 2, \dots, n$. Веома је битно пажљиво бирати променљиве које ће бити укључене у модел. Основни вишеструки регресиони модел изгледа на следећи начин:

$$\hat{x}_{i1.23\dots m} = a_{1.23\dots m} + b_{12.34\dots m}x_{i2} + b_{13.24\dots m}x_{i3} + \dots + b_{1m.23\dots(m-1)}x_{im} + \varepsilon_i, \quad i = 1, 2, \dots, n$$

$a_{1.23\dots m}$ - слободни члан

$\hat{x}_{i1.23\dots m}$, $i = 1, 2, \dots, n$ – појединачне вредности регресије

x_{i2}, \dots, x_{ij} , $i = 1, 2, \dots, n$; $j = 1, 2, \dots, m$ – вредности независних променљивих

$b_{12.34\dots m}, b_{13.24\dots m}, b_{1m.23\dots(m-1)}$ – регресиони коефицијенти

ε_i , $i = 1, 2, \dots, n$ – латентна променљива (случајна грешка)

m , број независних променљивих

n , величина узорка

Веома често у литератури се за означавање зависне променљиве користи симбол „Y“. У овом тексту ће наизменично бити употребљавана оба симбола.

Овај модел даје најбоље могуће предвиђање вредности зависне променљиве на основу вредности независних променљивих, ако су све претпоставке испуњене. На основу величине регресионих коефицијената можемо закључити колики је релативни утицај или важност сваке независне променљиве ако се ти коефицијенти конвертују у бета коефицијенте β . Ови коефицијенти се добијају када се све вредности променљивих стандардизују.

Једна од претпоставки за употребу регресионе анализе јесте постојање линеарне зависности између променљивих. Она је неопходна јер анализа започиње израчунавањем коефицијената просте корелације (биваријантних корелација) за све парове променљивих, а сва ова израчунавања захтевају линеаран однос између парова променљивих.

Вишеструка регресија ће бити приказана на хипотетичком примеру са 7 независних променљивих. Приликом корелационе анализе, од посебног интереса је одређивање степена повезаности између променљивих. Корелациона анализа нам пружа следеће:

1. Релативну важност сваке независне променљиве у предвиђању или утицају на зависну променљиву

2. Степен до којег све независне променљиве комбиновано објашњавају варијације зависне променљиве

Одговоре на ова питања добијамо преко величине стандардизованог регресионог коефицијента β и коефицијента Пирсонове корелације r . У примеру са 7 независних променљивих ови показатељи су израчунати и приказани у табели 8.3.

Табела 8.3 Вредности коефицијената за пример са 7 независних променљивих			
Променљива	Коефицијент Пирсонове корелације r	Стандардизовани регресиони коефицијент β	Регресиони коефицијент b
1	2	3	4
X_1	0,63	0,55	2,89
X_2	0,52	0,27	10,41
X_3	0,40	0,15	6,62
X_4	0,21	0,17	1,32
X_5	0,11	-0,04	-5,08
X_6	0,06	0,22	3,44
X_7	0,03	0,01	4,45

У другој колони коефицијенти Пирсонове корелације показују јачину везе између сваке независне променљиве посебно са зависном променљивом Y . Овај коефицијент се креће у интервалу од 0,03 до 0,63. Када се ове вредности подигну на квадрат добијају се коефицијенти детерминације који објашњавају колико дата независна променљива има удела у објашњавању варијација независне променљиве. На пример, ако се први регресиони коефицијент подигне на квадрат, добија се да независна променљива X_1 објашњава 39,7% варијација зависне променљиве Y .

Мултиколинеарност. Пошто се коефицијенти корелације и бета коефицијенти узимају као мере релативне важности сваке независне променљиве, вредности у другој и трећој колони табеле 8.3 би требале да буду пропорционалне или бар да опадају истим редом. Међутим, види се да то није случај. Разлог лежи у мултиколинеарности или просто колинеарности. Мултиколинеарност показује колика је међузависност између независних променљивих. Што је већа мултиколинеарност, то се више одражава на бета коефицијенте и они све мање могу да се употребе као показатељи релативног утицаја сваке независне променљиве.

Разлог лежи у томе што се регресиони коефицијенти, b и β , увек израчунавају тако да дају најбоље могуће предвиђање зависне променљиве Y , а не да покаже релативну важност сваке независне променљиве X . Када је мултиколинеарност мала и не постоји онда су регресиони коефицијенти приближно пропорционални коефицијентима просте корелације па и једни и други дају сличну представу о релативној важности независних променљивих. Ако постоји значајна мултиколинеарност, онда ће најзначајнијој независној променљивих бити додељена права вредност бета коефицијента, док ће код осталих независних променљивих бета вредност бити много мања да би се избегла међузависност и међусобни утицај независних променљивих.

У табели 8.3, пошто величине бета коефицијената нису пропорционалне са коефицијентима корелације, може се закључити да постоји значајна мултиколинеарност. На пример, видимо да најзначајнија независна променљива X_1 има висок коефицијент корелације и бета коефицијент, али већ X_2 има нешто мањи коефицијент корелације али дупло мању вредност бета коефицијента. Несразмера се понавља и код других променљивих у моделу. То је због тога што се преклапа утицај независних променљивих па су због тога бета коефицијенти свих променљивих осим X_1 пуно мањи. Овај проблем се може решити преко постепене регресије.

Прихватљиви ниво мултиколинеарности није лако одредити. Он зависи од броја независних променљивих у моделу, колико њих је корелисано и у ком обиму. Потребно је на

почетку израдити таблицу Пирсонових коефицијената корелације између свих променљивих. Коефицијенти Пирсонове корелације до 0,5 између неколико независних променљивих обично не би требало да утичу на регресионе коефицијенте. Ако су поменути коефицијенти просте корелације већи од 0,7 онда је у питању озбиљан проблем. Могућа решења су следећа (Myers & Mullet, 2003, стр. 89):

1. Израдити табелу са свим променљивама и њиховим коефицијентима Пирсонове корелације. Ако код неког пара променљивих коефицијент прелази 0,7, онда се једна од две променљиве елиминише, обично она која има мању корелацију са зависном променљивом Y .
2. Уколико три или више независних променљивих имају велику међусобну корелацију, изабере се она са највећом корелацијом са Y и онда се елиминишу све остале или се изради нова заједничка променљива од свих међузависних променљивих (на основу ваганих вредности или на основу пропорција у корелацији са Y).
3. Изради се анализа главних компоненти за све независне променљиве. Ова техника тражи групу од две или више променљивих које су високо или осредње међусобно корелисане, али су истовремено неповезане са осталим променљивама. За сваку од ових група израђују се вредности које се зову фактор скорови што је врста ваганих просека. Пошто су ови фактор скорови некорелисани и садрже већину информација из оригиналних променљивих, они могу да се употребе као нови сет независних променљивих у вишеструком регресионом моделу. Ова опција је најбоља и препоручује се посебно ако је у питању велики број променљивих (преко 50). Ипак, овим се губи могућност да се посматра свака оригинална променљива појединачно.

Индекс детерминације. Вишеструка регресија такође показује колики је јака међузависност зависне променљиве са свим независним променљивама преко индекса корелације r . Индекс детерминације R^2 показује колики је проценат варијабилитета зависне променљиве објашњен варијабилитетом независних променљивих. У примеру из табеле 8.3 индекс детерминације је 48% што је далеко од пожељне величине од 70%. То значи да неке променљиве које имају значајну повезаност са зависном променљивом Y недостају у моделу, али није познато које су то променљиве. Пошто се индекс корелације и индекс детерминације рачунају на основу података који су прикупљени, дакле пост-фестум, не може се ништа учинити на њиховом побољшању. Ипак, у пракси се препоручује да се прво уради пилот истраживање где се на мањем узорку тестира што већи број променљивих да би се идентификовале све оне које имају најзначајнији утицај, а затим се уради велико посматрање на комплетном узорку где се прикупљају подаци о тим променљивама.

Мултиколинеарност се може утврдити и преко специфичних показатеља као што је, на пример, ниво толеранције. Ниво толеранције је пропорција варијансе променљиве која није повезана са осталим променљивама у регресионом моделу. Висок ниво толеранције, преко 0,8 значи да је та променљива релативно некорелисана са осталим променљивама. Низак ниво толеранције, до 0,2 указује на велику мултиколинеарност и да та променљива мало доприноси објашњавању зависне променљиве у моделу.

Значај вишеструке регресије. Према томе, на основу претходно реченог, вишеструка регресија се користи за добијање одговора на следећа питања:

1. Колико добро све независне променљиве комбиновано објашњавају или им се може приписати разлог за варијације зависне променљиве (R^2).
2. Колика је релативна важност сваке независне променљиве у објашњавању варијација зависне променљиве (бета коефицијенти), под условом да не постоји значајна мултиколинеарност.
3. Која је најбоља предвиђена вредност зависне променљиве за било коју комбинацију независних променљивих.
4. Који се обим промене зависне променљиве може очекивати за сваку јединицу промене сваке независне променљиве (коефицијенти Пирсонове корелације).

Претпоставке на којима се заснива модел вишеструке регресије су сличне онима које важе за линеарну регресију и оне гласе:

5. Облик зависности између свих променљивих је линеаран односно права линија. Ово је поготово важно за однос независних променљивих са зависном променљивом. Све променљиве су непрекидне.
6. Све променљиве имају интервал варијације, дисперзију односно варијансу које имају смисла, односно већина опсервација није једна вредност или интервал.
7. У бази се налази барем три до пет пута више јединица посматрања него што је променљивих јер би у супротном регресиони коефицијенти били непоуздани.
8. Мултиколинеарност између променљивих је мала или не постоји.

Тестирање статистичке значајности. Пре објашњавања резултата потребно је тестирати њихову статистичку значајност. Ако R , b и β нису статистички значајне, закључује се да ниједна независна променљива нема стварну повезаност са зависном променљивом. То значи да добијени модел нема практичну вредност. Већина статистичких софтвера има опцију тестирања.

Уколико су сви регресиони коефицијенти b статистички значајни, онда ће и индекс корелације R бити сигурно значајан. У обрнутом случају то не мора да се деси јер је могуће да се због великог броја променљивих добије статистички значајно R , а да b коефицијенти нису значајни.

Униформно оцењивање. Још један проблем који може да се јави јесте када за неку јединицу посматрања не постоје варијације у прикупљеним вредностима променљивих. На пример, испитаник је на сва или скоро сва питања одговорио истом оценом (на скали од 1 до 10 он је на сва питања заокружио оценом 5). Пошто у том случају не постоје варијације за дату јединицу посматрања, не долази до коваријације са осталим променљивама и јединицама посматрања. Повећава се само величина узорка n или скупа N , али се не повећава коваријанса. На тај начин се вештачки снижава корелација. Ни овде не постоји идеално решење. Уколико су све вредности једнаке боље је такву јединицу елиминисати из анализе. Уколико је присутан део вредности који се понавља за дату јединицу посматрања може се урадити следеће:

1. Елиминисати јединицу посматрања код које не постоји интервал у вредностима променљивих у довољној мери. На пример на мерној скали са 10 вредности интервал за ту јединицу посматрања су само три суседне вредности.
2. Елиминисати јединицу посматрања код које постоји мали број варијација у односу на најчешћу вредност, на пример до 25% посматраних променљивих.
3. Израчунати стандардну девијацију свих вредности променљивих за сваку јединицу посматрања и елиминисати оне јединице посматрања код којих су израчунате вредности близу нуле.

Категоричке вредности. У пракси се често дешава да нису све променљиве изражене на метричкој скали, а да је потребно извести регресиону анализу. Типичан пример таквих променљивих брачни статус, пол, професија, стручна спрема, место становања, држава рођења, итд. Један начин за рад са таквим променљивама је њихово преводјење у категоричке променљиве (*dumty variables*) на следећи начин:

1. Свака категорија (модалитет) се посматра као посебна независна променљива.
2. За сваку јединицу посматрања се додељује вредност „1“ ако јединица поседује неку карактеристику, а „0“ ако је не поседује. На пример, код брачног статуса „1“ за „у браку“ и „0“ за „није у браку“.
3. Нове променљиве се убацују у регресиони модел, али тако да једна категорија из сваке оригиналне променљиве мора бити искључена из анализе.

Разлог за ово искључивање је да се избегне да вредности променљиве буду међусобна линеарна комбинација. На пример, ако имамо четири различита брачна статуса (самац, у браку, разведен(а), удовац-удовица) онда мора једна категорија да има вредност нула и да буде искључена из рачуна. Неки софтвери то раде рачунски ако се променљива на почетку дефинише као категоријска (думму). Ако желимо да укључимо самце у наш регресиони модел,

код брачног статуса са четири модалитета имали би четири променљиве обележене на следећи начин:

Променљиве	Унете вредности
Самац	1
У браку	0
Разведен(а)	0
Удовац-удовица	0

Постепена регресија

Вишеструка регресија нам даје модел у који су укључене све променљиве са којима је анализа и започета, без обзира на њихов различити значај, а такође и у случају када је присутна велика мултиколинеарност. Постепена регресија нам омогућује да се изборимо са проблемом мултиколинеарности и са независним променљивама које су од малог значаја.

Када је мултиколинеарност велика, онда многе променљиве имају слично значење, па није потребно да све оне буду укључене у модел. Постепена регресија омогућава да се елиминишу променљиве које се преклапају са другима и због тога мало или уопште не доприносе тачности у предвиђању модела. Као резултат овог приступа добија се нови модел са мањим бројем независних променљивих који је исто толико добар колико и модел у којем се налазе све независне променљиве.

Типични ток постепене регресије се одвија на следећи начин (Myers & Mullet, 2003, стр. 92):

1. Рачунар изабере једну независну променљиву која има највећу корелацију са зависном променљивом.
2. Рачунар бира између осталих променљивих ону која највише доприноси тачности предвиђања првој која је изабрана. Овај корак се изводи све док не остане ни једна променљива која би допринела још више тачности модела.
3. При сваком кораку израчунава се тест статистичке значајности за онај ниво предвиђања који додаје нова променљива. Ако је тај ниво предвиђања испод значајности коју је унапред одредио аналитичар, та променљива се искључује из модела.
4. Рачунар даје финални регресиони модел са b коефицијентима. Ако је мултиколинеарност била висока, модел ће имати мање променљивих у односу на оригинални модел.

9. Методе базиране на поретку ранга

9.1 Не-параметарске методе

У деловима 7 и 8 описали смо одређени број метода анализе који су се ослањали на претпоставку да подаци долазе из Нормалне расподеле. Да будемо прецизнији, могли бисмо рећи да подаци долазе из једне од Нормалних породица расподеле, одређене Нормалне расподеле која је у питању и која је дефинисана својом средином и стандардним одступањем, параметрима расподеле. Ови методи се називају **параметарским (parametric)** јер процењујемо параметре основне Нормалне расподеле. За методе које не претпостављају одређену породицу расподела за податке се каже да су **не-параметарске (non-parametric)**. У овом и следећем делу ћемо размотрити неке не-параметарске тестове значајности. Постоје многи други, али ови ће илустровати општи принцип. Ми смо се већ упознали са једним не-параметарским тестом, тестом предзнака (део 6.2). Тест Нормале великог узорка може се такође сматрати не-параметарским тестом.

Корисно је направити разлику између три типа скала мерења.

На **интервалној скали (interval scale)**, величина разлике између две вредности на скали има доследно значење. На пример, разлика у температури између 1°C и 2°C је иста као разлика између 31°C и 32°C . Погодна је за непрекидне (континуалне) податке.

На **номиналној скали (nominal scale)** имамо квалитативне или категоријске променљиве, где су појединци груписани, али не и обавезно уређени. Боја очију је добар пример. Користи се за категоријске податке.

На **ординалној скали (ordinal scale)**, посматрања су уређена, али разлике могу да немају значење. На пример, анксиозност се обично мери коришћењем скупа питања, број позитивних одговора даје скалу анксиозности. Скуп од 36 питања ће дати скалу од 0 до 36. Разлика у анксиозности између резултата 1 и 2, није обавезно иста као разлика између резултата 31 и 32. Користи се када је важан поредак могућих вредности.

9.2 Mann-Whitney U тест

То је не-параметарски аналог t теста два-узорка (део 7.3). Он ради на следећи начин. Размотрите следеће вештачке податке који приказују посматрања променљиве у две независне групе, А и В:

А 7 4 9 17

В 11 6 21 14

Ми желимо да знамо да ли постоји било који доказ да су А и Б извучени из популација са различитим нивоима променљиве. Нулта хипотеза је да не постоји тенденција за чланове једне популације да прекораче чланове друге популације. Алтернатива је да постоји таква тенденција, у једном смеру или другом. Прво уређујемо посматрања у растућем редоследу, односно рангирамо их:

4 6 7 9 11 14 17 21

А В А А В В А В

Сада изаберемо једну групу, рецимо А. За свако А, израчунамо колико В им претходи. За прво А, 4, не претходи ниједно В. За друго А, 7, претходи једно В, за треће А, 9, претходи једно В, за четврто А, 17, претходи три В. Сабирамо ове бројеве претходних В заједно да добијемо $U = 0 + 1 + 1 + 3 = 5$. Сада, ако U је веома мало, скоро сви А су мањи него скоро сви В. Ако је U велико, скоро сви А су већи од скоро свих В. Осредње вредности U значе да су А и В мешовити. Минимално U је 0, када сви В превазилазе све А, а максимално U је $n_1 n_2$ када сви А надмашују све В. Величина U има значење, јер $U/n_1 n_2$ је предвиђање вероватноће да ће посматрање извучено случајно из популације А премашити посматрање извучено случајно из популације В.

Постоји још једна могућа вредност U , коју ћемо звати U' , која је добијена израчунањем колико има А пре сваког В, пре него колико има В пре сваког А. Ово ће бити $1 + 3 + 3 + 4 = 11$. Две могуће вредности U и U' , повезане су једначином $U + U' = n_1 n_2$. Тако да одузимамо U' , од $n_1 n_2$ да добијемо $U = n_1 n_2 - U' = 4 \times 4 - 11 = 5$.

Ако знамо расподелу U , по нултој хипотези да узорци долазе из исте популације, можемо рећи са којом вероватноћом би могли да се појаве ови подаци да није било никакве разлике. Можемо да спроведемо тест значајности. Расподела U по нултој хипотези се може лако наћи. Два скупа од четири посматрања могу се уредити на 70 различитих начина, од AAAABBBB до BBBBAAAA ($8! / 4! 4! = 70$). По нултој хипотези оваква уређења су сва подједнако могућа и, самим тим, имају вероватноћу $1/70$. Свако уређење има своју сопствену вредност U , од 0 до 16, и израчунајући број уређења која дају сваку вредност U можемо наћи вероватноћу те вредности. На пример, $U = 0$ настаје само из редоследа AAAABBBB и тако има вероватноћу од $1/70 = 0.014$. $U = 1$ произилази само из редоследа AAABABBB и тако има вероватноћу од $1/70 = 0.014$. $U = 2$ може да настане на два начина: AAABBBAB и AABAABBB. Оно има вероватноћу од $2/70 = 0.029$. Комплетан скуп вероватноћа је приказан у табели 9.1.

Ово примењујемо на пример. За групе А и В, $U = 5$, а вероватноћа за ово је 0.071. Као што смо урадили за тест предзнака (део 6.2) разматрамо вероватноћу екстремнијих вредности U , $U = 5$ или мање, што даје $0.071 + 0.071 + 0.043 + 0.029 + 0.014 + 0.014 = 0.242$.

Ово даје једностран тест. За двострани тест, морамо размотрити вероватноће као екстремне разлике у супротном смеру. Можемо видети из табеле 9.1 да је расподела U симетрична, тако да је вероватноћа подједнако екстремних вредности у супротном смеру такође 0.242, стога је двострана вероватноћа $0.242 + 0.242 = 0.484$. Тако би посматрана разлика била сасвим вероватна ако је нулта хипотезе тачна, и два узорка би могла да дођу из исте популације.

Табела 9.1 Расподела Mann-Whitney U статистике, за два узорка величине 4

U	Вероватноћа	U	Вероватноћа	U	Вероватноћа
0	0.014	6	0.100	12	0.071
1	0.014	7	0.100	13	0.043
2	0.029	8	0.114	14	0.029
3	0.043	9	0.100	15	0.014
4	0.071	10	0.100	16	0.014
5	0.071	11	0.071		

У пракси, нема потребе да се изврши сабирање вероватноћа горе описаних, јер су оне већ стављене у табелу. Табела 9.2 приказује 5% тачке U за сваку комбинацију величина узорака n_1 и n_2 до 20. За наше групе А и В, $U = 5$. Проналазимо $n_1 = 4$ ред и $n_2 = 4$ колону. Из овога можемо да видимо да 5% тачка за U је 0, и тако $U = 5$ није од значаја. Да смо израчунали већу од две вредности U' , 11, могли смо користити табелу 9.2 да пронађемо нижу вредност, $n_1 n_2 - U' = 16 - 11 = 5$.

Табела 9.2 Двостране 5% тачке за расподелу мање вредности U у Mann-Whitney U тесту

n_1	n_2																		
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	-	-	-	-	-	-	0	0	0	0	1	1	1	1	1	2	2	2	2
3	-	-	-	0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
4	-	-	0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	13

5	-	0	1	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6	-	1	2	3	5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
7	-	1	3	5	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
8	0	2	4	6	8	10	13	15	17	19	22	24	26	29	31	34	36	38	41
9	0	2	4	7	10	12	15	17	20	23	26	28	31	34	37	39	42	45	48
10	0	3	5	8	11	14	17	20	23	26	29	33	36	39	42	45	48	52	55
11	0	3	6	9	13	16	19	23	26	30	33	37	40	44	47	51	55	58	62
12	1	4	7	11	14	18	22	26	29	33	37	41	45	49	53	57	61	65	69
13	1	4	8	12	16	20	24	28	33	37	41	45	50	54	59	63	67	72	76
14	1	5	9	13	17	22	26	31	36	40	45	50	55	59	64	67	74	78	83
15	1	5	10	14	19	24	29	34	39	44	49	54	59	64	70	75	80	85	90
16	1	6	11	15	21	26	31	37	42	47	53	59	64	70	75	81	86	92	98
17	2	6	11	17	22	28	34	39	45	51	57	63	67	75	81	87	93	99	105
18	2	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99	106	112
19	2	7	13	19	25	32	38	45	52	58	65	72	78	85	92	99	106	113	119
20	2	8	13	20	27	34	41	48	55	62	69	76	83	90	98	105	112	119	127

ако је U мање него или једнако табеларној вредности разлика је значајна

Табела 9.3. Поткожно ткиво бицепса (мм) у две групе болесника

Кронова (Crohn's) болест				Целијачна (Coeliac) болест	
1.8	2.8	4.2	6.2	1.8	3.8
2.2	3.2	4.4	6.6	2.0	4.2
2.4	3.6	4.8	7.0	2.0	5.4
2.5	3.8	5.6	10.0	2.0	7.6
2.8	4.0	6.0	10.4	3.0	

Сада се можемо окренути практичној анализи неких стварних података. Размотрићемо податке о поткожном ткиву бицепса из табеле 7.4, репродукованих као табела 9.3. Ми ћемо их анализирати користећи Mann-Whitney U тест. Означимо групу са Кроновом болешћу са А, и групу са целијакијом са В. Здружени редослед је као што следи:

1.8	1.8	2.0	2.0	2.0	2.2	2.4	2.5	2.8	2.8
A	B	B	B	B	A	A	A	A	A
3.0	3.2	3.6	3.8	3.8	4.0	4.2	4.2	4.4	4.8
B	A	A	A	B	A	B	A	A	A
5.4	5.6	6.0	6.2	6.6	7.0	7.6	10.0	10.4	
B	A	A	A	A	A	B	A	A	

Хајде да израчунамо број U пре сваког B . Одмах имамо проблем. Прво A и прво B имају исту вредност. Да ли прво A долази пре првог B или после њега? Ову дилему решавамо рачунајући једну половину везаног A . Везе између другог, трећег и четвртог B нису битне, пошто можемо да израчунамо број A пре сваког B , без тешкоћа. Имамо за U статистику:

$$U = 0.5 + 1 + 1 + 1 + 6 + 8.5 + 10.5 + 13 + 18 = 59.5$$

То је нижа вредност, пошто је $n_1 \times n_2 = 9 \times 20 = 180$, тако да је средња вредност 90. Стога можемо упутити U на табелу 9.2. Критична вредност на нивоу од 5% за групе величине 9 и 20 је 48, што наша вредност превазилази. Стога разлика није значајна на нивоу од 5%, а подаци су у складу са нултом хипотезом да не постоји тенденција за припаднике једне популације да превазиђу припаднике друге популације. Ово је исто као резултат t теста из дела 7.4.

За веће вредности n_1 и n_2 израчунавање U може бити прилично заморно. Једноставна формула за U се може наћи користећи рангове. Ранг најнижег посматрања је 1, ранг следећег посматрања је 2, и тако даље. Ако је одређени број посматрања везан, а свако има исту вредност и самим тим је истог ранга, дајемо сваком посматрању просек рангова које би имали да су уређени. На пример, у подацима о поткожном ткиву, прва два посматрања су свако појединачно 1.8. Свако од њих добија ранг $(1 + 2) / 2 = 1.5$. Треће, четврто и пето посматрање су везани у 2.0, дајући сваком од њих ранг $(3 + 4 + 5) / 3 = 4$. Шесто посматрање, 2.2, није везано па има ранг 6. Рангови за податке о поткожном ткиву су следећи:

поткожно ткиво	1.8	1.8	2.0	2.0	2.0	2.2	2.4	2.5	2.8	2.8
група	A	B	B	B	B	A	A	A	A	A
ранг	1.5	1.5	4	4	4	6	7	8	9.5	9.5
	r_1		r_2	r_3	r_4					
поткожно ткиво	3.0	3.2	3.6	3.8	3.8	4.0	4.2	4.2	4.4	4.8
група	B	A	A	A	B	A	A	B	A	A
ранг	11	12	13	14.5	14.5	16	17.5	17.5	19	20
	r_5			r_6			r_7			
поткожно ткиво	5.4	5.6	6.0	6.2	6.6	7.0	7.6	10.0	10.4	
група	B	A	A	A	A	A	B	A	A	
ранг	21	22	23	24	25	26	27	28	29	
	r_8					r_9				

Означимо рангове групе B са r_1, r_2, \dots, r_{n_1} . Број A који претходи првом B мора бити $r_1 - 1$, пошто не постоји B пре њега и то је r_1 посматрање. Број A који претходи другом B је $r_2 - 2$, пошто је то r_2 посматрање, а једно посматрање које претходи је B . Слично, број који претходи трећем B је $r_3 - 3$, а број које претходи i -том B је $r_i - i$. Стога имамо:

$$U = \sum_{i=1}^{n_1} (r_i - i) = \sum_{i=1}^{n_1} r_i - \sum_{i=1}^{n_1} i = \sum_{i=1}^{n_1} r_i - \frac{n_1(n_1 + 1)}{2}$$

То јест, спајамо заједно рангове свих n_1 посматрања, одузимамо $n_1(n_1 + 1)/2$ и имамо U . На пример, имамо

$$U = 1.5 + 4 + 4 + 4 + 11 + 14.5 + 17.5 + 21 + 27 - \frac{9 \times (9 + 1)}{2} = 104.5 - 45 = 59.5$$

као и пре. Ова формула је понекад написана као

$$U' = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - \sum_{i=1}^{n_1} r_i$$

Али, ово се једноставно заснива на другој групи, пошто је $U + U' = n_1 \times n_2$. За тестирање користимо мању вредност, као и пре.

Док се n_1 и n_2 повећавају, израчунавање тачне расподеле вероватноће постаје све теже. Када не можемо да користимо табелу 9.2, ми користимо апроксимацију великог узорка уместо ње. Зато што је U пронађено сабирањем заједно независних, идентично распоређених случајних променљивих, примењује се централна гранична теорема. Ако је нулта хипотеза тачна, расподела U апроксимира Нормалну расподелу са средином $n_1 n_2 / 2$ и стандардним одступањем $\sqrt{n_1 n_2 (n_1 + n_2 + 1) / 12}$. Стога

$$\frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$$

је посматрање из Стандардизоване Нормалне расподеле. На пример, за $n_1 = 9$ и $n_2 = 20$ имамо

$$\frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}} = \frac{59.5 - \frac{9 \times 20}{2}}{\sqrt{\frac{9 \times 20 \times (9 + 20 + 1)}{12}}} = -1.44$$

Из Табеле 4.1 ово даје двострану вероватноћу = 0.15, сличну оној пронађеној помоћу t теста два узорка (део 7.3).

Нити табела 9.2 нити горња формула за стандардно одступање U не узимају везе у обзир; обе претпостављају да се подаци могу у потпуности рангирати. Њихова употреба за податке са везама је апроксимација. За мале узорке ово морамо прихватити. За Нормалну апроксимацију, везе се могу дозволити уз коришћење следеће формуле за стандардно одступање U када је нулта хипотеза тачна:

$$\sqrt{\frac{n_1 n_2}{(n_1 + n_2)(n_1 + n_2 - 1)} \sum_{i=1}^{n_1 + n_2} r_i^2 - \frac{n_1 n_2 (n_1 + n_2 + 1)^2}{4(n_1 + n_2 - 1)}}$$

где је $\sum_{i=1}^{n_1 + n_2} r_i^2$ збир квадрата рангова за сва посматрања, односно за обе групе (погледајте

Conover 1980). Mann-Whitney U тест није независан од претпоставки које могу бити нарушене. Претпостављамо да подаци морају бити у потпуности уређени, што у случају веза није тако.

Mann-Whitney U тест је не-параметарски аналог t теста два узорка. Предност у односу на t тест је да једина претпоставка о расподели података је да се посматрања могу рангирати, док за t тест морамо претпоставити да су подаци из Нормалне расподеле са униформном варијансом. Постоје недостаци. За податке који су Нормално расподељени, U тест је мање моћан од t теста, односно t тест, када је исправан, може да открије мање разлике за дату величину узорка. U тест је готово исто толико моћан за осредње и велике узорке, а ова разлика је важна само за мале узорке. За врло мале узорке, на пример, две групе од три посматрања, тест је бескористан пошто све могуће вредности U имају вероватноћу изнад 0.05 (табела 9.2). U тест је пре свега тест значајности. t метод нам такође омогућава да проценимо величину разлике и даје интервал поверења. Иако, као што је наведено горе $U / n_1 n_2$ има тумачење, не можемо, колико знамо, пронаћи интервал поверења за то.

Табела 9.4 Учесталост расподеле броја чворова укључених у карцином дојке откривених скринингом и откривених у интервалима између скрининга (подаци Mohammed Raja)

Скрининг тумори		Интервал тумори	
Чворови	Учесталост	Чворови	Учесталост
0	291	0	66
1	43	1	22
2	16	2	7
3	20	3	7
4	13	4	2
5	3	5	4
6	1	6	4
7	4	7	3
8	3	8	3
9	1	9	2
10	1	10	2
11	2	12	2
12	1	13	1
15	1	15	1
16	1	16	1
17	2	20	1
18	2		
20	1		

27	1	
33	1	
Укупно	408	128
Средина	1.21	2.19
Медијана	0	0
75% болест	1	3

Man - Whitney U тест

$U = 31555.00$, $SE = 1281.33$

$$z = \frac{U - n_1 n_2 / 2}{SE} = \frac{31555.00 - 408 \times 128 / 2}{1281.33} = 4.25$$

$P < 0.0001$

Нулта хипотеза Mann-Whitney U теста је понекад представљена као да популације имају исту медијану. Постоји чак интервал поверења за разлику између две медијане на основу Mann-Whitney U теста (Campbell и Gardner 1989). То је изненађујуће, јер медијане нису укључене у израчунавање. Осим тога, можемо имати две групе које се значајно разликују и које коришћењем Mann-Whitney U теста још увек имају исту медијану. Табела 9.4 показује пример. Већина посматрања у обе групе су нуле, тако да трансформација на Нормалну није могућа. Мада су узорци прилично велики, расподела је толико искривљена да метод рангирања, на одговарајући начин прилагођен за везе, може бити сигурнији него метод из дела 6.7. Mann-Whitney U тест је био веома значајан, пошто су обе медијане нула. Пошто су медијане биле једнаке, предложио сам 75 процената као меру положаја за расподелу.

Разлог за ова два различита погледа Mann-Whitney U теста лежи у претпоставци коју правимо у вези расподеле у две популације. Ако не правимо претпоставке, можемо тестирати нулту хипотезу: да је вероватноћа да ће члан прве популације изабран случајно прекорачити члана друге популације изабраног случајно, једна половина. Неки људи бирају да праве претпоставку о расподелама: да имају исти облик и разликују се само у положају (средина или медијана). Ако је ова претпоставка тачна, онда ако су расподеле различите медијане морају бити различите. Средина мора да се разликује за исти износ. То је веома јака претпоставка. На пример, ако је ово тачно тада варијансе морају бити исте у две популације. Из разлога наведених у делу 7.5, мало је вероватно да бисмо добили ово ако расподела није била Нормална. Под овом претпоставком Mann-Whitney U теста ће ретко бити важећи ако t тест два узорка није такође валидан.

Постоје и други не-параметарски тестови који тестирају исте или сличне нулте хипотезе. Два од ових тестова, Wilcoxon-ов тест упарених парова и Kendall Tau тест, су различите верзије Mann-Whitney U теста који је био развијен отприлике у исто време, и касније је показано да је идентичан. Ови називи се понекад користе као синоними. Тест статистике и табеле нису исте, и корисник мора бити веома опрезан да израчунавање тест статистике која се користи одговара табели на коју се односи. Још једна потешкоћа са табелама је да су неке тако направљене да за значајну разлику, U мора бити мање или једнако табеларној вредности (као у табели 9.2), за друге табеле U мора бити стриктно мање од табеларне вредности.

За више од две групе, за анализу рангова користи се Kruskal-Wallis-ов тест, погледајте Conover (1980) и Siegel (1956).

9.3 Wilcoxon-ов тест упарених (еквивалентних) парова

Овај тест је аналог t теста за везане узорке.

Табела 9.5 Резултати испитивања пронеталола за превенцију ангине пекторис (Pritchard *et al.* 1963), по реду рангова разлика

Број напада за		Разлика плацебо – пронеталол	Ранг разлике		
Плацебо	Пронеталол		Сви	Позитивно	Негативно
2	0	2	1.5	1.5	
17	15	2	1.5	1.5	
3	0	3	3	3	
7	2	5	4	4	
8	1	7	6	6	
14	7	7	6	6	
23	16	7	6	6	
34	25	9	8	8	
79	65	14	9	9	
60	41	19	10	10	
323	348	-25	11		11
71	29	42	12	12	
Сума рангова				67	11

Имамо узорак измерен под два услова, а нулта хипотеза је да не постоји тенденција да исход под једним условом буде већи или мањи од исхода под другим условом. Алтернативна хипотеза је да исход под једним условом, тежи да буде већи или мањи од другог исхода. Како се тест заснива на величини разлика, подаци морају бити интервал.

Размотрићемо податке из табеле 9.5, о којима је претходно било речи у делу 6.2, где смо користили тест предзнака за анализу. У тесту предзнака, игнорисали смо величину разлика, и узели смо у обзир само њихове знаке. Да можемо користити информације о величини, надали бисмо се снажнијем тесту. Несумњиво, морамо имати податке о интервалу да би то урадили. Да би избегли прављење претпоставки о расподели разлика, користимо њихов ред рангова на сличан начин као Mann–Whitney U тест.

Табела 9.6 Двостране 5% и 1% тачке за расподелу T (ниже вредности) у Wilcoxon-овом тесту једног узорка

Величина узорка n	Вероватноћа да је $T \leq$ табеларној вредности		Величина узорка n	Вероватноћа да је $T \leq$ табеларној вредности	
	5%	1%		5%	1%
5	-	-	16	30	19
6	1	-	17	35	23
7	2	-	18	40	28
8	4	0	19	46	32
9	6	2	20	52	37
10	8	3	21	59	43
11	11	5	22	66	49
12	14	7	23	73	55
13	17	10	24	81	61
14	21	13	25	90	68
15	25	16			

Прво, рангирамо разлике по њиховим апсолутним вредностима, тј. игноришући знак. Као и у делу 9.2, везана посматрања дају просечне вредности њихових рангова. Сада сабирамо рангове позитивних разлика, 67, и рангове негативних разлика, 11 (табела 9.5). Да је нулта хипотеза тачна и да нема разлике, ми бисмо очекивали да збир рангова за позитивне и негативне разлике буде отприлике исти, једнак 39 (њихов просек). Тест статистика је мања од ове две суме, T . Што је T мање, то је нижа вероватноћа да подаци настану случајно.

Расподела T када је нулта хипотеза тачна може се пронаћи набрајањем свих могућности, као што је описано за Mann–Whitney U статистику. Табела 9.6 даје 5% и 1% тачке за ову расподелу, за величину узорка n све до величине 25. На пример, $n = 12$ и тако би разлика била значајна на 5% нивоу да је T мање од или једнако 14. Имамо да је $T = 11$, тако да подаци нису у складу са нултом хипотезом. Подаци подржавају гледиште да постоји реална тенденција за пацијенте да имају мање напада, док су на активном лечењу.

Из табеле 9.6, можемо видети да вероватноћа да је $T \leq 11$ лежи између 0.05 и 0.01. Ово је већа вероватноћа од оне која је дата тестом предзнака, која је била 0.006 (део 6.2). Обично бисмо очекивали већу снагу, а тиме и ниже вероватноће када је нулта хипотеза лажна, када користимо више информација. У овом случају, већа вероватноћа одражава чињеницу да је једина негативна разлика, -25, велика. Прегледање оригиналних података приказује да је овај

појединац имао велики број напада у оба лечења, и чини се могућим да он можда припада различитој популацији од других једанаест.

Као и табела 9.2, табела 9.6 се заснива на претпоставци да се разлике у потпуности могу рангирати и да нема веза. Везе се у овом тесту могу појавити на два начина. Прво, веза се може јавити у смислу рангирања. У примеру смо имали две разлике од +2 и три од +7. Оне су биле једнако рангиране: 1.5 и 1.5, и 6, 6 и 6. Када су везе присутне између негативних и позитивних разлика, табела 9.6 само апроксимира расподелу T .

До веза такође може доћи између везаних посматрања, где је примећена разлика нула. На исти начин као и за тест предзнака, изостављамо нула разлике (део 6.2). Табела 9.6 се користи са n као бројем само за разлике без-нуле, а не за укупни број разлика. Ово изгледа чудно, због тога што изгледа да много нула разлика подржава нулту хипотезу. На пример, да смо у табели 9.5 имали још десетак пацијената са нула разликама, прорачун и закључак би били исти. Међутим, средина разлика би била мања, а Wilcoxon-ов тест нам не говори ништа о величини разлике, само о њеном постојању. Ово илустрuje опасност од допуштања тестовима значајности да надмаше све остале начине посматрања података.

Како n расте, расподела T по нултој хипотези тежи Нормалној расподели, као што то чини и она из Mann–Whitney U статистике. Збир свих рангова, независно од знака, је $n(n+1)/2$, тако да очекивана вредност T по нултој хипотези је $n(n+1)/4$, пошто би две суме требало да буду једнаке. Ако је нулта хипотеза тачна, стандардно одступање од T је $\sqrt{1/4 \sum r_i^2}$, где r_i је ранг i -те разлике, која је $\sqrt{n(n+1)(2n+1)/24}$, када нема веза. Стога

$$\frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

је из Стандардне Нормалне расподеле, ако је нулта хипотеза тачна. За пример из табеле 9.5, имамо

$$\frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} = \frac{11 - \frac{12 \times 13}{4}}{\sqrt{\frac{12 \times 13 \times 25}{24}}} = -2.197$$

Из табеле 4.2 ово даје дво-страну вероватноћу од 0.028, сличну оној добијеној из табеле 9.6. Имамо три могућа теста за везане податке, Wilcoxon-ов, тест предзнака и t методе за везане узорке. Ако су разлике Нормално расподељене, t тест је најмоћнији тест. Међутим Wilcoxon-ов тест је готово исто толико моћан, и у пракси разлика није велика, осим за мале узорке. Као и Mann–Whitney U тест, Wilcoxon-ов тест је бескористан за веома мале узорке. Тест предзнака је по снази сличан Wilcoxon-овом за веома мале узорке, али како се величина узорка повећава Wilcoxon-ов тест постаје много јачи. Wilcoxon-ов тест користи величину разлика, и стога захтева податке о интервалима. То значи да ћемо, што се тиче t метода, добити различите резултате ако трансформишемо податке. За заиста редне податке треба користити тест предзнака. t метод за везане узорке такође даје интервал поверења за разлику.

9.4 Spearman-ов коефицијент корелације ранга, ρ

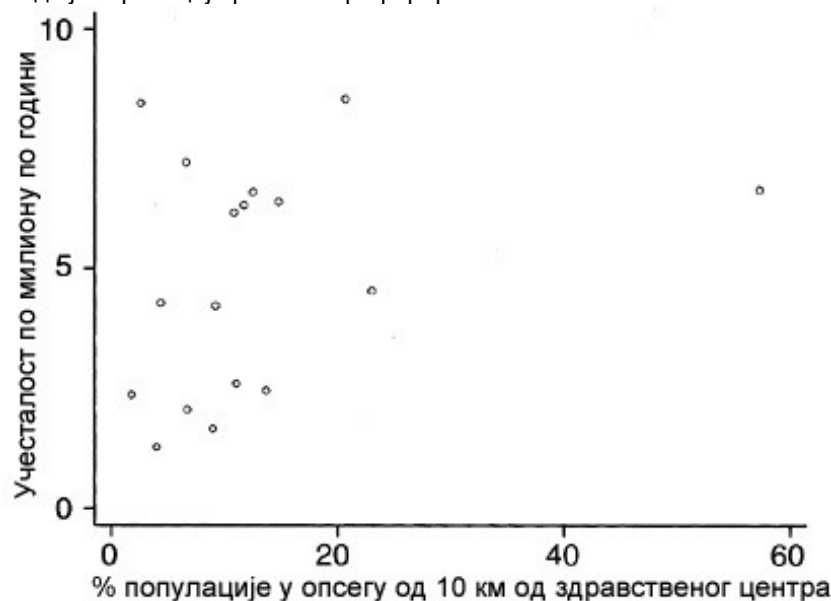
Забележили смо у делу 8 осетљивост на претпоставке о Нормалности производа момента коефицијента корелације, r . То је довело до развоја не-параметарских приступа заснованих на ранговима. Spearman-ов приступ је био директан. Прво рангирамо посматрања, а затим израчунавамо производ момента корелације рангова, пре него за сама посматрања. Резултујућа статистика има расподелу која не зависи од расподеле оригиналних променљивих. Обично се означава грчким словом ρ , изговара се “ро”.

Табела 9.7 приказује податке из студије о географској расподели тумора, Капошијевог саркома (*Kaposi's sarcoma*), у копненој Танзанији. Стопе учесталости су израчунате на основу података о регистрованом раку и било је значајне сумње да ли су сви случајеви били регистровани. Могуће је да је степен извештавања о случајевима имао везе са густином популације или доступношћу здравствених услуга.

Табела 9.7 Учесталост Капошијевог саркома и приступ становништва здравственим центрима за сваки регион копнене Танзаније (Bland *et al.* 1977)

Регион	Учесталост по милиону по години	Проценат популације унутар 10 км од здравственог центра	Редослед рангова	
			Учесталост %	Популација %
Coast	1.28	4.0	1	3
Shinyanga	1.66	9.0	2	7
Mbeya	2.06	6.7	3	6
Tabora	2.37	1.8	4	1
Arusha	2.46	13.7	5	13
Dodoma	2.60	11.1	6	10
Kigoma	4.22	9.2	7	8
Mara	4.29	4.4	8	4
Tanga	4.54	23.0	9	16
Singida	6.17	10.8	10	9
Morogoro	6.33	11.7	11	11
Mtwara	6.40	14.8	12	14
Westlake	6.60	12.5	13	12
Kilimanjaro	6.65	57.3	14	17
Ruvuma	7.21	6.6	15	5
Iringa	8.46	2.6	16	2
Mwanza	8.54	20.7	17	15

Поред тога, учесталост је била у блиској вези са годиштем и полом (тамо где је забележена) и тако је могла да се доведе у везу са расподелом годишта и пола у региону. Да би проверили да ништа од овога не ствара лажну слику у географској расподели, израчунали смо корелацију рангова учесталости болести са сваком од могућих променљивих које објашњавају. Табела 9.7 приказује однос учесталости у односу на проценат популације која живи у опсегу од 10 км од здравственог центра. Слика 9.1 приказује дијаграм растурања ових података. Проценат у опсегу од 10 км од здравственог центра је веома искошен, док учесталост болести изгледа донекле двомодална. Претпоставка о корелацији производа момента изгледа да није испуњена, тако да је корелација рангова преферирана.



Слика 9.1 Учесталост Капошијевог саркома по милиону по години и проценат популације у опсегу од 10 км од здравственог центра, за 17 подручја копнене Танзаније

Прорачун Спирмановог ρ се изводи као што следи. Проналазе се рангови за две променљиве (табела 9.7). Примењујемо формулу за корелацију производа момента на ове рангове. Дефинишемо:

$$\rho = \frac{\text{сума производа око средине рангова}}{\sqrt{\text{сума квадрата рангова за прву променљиву} \times \text{сума квадрата рангова за другу променљиву}}}$$

Прорачун је изведен као што је описано у делу 8.5 и он даје да је $\rho = 0.38$. Сада можемо тестирати нулту хипотезу да су променљиве независне, уз алтернативу да или се једна променљива повећава док се и друга повећава, или да се једна смањује док се друга повећава. Као и обично са статистиком рангова, расподела ρ за мале узорке се може пронаћи набрајањем свих могућих пермутација и њихових вредности ρ . За узорак величине n постоји, наравно, $n!$ могућности.

Табела 9.8 приказује критичну вредност ρ за величине узорака све до 10. Како се n повећава, тако ρ тежи Нормалној расподели, када је нулта хипотеза тачна, са очекиваном вредношћу 0 и варијансом $1/(n-1)$. Тако $\rho / \sqrt{1/(n-1)} = \rho \sqrt{n-1}$ долази из Стандардизоване Нормалне расподеле. Апроксимација је разумна за $n > 10$.

За наше податке имамо да је $0.38 \sqrt{17-1} = 1.52$, што из табеле 4.1, има двострану вероватноћу од 0.13. Дакле нисмо пронашли никакве доказе о постојању односа између запажене учесталости Капошијевог саркома и приступа здравственим центрима. У овој студији није било значајног односа са било којом од могућих променљивих које дају објашњење и закључили смо да се није испоставило да је посматрана географска расподела лажна слика расподеле становништва или дијагностичких одредби.

Табела 9.8 Двостране 5% и 1% тачке расподеле Spearman-овог ρ

Величина узорка n	Вероватноћа да је ρ далеко или удаљеније од 0 него табеларна вредност	
	5%	1%
4	-	-
5	1.00	-
6	0.89	1.00
7	0.82	0.96
8	0.79	0.93
9	0.70	0.83
10	0.68	0.81

9.5 Kendall-ов коефицијент корелације ранга, τ

Spearman-ова корелација рангова је сасвим задовољавајућа за тестирање нулте хипотезе да нема повезаности, али је тешка за тумачење као мера јачине повезаности. Kendall је развио различит коефицијент корелације ранга. Kendall-ово τ , има неке предности у односу на Spearman-ово ρ (грчко слово τ се изговара "тау"). Прилично је напорно да се израчуна Kendall-ово τ у поређењу са Spearman-овим коефицијентом, али у доба рачунара то једва да је битно. За сваки пар субјеката посматрамо да ли су субјекти поређани на исти начин помоћу две променљиве, **складан (concordant)** пар, поређани у супротним смеровима, **нескладан (discordant)** пар, или једнаки за једну од променљивих и ако уопште нису поређани, **везани (tied)** пар. Кендалово τ је пропорција складних парова минус пропорција нескладних парова. τ ће бити један, ако су рангови идентични, пошто ће сви парови бити уређени на исти начин, а минус један ако су рангови управо супротни, пошто ће сви парови бити уређени у супротном смеру.

Означимо број складних парова (уређених на исти начин) са n_c , број нескладних парова (уређених у супротним смеровима) са n_d , и разлику $n_c - n_d$, са S . Има $n(n-1)/2$ парова укупно, тако да је

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)} = \frac{S}{\frac{1}{2}n(n-1)}$$

Када нема веза, $n_c + n_d = n(n-1)/2$.

Најједноставнији начин за израчунавање n_c је уређивањем посматрања помоћу једне од променљивих, као у табели 9.7 која је уређена учесталошћу болести. Сада размотрите друго уређење рангова (% популације у опсегу од 10 км од здравственог центра). Прва регија, Coast, има 14 регија испод ње које имају већи ранг, тако да ће парови бити формирани помоћу прве регије и они ће бити у правом редоследу. Постоје 2 регије испод ње које су нижег ранга, тако да ће парови бити формирани помоћу прве регије и они ће бити у супротном редоследу. Друга регија, Shinyanga, има 10 регија испод ње са већим рангом и тако придодaje 10 парова више у исправном редоследу. Обратите пажњу на то да су пар "Coast и Shinyanga" већ израчунати.

Има 5 парова у супротном редоследу. Трећа регија, Мбеуа, има 10 регија испод ње у истом редоследу и 4 у супротном редоследу, и тако даље. Додајемо ове бројеве да би добили n_c and n_d :

$$n_c = 14 + 10 + 10 + 13 + 4 + 6 + 7 + 8 + 1 + 5 + 4 + 2 + 2 + 0 + 1 + 1 + 0 = 88$$

$$n_d = 2 + 5 + 4 + 0 + 8 + 5 + 3 + 1 + 7 + 2 + 2 + 3 + 2 + 3 + 1 + 0 + 0 = 48$$

Број парова је

$$\frac{n(n-1)}{2} = \frac{17 \times 16}{2} = 136$$

Зато што нема веза, могли бисмо такође израчунати n_d помоћу

$$n_d = \frac{n(n-1)}{2} - n_c = 136 - 88 = 48$$

$$S = n_c - n_d = 88 - 48 = 40.$$

Стога је

$$\tau = \frac{S}{n(n-1)/2} = \frac{40}{136} = 0.29.$$

Када постоје везе, τ не може да буде један. Међутим, имали бисмо савршену корелацију да су везе биле између истих испитаника за обе променљиве. Да би ово дозволили, користимо неку другу верзију τ , τ_b . Размотрите именилац. Има $n(n-1)/2$ могућих парова. Ако постоје t појединци везани за одређени ранг променљиве X , ниједан пар од ових t појединаца не доприноси S . Постоји $t(t-1)/2$ таквих парова. Ако узмемо у обзир све групе повезаних појединаца имамо $\sum t(t-1)/2$ парова који не доприносе S , сабирајући све групе повезаних рангова. Стога укупан број парова који могу да доприносе S је $n(n-1) - \sum t(t-1)/2$, и S не може бити веће од $n(n-1)/2 - \sum t(t-1)/2$. Величина S је такође ограничена везама у другом уређењу рангова. Ако број појединаца са истом вредношћу Y означимо са u , тада број парова који могу да доприносе S је $n(n-1)/2 - \sum u(u-1)/2$. Сада дефинишемо τ_b помоћу

$$\tau_b = \frac{S}{\sqrt{\left(\frac{n(n-1)}{2} - \sum \frac{t(t-1)}{2}\right) \left(\frac{n(n-1)}{2} - \sum \frac{u(u-1)}{2}\right)}}$$

Обратите пажњу да ако нема веза, $\sum t(t-1)/2 = 0 = \sum$ Када су рангови идентични $\tau_b = 1$, без обзира на то колико веза има. Kendall (1970) такође расправља о друга два начина која се баве везама, добијајући коефицијенте τ_a и τ_c , али њихова употреба је ограничена.

Често желимо да тестирамо нулту хипотезу да не постоји однос између две променљиве у популацији из које је наш узорак извучен. Као и обично, нас интересује вероватноћа S која је екстремна као или екстремнија (тј. далеко од нуле) од посматране вредности. Табела 9.9 израчуната је на исти начин као и табеле 9.1 и 9.2. Она приказује вероватноћу да буде екстремна као и посматрана вредност S за n све до 10. Због практичности, S је табелирано пре него τ . Када су везе присутне ово је само апроксимација.

Када је величина узорка већа од 10, S има приближно Нормалну расподелу по нултој хипотези, са средином нула. Ако нема веза, варијанса је

$$VAR(S) = \frac{n(n-1)(2n+5)}{18}$$

Када везе постоје, формула варијансе је врло компликована (Kendall 1970). Ми ћемо је изоставити, пошто би у пракси ова израчунања свакако била урађена помоћу рачунара. Ако нема много веза то неће направити много разлике, ако се користи једноставна форма. На пример, $S = 40$, $n = 17$ и нема веза, тако да је Стандардно Нормално одступање

$$\frac{S}{\sqrt{VAR(S)}} = \frac{S}{\sqrt{n(n-1)(2n+5)/18}} = \frac{40}{\sqrt{17 \times 16 \times 39/18}} = 1.55$$

Табела 9.9 Двостране 5% и 1% тачке расподеле S за Kendall-ово τ

Величина узорка n	Вероватноћа да је S далеко или удаљеније од 0 него табеларна вредност	
	5%	1%
4	-	-
5	10	-
6	13	15
7	15	19
8	18	22
9	20	26
10	23	29

Из табеле 4.2 Нормалне расподеле налазимо да је двострана вероватноћа вредности екстремне као што је ова $0.06 \times 2 = 0.12$, која је врло слична оној пронађеној помоћу Spearman-овог ρ . Корелација производа момента r , даје $r = 0.30$, $P = 0.24$, али наравно расподела променљивих различита од Нормалне чини ово P неважећим.

Зашто имати два различита коефицијента корелације ранга? Spearman-ово ρ је старије од Kendall-овог τ , и може се посматрати као једноставни аналог коефицијента корелације производа момента, Pearson-овог r . τ је део општијег и конзистентнијег система метода рангова, и има директну интерпретацију, као разлика између пропорција складних и нескладних парова. У принципу, нумеричка вредност од ρ је већа од τ . Није могуће израчунати τ из ρ , или ρ из τ , они мере различите врсте корелација. ρ даје више тежине преокретању редоследа када су подаци више удаљени у рангу него онда када постоји преокретање која је ближе заједно у рангу, а τ то не чини. Међутим, у погледу тестова значајности оба имају исту снагу да одбију лажну нулту хипотезу, тако да за ову сврху није важно које се користи.

9.6 Исправке континуитета

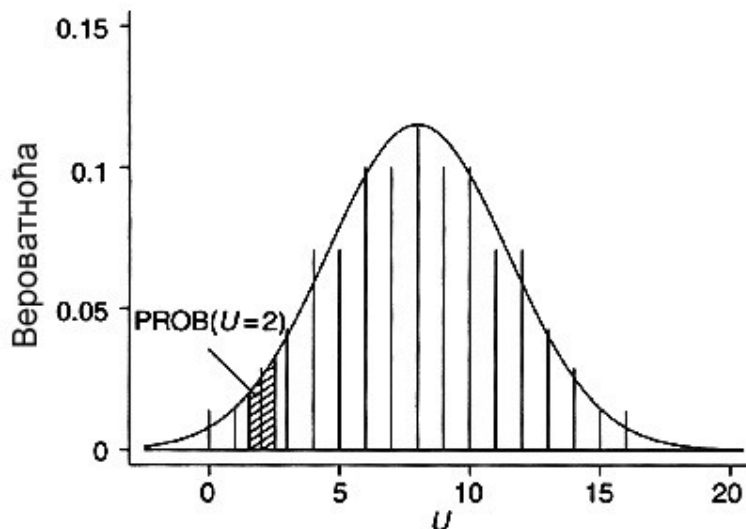
У овом делу, када су узорци били велики користили смо непрекидну расподелу, Нормалну, да апроксимирамо дискретну расподелу, U , T или S . На пример, слика 9.2 приказује расподелу Mann-Whitney U статистике за $n_1 = 4$, $n_2 = 4$ (Табела 9.1) са одговарајућом Нормалном кривом.

Од тачне расподеле, вероватноћа да је $U < 2$ је $0.014 + 0.014 + 0.029 = 0.057$. Одговарајуће Стандардно Нормално одступање је

$$\frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}} = \frac{2 - \frac{4 \times 4}{2}}{\sqrt{\frac{4 \times 4 \times 9}{12}}} = -1.732$$

Оно има вероватноћу од 0.048, интерполирањем у табели 4.1. Ово је мање од тачне вероватноће која је 0.057. Неподударане настаје због тога што непрекидна расподела даје вероватноћу вредностима које нису цели бројеви 0, 1, 2, итд. Процењена вероватноћа за $U = 2$ се може наћи помоћу површине испод криве између $U = 1.5$ и $U = 2.5$. Одговарајућа Нормална

одступања су -1.876 и -1.588 , која имају вероватноће из табеле 4.1 од 0.030 и 0.056 . Ово даје процењену вероватноћу за $U = 2$ од $0.056 - 0.030 = 0.026$, која се упоређује прилично добро са тачном цифром од 0.029 (Табела 9.1). Стога да би проценили вероватноћу да је $U < 2$, проценили смо део испод $U = 1.5$, а не испод $U = 2$. То нам даје Стандардно Нормално одступање од -1.588 , као што је већ речено, а тиме и вероватноћу од 0.056 . Ово изузетно добро одговара тачној вероватноћи од 0.057 , посебно када узмемо у обзир колико су мали n_1 и n_2 .



Слика 9.2 Расподела Mann-Whitney U статистике, $n_1 = 4$, $n_2 = 4$, када је нулта хипотеза тачна, са одговарајућом Нормалном расподелом и проценом површине $\text{PROB}(U = 2)$

Добићемо бољу апроксимацију од нашег Стандардног Нормалног одступања ако приближимо U његовој очекиваној вредности преко $1/2$. У принципу, добијамо боље уклапање ако учинимо посматране вредности статистике ближим њиховој очекиваној вредности помоћу пола интервала између суседних дискретних вредности. Ово је **корекција континуитета (continuity correction)**.

За S , интервал између суседних вредности је 2, а не 1, за $S = n_c - n_d = 2n_c - n(n-1)/2$, и n_c је цео број. Промена једне јединице у n_c производи промену две јединице у S . Корекција континуитета је, дакле, половина од 2, што је 1. Приближавамо S очекиваној вредности 0 преко 1 пре примене Нормалне апроксимације. За податке о Капошијевом саркому, имали смо да је $S = 40$, са $n = 17$. Коришћењем корекције континуитета добијамо

$$\frac{S-1}{\sqrt{\text{VAR}(S)}} = \frac{40-1}{\sqrt{17 \times 18 \times 39 / 18}} = \frac{39}{25.75} = 1.513$$

Ово даје двострану вероватноћу $0.066 \times 2 = 0.13$, мало већу од неисправљене вредности 0.12 . Корекције континуитета су од значаја за мале узорке; за велике узорке оне су занемарљиве. Упознаћемо се са тим још једном у делу 10.

9.7 Параметарске или не-параметарске методе?

За многе статистичке проблеме постоји неколико могућих решења, као што и за многе болести постоји неколико могућности за лечење, сличних можда у њиховој свеукупној ефикасности, али уз варијације у њиховим нежељеним дејствима, у њиховој интеракцији са другим болестима или лечењима и у њиховој подобности за различите типове пацијената. Ту често не постоји један прави третман, већ се често третман бира на основу процене ових ефеката од стране оног који третман преписује, на основу прошлог искуства и обичне предрасуде. Многи проблеми у статистичкој анализи су попут овог. У поређењу средине две мале групе, на пример, могли бисмо да користимо t тест, t тест са трансформацијом, Mann-Whitney U тест, или један од неколико других тестова. Наш избор метода зависи од уверљивости Нормалних претпоставки,

важности добијања интервала поверења, једноставности израчунања, и тако даље. Такође зависи од обичне предрасуде. Неки корисници статистичких метода су веома забринуте због импликација Нормалних претпоставки и залагају се за не-параметарске методе где год је то могуће, док су други превише необазриви на грешке које могу бити уведене када се претпоставке не задовоље.

Постоји распрострањено погрешно схватање да, када је број посматрања веома мали, обично мањи од шест, методе Нормалне расподеле као што су t тестови и регресија се не смеју користити и да методе рангова треба користити уместо њих. Никада није приказано да је било који аргумент изнешан у прилог овоме, и преглед табела 9.2, 9.6, 9.8, и 9.9 ће показати да је то бесмислица. За такве мале узорке тестови рангова не могу бити од било какве важности на уобичајеном 5% нивоу. Ако неком затребају статистичке анализе тако малих узорака, онда се захтевају Нормалне методе.