Рационална терапија
2018, Vol. X, No. 1, стр. 31-35, **UDK:**
**DOI:**

**Стручни чланак/Professional Paper**

# АНАЛИЗА ФАКТОРА КОЈИ УТИЧУ НА ИСХОД ЛЕЧЕЊА ПОМОЋУ МУЛТИВАРИЈАНТНЕ ЛИНЕАРНЕ РЕГРЕСИЈЕ

**Слободан М. Јанковић**

Факултет медицинских наука, Универзитет у Крагујевцу

# ANALYSING FACTORS WHICH INFLUENCE TREATMENT OUTCOMES BY MULTIVARIABLE LINEAR REGRESSION

**Slobodan M. Janković**

Faculty of Medical Sciences, University of Kragujevac

Примљен/Received: 6.4.2018. Прихваћен/Accepted: 10.4.2018.

## САЖЕТАК

Мултиваријантна линеарна регресија је поступак изградње математичког модела праве са више независних варијабли које се сабирају међусобно и са једном константом да би дали вредност зависне варијабле, тј. исхода. Зависна варијабла је континуалног нумеричког карактера, док независне варијабле (предиктори) могу бит и нумеричког и категоријског типа. Мултиваријантна линеарна регресија је изузетно користан метод за испитивање утицаја већег броја варијабли на исход лечења (зависну варијаблу). Не само да је помоћу тог метода могуће проценити да ли нека варијабла утиче значајно на исход или не, већ се добија и квантитативна мера тог утицаја, као и релативни значај сваке од варијабли у односу на друге. Своју највећу примену мултиваријантна линеарна регресија има у опсервационим клиничким студијама, посебно онима типа "студије пресека".

**Кључне речи:** мултиваријантна линеарна регресија, колинеарност, коефицијент детерминације.

## ABSTRACT

Multivariable linear regression is a procedure of building mathematical model of straight line with several independent variables which are added one to another and to a constant to give value of dependent variable, i.e. of outcome. Dependent variable has to be continuous and numeric, while independent variables may be of both numeric and categorical type. Multivariable linear regression is very useful method for testing influence of multiple independent variables on a treatment outcome (dependent variable). With this method it is possible not only to estimate whether certain independent variable may influence the outcome significantly, but also to quantify this influence and compare strength of influence between the variables. Multivariable linear regression is especially useful method for analyzing data of observational clinical studies, particularly "cross-sectional" ones.

**Keywords:** multivariable linear regression, collinearity, coefficient of determination.

***Контакт:*** Слободан М. Јанковић,
Факултет медицинских наука, Универзитет у Крагујевцу, Светозара Марковића 69
e-mail: slobnera@gmail.com

## УВОД

Ако се исход лечења у некој клиничкој или опсервационој студији мери континуалном нумеричком варијаблом (тј. варијаблом чије вредности могу бити било где на скали којом се она мери), факторе који утичу на исход називамо предикторским или независним варијаблама. Предикторских варијабли обично има више, и да би објаснили како оне утичу на исход лечења, један од начина је да формирамо једначину (тј. математички модел) праве линије у којој ће учествовати све предикторске варијабле:

$$y = k_0 + a * k_1 + b * k_2 + \cdots + aa * k_n$$

У овој једначини "y" је ознака за исход, тј. зависну варијаблу, "a", "b" и "aa" су ознаке за вредности предикторских варијабли, а "$k_0$", "$k_1$", "$k_2$" и "$k_n$" су ознаке за коефицијенте испред предикторских варијабли који одређују колики утицај свака од њих има на исход. Предикторске варијабле могу бити и нумеричког, и категоријског карактера, и формирање оваквог модела омогућава да се квантификује утицај сваке од њих на исход, и то прилагођен за истовремено дејство свих осталих варијабли. Овакав тип статистичке анализе се назива и мултиваријантном, јер узима у обзир дејство свих предикторских варијабли заједно.

## УСЛОВИ ЗА СПРОВОЂЕЊЕ МУЛТИВАРИЈАНТНЕ ЛИНЕАРНЕ РЕГРЕСИЈЕ

Да би могао да се направи модел мултиваријантне линеарне регресије неопходно је да буду испуњени одређени услови, којих има укупно четири. Први услов је да постоји линеарна веза између предикторских варијабли и исхода, тј. да повећање или смањење предикторске варијабле доводи до пропорционалног повећања или смањења исхода. Други услов је да варијабилност података око регресионе линије треба да буде једнакомерна у свим деловима те линије (тј. да постоји хомосцедастичност). Трећи услов се односи на варијабилност података око сваке тачке на регресионој линији, која треба да следи нормалну дистрибуцију. Најзад, четврти услов је испуњен ако је одступање сваког појединачног податка од регресионе линије независно од одступања других података[1]. Ако су ови услови испуњени, модел мултиваријантне линеарне регресије се може успоставити, а колико он заиста објашњава утицај предикторских варијабли на исход говори нам коефицијент детерминације ($R^2$), тј. фракција укупне варијабилности која се може објаснити моделом регресије. Што је коефицијент детерминације ближи јединици, то модел мултиваријантне линеарне регресије више објашњава варијације исхода помоћу варијација предиктора.

## КОЛИНЕАРНОСТ

Предикторске варијабле могу бити у међусобној вези, тј. вредности две или више њих могу међусобно корелирати. Такву појаву називамо колинеарношћу, и она значајно утиче на модел мултиваријантне линеарне регресије. Када се у модел надодаје нова предиктоска варијабла која корелира са другим предиктором што је већ у моделу, онда ће њено додавање утицати на коефицијент испред тог другог предиктора, мењајући му вредност и повећавајући његову стандардну грешку. С друге стране, уколико нови предиктор не корелира ни са једним који се већ налази у моделу, тј. ако нема колинеарности, његовоувођење неће имати никакав утицај ни на вредност коефицијената осталих предиктора, ни на стандардне грешке тих вредности. Зато се приликом избора предикторских варијабли које уносимо у модел увек трудимо да изаберемо оне које међусобно нису колинеарне (тј. које су међусобно независне), јер ћемо тиме повећати статистичку снагу наше студије, чак иако број испитаника остаје исти. Добар избор предиктора које ћемо укључити у модел подразумева проналажење свих релевантних варијабли које уопште имају утицај на исход, али укључивање само оних који уносе независну информацију у модел. У том смислу је најбоље избећи аутоматске системе за селекцију предикторских варијабли (као што су скрининг ефеката појединачних варијабли и додавање или одузимање варијабли корак по корак), већ варијабле изабрати на основу наших концептуалних знања о проблему који истражујемо.

Да ли постоји колинеарност између предиктора може се испитати на два начина: (1) формирањем матрикса коефицијената корелације између свих предиктора понаособ (биваријантна корелација) – колинеарност не постоји ако ниједан од коефицијената не превазилази вредност 0.8; (2) израчунавањем фактора инфлације варијансе појединачног предиктора који указује у којој мери колинеарност повећава варијансу вредности тог пре-

диктора; фактор инфлације варијансе (engl. variance inflation factor – VIF) се израчунава по формули $VIF = \frac{1}{1-R_k^2}$, где је $R_k^2$ коефицијент детерминације регресије између предиктора за кога се рачуна VIF и свих осталих предиктора са друге стране[2]. Уколико је фактор инфлације варијансе већи од 10, може се са сигурношћу рећи да колинеарност постоји, што значи да неке предикторе морамо елиминисати. Понекада се уместо фактора инфлације варијансе користи његова реципрочна вредност ($\frac{1}{VIF}$), што се назива толеранција; колинеарност постоји ако је толеранција мања од 0,1.

## ЕЛИМИНАЦИЈА ЕКСТРЕМНИХ ПОДАТАКА

Када тражимо модел мултиваријантне линеарне регресије за одређену групу података обично приметимо да подаци од неколико јединица посматрања узимају екстремне вредности. Такве, екстремне, податке је најбоље елиминисати из анализе, јер они имају непропорционално велики утицај на коефицијенте регресионе формуле. Колико је велики тај утицај најбоље се може проценити тзв. Куковом дистанцом, тј. збиром промена свих коефицијената у случају елиминације једног екстремног податка (плус одсечак на Y-оси); када је Кукова дистанца велика, то указује да такав екстремни податак свакако треба елиминисати из даље анализе[3].

## ИНТЕРАКЦИЈА ИЗМЕЂУ ПРЕДИКТОРСКИХ ВАРИЈАБЛИ

Понекада два предиктора нису само у корелацији, већ могу појачавати или слабити ефекат један другом на исход. Када постоји таква појава, она се назива интеракцијом предиктора, и тада морамо унети као посебну, додатну предикторску варијаблу комбиновано дејство та два предиктора (нпр. ако су у интеракцији предиктори a и b, онда у моделу мултиваријантне регресије морају бити и a, и b, и ab као предиктори). Када унесемо и комбинацију два предиктора као посебан предиктор, наш модел мултиваријантне линеарне регресије ће бити знатно бољи, тј. имаће већи коефицијент детерминације[3].

## ПАРАМЕТРИ МУЛТИВАРИЈАНТНЕ ЛИНЕАРНЕ РЕГРЕСИЈЕ

Као што је већ поменуто, коефицијент детерминације ($R^2$) је фракција варијабилности исхода коју објашњава модел мултиваријантне лилнеарне регресије, и може се израчунати на следећи начин: $R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i-\bar{y})^2}{\sum_{i=1}^{n}(Y_i-\bar{y})^2}$, где је $\bar{y}$ средња вредност свих опсервираних вредности исхода, $\hat{Y}_i$ вредност исхода за појединачног пацијента израчуната из модела линеарне регресије, а $Y_i$ опсервирана вредност исхода за појединачног пацијента[4]. Из једначине се јасно види да је у бројиоцу укупна варијабилност коју објашњава регресиона линија, а у имениоцу укупна варијабилност исхода. Корен из коефицијента детерминације је коефицијент корелације (R) између исхода и вредности из модела $a * k_1 + b * k_2 + \cdots + aa * k_n$, тј. он нам говори у којој мери су повезани исход и збир вредности свих предикторских варијабли помножених са одговарајућим коефицијентима. Више вредности и коефицијента детерминације, и коефицијента корелације, указују на бољу предиктивну моћ модела, тј. независне варијабле заједно боље предвиђају вредности исхода.

Коефицијент детерминације израчунат на горе поменут начин може бити нереално висок, посебно у случају ако модел има много предикторских варијабли; зато се израчунава тзв. прилагођени коефицијент детерминације, који се коригује за број унетих варијабли, и тако представља бољу процену квалитета модела мултиваријантне линеарне регресије од неприлагођеног коефицијента детерминације.

Коефицијенти испред предикторских варијабли у моделу ("$k_1$", "$k_2$". … "$k_n$")указују у којој мери њихове предикторске (независне) варијабле утичу на исход (зависну варијаблу) НЕЗАВИСНО од осталих предикторских варијабли, тј. приликом њиховог израчунавања утицај осталих варијабли је искључен. Када је модел формиран, коефицијенти омогућавају да се утицај сваког од предиктора на исход тачно квантификује, нпр. ако је коефицијент испред метформина (независна варијабла) као предиктора гликемије (исход) -2.3, то значи да примена метформина снижава гликемију за 2.3 милимола по литру. У резултатима обраде мултиваријантне линеарне регресије коефицијенти из модела (које нази-

вамо "сировим") се обично означавају словом "B" са субскриптом ($B_1$, $B_2$,…$B_n$), док се њихов стандардизовани облик означава грчким словом бета ("β"). Стандардизовани облик коефицијената се добија када се они помноже са односом (разломком) варијансе исхода и варијансе тог предиктора, и омогућава да се варијабле међусобно пореде по значају на основу величине стандардизованог коефицијента. Што је варијанса предиктора већа, стандардизована вредност коефицијента ће бити мања, а тиме и значај варијабле чијој је то коефицијент. Дакле, "сирови" коефицијенти нам омогућавају да квантификујемо утицај сваког предиктора на исход, а стандардизовани коефицијенти да упоредимо међусобно значај предикторских варијабли[5].

У резултатима мултиваријантне линеарне регресије се за сваку предикторску варијаблу израчунава вредност "T" и њој одговарајућа вредност "p". "T" је број који се добија када се коефицијент испред предикторске варијабле подели са стандардном грешком коефицијента; што је "T" веће, тј. удаљеније од нуле, то је вероватноћа нулте хипотезе (да та предикторска варијабла не утиче значајно на исход), тј. "p", мања. Дакле, "T" вредност нам говори да ли предикторска варијабла значајно утиче на исход, што се може закључити уколико је p мање од 0,05[6].

За цео модел израчунава се помоћу анализе варијансе вредност "F", која представља количник средњег квадратног одступања због модела (тј. збира квадрата одступања израчунатих тачака на регресионој линији од средње вредности свих исхода, подељеног са бројем пацијената у узорку минус 1) и средњег случајног одступања (тј. збира квадрата одступања опсервираних вредности исхода од израчунатих тачака на регресионој линији, подељенод са бројем пацијената у узорку минус 1). "F" вредност нам уствари говори колико пута је варијабилност објашњена моделом регресије већа од случајне варијабилности, и постаје значајна (а тиме и наш модел) уколико превазиђе граничне вредности у "F" дистрибуцији, изнад којих је вероватноћа нулте хипотезе (тј. вероватноћа да регресиони модел није значајан) мања од 0,05[7].

## ПОСТУПАК ИЗГРАДЊЕ МОДЕЛА

Модел изграђујемо тако што укључујемо предикторске варијабле и затим пратимо параметер модела, тежећи да они добију што повољније вредности. Предикторе можемо укључивати постепено, један по један (енглески назив метода "forward selection"), све дотле док нови предиктори значајно додају способности модела да објасни исход. Могућа је употреба и обрнуте методе: све потенцијалне предикторске варијабле се унесу у модел, да би се затим једна по једна елиминисале, и пратило како то утиче на параметер модела (задржавамо предикторе чије отклањање значајно смањује предиктивну моћ модела). Овакав метод се енглески назива "backward deletion". Метод "stepwise selection" је комбинација претходна два метода, где почињемо са уносом једне по једне варијабле, а онда повремено елиминишемо варијабле уните у ранијим корацима да би проверили да ли све уните заиста треба да остану у моделу – резултат је модел са најповољнијим вредностима параметара. Најзад, када смо сигурни да неке варијабле треба да буду у моделу јер постоји концептуално оправдање за то, можемо их унети све истовремено, "у блоку", и тиме добити почетну форму модела на коју можемо даље надодавати и одузимати нове варијабле методом "stepwise selection".

## ПРОВЕРА КВАЛИТЕТА МОДЕЛА МУЛТИВАРИЈАНТНЕ ЛИНЕАРНЕ РЕГРЕСИЈЕ

На почетку је поменуто да је један од услова изградње доброг модела мултиваријантне линеарне регресије једнакомерност варијабилности података око регресионе линије, тј. да резидуалне вредности (разлике између опсервираних вредности исхода и вредности израчунатих из модела мултиваријантне регресије) буду сличне у целом опсегу предикторских варијабли. Једнакомерност се може проверити формирањем графика где ће на x-оси бити стандардизоване израчунате вредности исхода из модела (стандардизоване значи да су апсолутне вредности исхода подељене са својом стандардном девијацијом), а на y-оси стандардизоване резидуалне вредности (тј. апсолутне вредности резидуа подељене са стандардном девијацијом резидуа). Уколико се на графику види "облак" од тачака, тј. нема никаквог посебног облика или тренда, једнакомерност је постигнута[5].

Независност варијабилности појединачних података (тј. података од појединачних пацијената) једних од других се може проверити Дурбин-Вотсон тестом, чији резултати варирају од нуле до 4. Вредности око 2 значе

да су подаци независни, док вредности блиске нули указују на значајну позитивну корелацију, а вредности блиске 4 говоре о значајној негативној корелацији[5].

На крају, треба проверити и да ли су резидуалне вредности нормално дистрибуиране. Прво се апсолутне резидуалне вредности дељењем са стандардном девијацијом преведу у стандардизоване резидуалне вредности, а затим се нормалност њихове расподеле може проветити Колмогоров- Смирновљевим тестом или Шапиро-Вилковим тестом.

У закључку се може рећи да је мултиваријантна линеарна регресија изузетно користан метод за испитивање утицаја већег броја варијабли на исход (зависну варијаблу) лечења који је континуална нумеричка величина. Не само да је помоћу тог метода могуће проценити да ли нека варијабла утиче значајно на исход или не, већ се добија и квантитативна мера тог утицаја, као и релативни значај сваке од варијабли у односу на друге.

## ЗАХВАЛНИЦА

Рад је делимично финансиран средствима пројекта ОИ 175007 који је одобрило Министарство просвете, науке и технолошког развоја.

## ЛИТЕРАТУРА

1. Marill KA. Advanced statistics: linear regression, part I: simple linear regression. Academic emergency medicine. 2004;11(1):87-93.
2. 12.4 - Detecting Multicollinearity Using Variance Inflation Factors | STAT 501 [Интернет]. [цитирано 06. Април 2018.]. Available at: https://onlinecourses.science.psu.edu/stat501/node/347.
3. Marill KA. Advanced statistics: linear regression, part II: multiple linear regression. Academic emergency medicine. 2004; 11(1): 94-102.
4. Schneider A, Hommel G, Blettner M. Linear regression analysis: part 14 of a series on evaluation of scientific publications. Deutsches Ärzteblatt International. 2010; 107(44): 776-82.
5. Chan YH. Biostatistics 201: Linear Regression Analysis. Singapore Med J 2004; 45(2): 55-61.
6. Kaya Uyanik G, Guler N. A study on multiple linear regression analysis. Procedia - Social and Behavioral Sciences 2013; 106: 234 – 240.
7. Dallal GE. How to Read the Output From Multiple Linear Regression Analyses [Интернет]. [цитирано 05. Април 2018.]. Available at: http://www.jerrydallal.com/lhsp/regout.htm.