

ПРЕДМЕТ

< МЕДИЦИНСКА СТАТИСТИКА И ИНФОРМАТИКА >

Предавање број 15

**<** **РЕГРЕСИЈА И КОРЕЛАЦИЈА >**

|  |  |  |  |
| --- | --- | --- | --- |
| Недеља | Наставна јединица | Тематске јединице | Резултат – знања или вештине које студент треба да добије |
| 15 | Регресија и корелација. | Дијаграми растурања. Регресија. Метода најмањих квадрата. Стандардна грешка коефицијента регресије. Корелација. Значај теста и интервал поверења за r. Коришћење коефицијента корелације. | Упознавање са регресијом и корелацијом. |

Copyright © 2012 – Факултет медицинских наука Универзитета у Крагујевцу. Сва права задржана. Без претходне писмене дозволе од стране Факултета медицинских наука забрањена је репродукција, трансфер, дистрибуција или меморисање неког дела или читавих садржаја овог документа, копирањем, снимањем, електронским путем, скенирањем или на било који други начин.

Copyright © 2012 – Faculty of Medical Sciences of University of Kragujevac. All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying,, recording, scanning or otherwise, without the prior written permission of Faculty of Medical Sciences.

**САДРЖАЈ**

[8. Регресија и корелација 2](#_Toc291506715)

[8.1 Дијаграми растурања 2](#_Toc291506716)

[8.2 Регресија 4](#_Toc291506717)

[8.3 Метода најмањих квадрата 4](#_Toc291506718)

[8.4 Стандардна грешка коефицијента регресије 7](#_Toc291506719)

[8.5 Корелација 8](#_Toc291506720)

[8.6 Коришћење коефицијента корелације 10](#_Toc291506721)

Предавање бр. 15

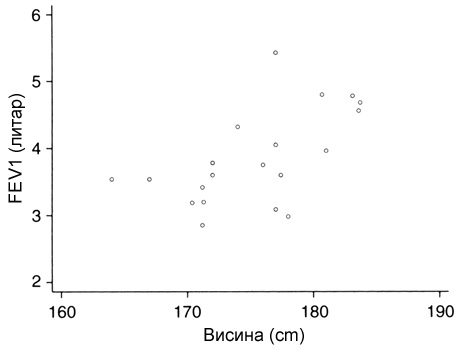
**<** **РЕГРЕСИЈА И КОРЕЛАЦИЈА >**

## 8. Регресија и корелација

### 8.1 Дијаграми растурања

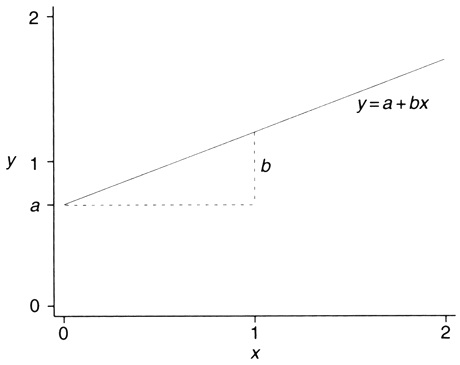
У овом делу ћемо погледати методе анализирања односа између две квантитативне променљиве. Размотрите табелу 8.1, која приказује податке прикупљене од стране групе студената медицине на часу физиологије. Преглед података указује на то да је могуће да постоји неки однос између FEV1 и висине. Пре покушаја 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 (cm) | 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 |  |  | |



Слика 8.1 Дијаграм растурања који приказује однос између FEV1 и висине за групу мушких студената медицине

Слика 8.1 приказује дијаграм растурања за FEV1 и висину. Преглед слике 8.1 указује да се FEV1 повећава са висином. Следећи корак је покушати и повући линију која најбоље представља однос. Најједноставнија линија је права линија.



Слика 8.2 Коефицијенти праве линије

Једнакост праве линије односа између променљивих *x* и *y* је *y* = *а* + b*x*, где су *а* и *b* константе. Прва константа, *а*, се зове **одсечак** (**intercept**). То је вредност *y* када *x* је 0. Друга константа, *b*, се зове **нагиб** (**slope**) или **градијент** (**gradient**) линије. То је повећање *y* које одговара повећању једне јединице *x*. Њихово геометријско значење је приказано на слици 8.2. Можемо наћи вредности *а* и b који се најбоље уклапају у податке регресионе анализе.

### 8.2 Регресија

**Регресија** (**regression**) је метода предвиђања нумеричког односа између променљивих. На пример, желимо да знамо која је средина или очекивана вредност FEV1 за студенте дате висине, и које повећање у FEV1 је повезано са јединицом повећања у висини.

Име ‘‘регресија’‘ је настало захваљујући Galton-у (1886), који је развио технику да се испита однос између висине деце и њихових родитеља. Oн је приметио да ако изаберемо групу родитеља дате висине, средина висина њихове деце ће бити ближа средини висине популације него што је дата висина родитеља. Другим речима, високи родитељи имају тенденцију да буду 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



где су *а* и *b* константе, а *Е* 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 *Е* није била нула, могли бисмо то учинити променом *а*. Претпостављамо да је *Е* 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 *а* 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. Oва одступања представљају грешку *Е* 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 *Е* минималном. Oво 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. Oво ће вероватно бити случај, пошто регресија има тенденцију да уклони из *Y* варијабилност између субјеката и остави грешку мерења, која ће вероватно бити Нормална.

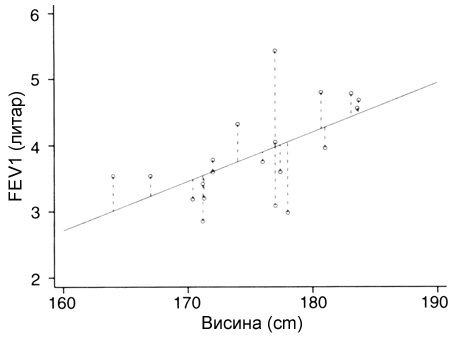
Многи корисници статистике су збуњени минимизирањем варијације само у једном смеру. Oбично су обе променљиве мерене са неким грешкама па опет изгледа да игноришемо грешку у *X*. Зашто не би смањили нормална одстојања до линије пре него вертикална одстојања? Постоје два разлога за то. Прво, ми налазимо 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

FEV1 (литар) = -9.33 + 0.075 x висина (цм)

Ако меримо висину у метрима уместо у центиметрима, добијамо

FEV1 (литар) = -34.70 + 22.0 x 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. Oво је очигледно незадовољавајуће и нећемо даље узимати у обзир овакав приступ.



Слика 8.3 Одступање од линије у правцу''*y*''

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



Затим проналазимо одсечак *а* преко



Приметите да линија мора да прође кроз тачку средине, (,). 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 regresion**).

Једначина *Y = а + b X* 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 **коефицијент** **регресије** (**regression coefficient**). Ми ћемо га израчунати за податке из табеле 8.1. Имамо да је







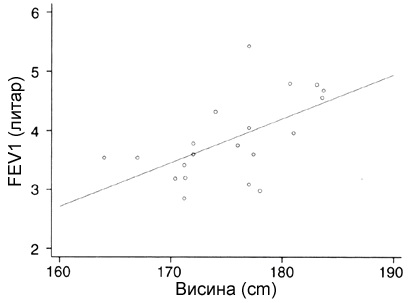
Не треба нам сума квадрата за *Y* још увек, али биће нам потребна касније.





Oтуда је једначина регресије FEV1 за висину

FEV (литар) = -9.19 + 0.0744 x висина (цм)

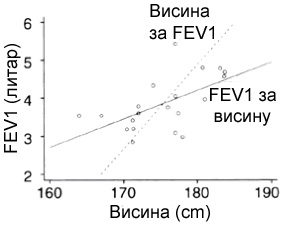


Слика 8.4 Регресија FEV1 за висину

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

FEV1 (литар) = -9.19 + 7.44 x висина (m)

Oво 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

висина = 158 + 4.54 x FEV1

Ово није исто као линија регресије FEV1 на висину. Ако преуредимо ову једначину дељењем сваке стране са 4.54 добијамо

FEV1 = -34.8 + 0.220 x висина

Нагиб регресиje висине на FEV1 је већи него FEV1 на висину (слика 8.5). У принципу, нагиб регресиje 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 *yi* и вредности предвиђених регресионом линијом. Oво је



је наравно, укупан збир квадрата око средине *yi*. Појам 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, *а* и *b*. Губимо два степена слободе, остављајући нас са . Стога је варијанса *Y* око линије, звана **резидуална варијанса** (**residual variance**)



Да бисмо предвидели варијацију око линије, морамо претпоставити да је иста све до краја линије, односно да је варијанса униформна. Oво је исто као и за *t* метод два узорка (део 7.3) и анализу варијансе. За FEV1 податке збир квадрата услед регресије је , а збир квадрата око регресије је . Има 20 – 2 = 18 степени слободе, тако да је варијанса око регресије. Стандардна грешка од *b* је дата помоћу

 литара/цм

Већ смо претпоставили да је грешка *Е* 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  до  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, односа 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,

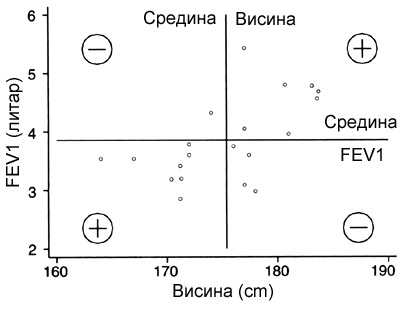


Из табеле 7.1 ово има двострану вероватноћу мању од 0.01. Рачунар нам говори да је вероватноћа око 0.007. Стога подаци су у супротности са нултом хипотезом и подаци пружају прилично добар доказ да постоји веза. Да је узорак био много већи, могли бисмо да га изделимо *t* расподелом и употребимо Стандардизовану 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*. Oво нам даје **производ момента коефицијента корелације** (**product moment correlation coefficient**), или укратко **коефицијент** **корелације** (**correlation coefficient**), обично означен са *r.*

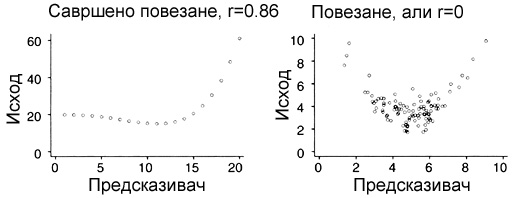
Ако су *n* парова посматрања обележени са (*xi*, *yi*), онда се *r* добија помоћу



За FEV1 и висину имамо



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. Oво може бити приказано стављањем на место *yi* 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 узимaмо 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. Oво 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 ako je регресија *X* на *Y*, онда је . Oво произилази из формулa за *r* и *b*. За FEV1 податке, *b* = 0.074389 и *b′* = 4.5424, тако да је , па је квадратни корен од тога једнак 0.58129, и то је вредност коефицијента корелације. Такође имамо



То је објаснила пропорција варијабилности, описана у делу 8.4.

### 8.6 Коришћење коефицијента корелације

Коефицијент корелације има неколико примена. Oн 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. Oво 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.

Proverite svetlosni indikator na tastaturi da biste videli da li je taster NUM LOCK uključen. Ako nije, pritisnite taster NUM LOCK.

Otkucajte prvi broj u izračunavanju pomoću numeričke tastature.

Na tastaturi otkucajte znak **+** za sabiranje, **-** za oduzimanje, **\*** za množenje ili **/** za deljenje.

Otkucajte sledeći broj u izračunavanju.

Pritisnite taster ENTER da biste završili izračunavanje.