Глава 12

МЕТОДИ ЗА ИЗУЧАВАНЕ НА ПРИЧИННИ ЗАВИСИМОСТИ

1. Корелационен анализ

1.1. Понятие за функционална и корелационна зависимости

Една от главните задачи на статистическия анализ в медицината и здравеопазването е разкриване и изучаване на взаимовръзките между явленията и причинната обусловеност на едни явления и процеси от други. От тази гледна точка може да се разглеждат ***явления причини и явления следствия.***

Явленията причини представляват условие за възникването, съществуването или промените на други явления. Променливите величини, които ги характеризират се наричат ***независими променливи*** и се отбелязват с ***x***.

Явленията следствия са резултат от въздействието на явленията причини. Променливите, които ги характеризират се наричат ***зависими променливи*** и се отбелязват с ***y.***

На основата на установените причинно-следствени връзки между явленията могат да се разработват и реализират мерки за отстраняване или намаляване на отрицателно влияещите фактори, както и мерки за използване на положително влияещите върху здравето фактори. Като правило, при разглеждане на причинната обусловеност на здравето се сблъскваме не с единични, изолирани фактори, а с множество комплексно действащи причини, което значително усложнява статистическия анализ и осъществяването на ефективни оздравителни мероприятия.

Различаваме две форми на проявление на причинно-следствените връзки между явленията: ***функционална и корелационна зависимости.***

Функционална зависимост е тази, при която на всяко значение на независимата променлива съответства точно определено значение на зависимата променлива, описваща резултата. Функционалната зависимост е характерна за математиката, физико-математическите процеси и други точни науки. Тя може да се изрази с точна математическа формула. В медицината и здравеопазването функционални зависимости почти не се срещат.

Корелационна зависимост е тази, при която определено изменение в явлението причина не винаги води до точно определена промяна в резултата, т.е. на определено значение на независимата променлива могат да съответстват няколко значения на зависимата променлива. При това измененията в зависимата променлива само отчасти зависят от съответните изменения във проучваните фактори, не са предопределени от тях, както при функционалната зависимост. При дадена стойност или категория на независимата променлива са възможни и се наблюдават често различни стойности на резултата. Напр., теглото на човека зависи главно от неговия ръст. Но освен ръста, върху теглото оказват влияние много други фактори като пол, възраст, хранене, енергоразход, здравно състояние и т. н. Затова при лица с еднакъв ръст, и даже на еднаква възраст, рядко се срещат лица с напълно еднакво тегло – колебанията варират в определени граници, т.е. има съотносителност.

Тъй като практически е невъзможно да се изучи влиянието на всички фактори върху дадено явление или процес, то при измерване на причинно-следствената връзка е нужно да се отдиференцират основните от второстепенните фактори. Необходимо е да се търси логичната връзка между явленията и да се избягва смесването на реалните корелационни зависимости с т. н. ***успоредност в измененията*** между две или повече явления, която може да бъде резултат от случайно съвпадение на някои обстоятелства, не свързани едно с друго, т. е. важно е ***да се отличава имагинерната (привидната, лъжливата) от реалната корелационна връзка.***

1.2. Видове корелационни зависимости

Според формата на проявлението си корелацията бива:

- ***праволинейна***
- ***криволинейна***

Праволинейна корелация – при нея равномерните изменения на факторите се придружават с равномерни изменения на следствието (резултата), т.е. налице е някаква постоянна пропорционалност между абсолютните размери на съответстващите изменения във факторите и следствието. Математически тя се представя чрез уравненията на правата линия –

$y = a + bx$ или $y = a – bx$.

Криволинейна корелация – при нея равномерните изменения на наблюдаваните фактори се придружават от неравномерни изменения в размера на следствието. В тези случаи въздействието на наблюдаваните фактори не е еднакво при всеки техен размер и затова се нарича сложно въздействие. Математически криволинейната корелация се представя чрез уравненията на различни видове криви линии – парабола, хипербола, експоненциална крива, реципрочна и др.

Според начина и посоката на влияние:

- ***пряка (положителна, еднопосочна***)
- ***обратна (отрицателна, разнопосочна***)

Пряка корелация се наблюдава, когато с увеличаване (намаляване) на стойностите на независимата променлива нарастват (намаляват) стойностите на зависимата променлива. Напр., с увеличаване срока на бременността нарастват стойностите на антропометричните показатели при новородените.

Обратна корелация – при нея с увеличаване (намаляване) стойностите на независимата променлива намаляват (увеличават се) стойностите на зависимата променлива. Напр., с увеличаване

степента на обхвата на децата с имунизации намалява нивото на острата заразна заболяемост.

Според начина на изследване на връзките:

- ***обикновена (проста, единична***)
- ***частична***
- ***множествена***

Обикновена корелация – при нея се измерва връзката между две променливи, характеризиращи една причина и едно следствие, без да се взема пред вид влиянието на други фактори и причини. Например зависимост между характера на жилищните условия и здравето на членовете на семейството, без да се отчита влиянието на множество други фактори.

Частична корелация – при нея се изследва връзката между две променливи, характеризиращи една причина и едно следствие, при условие, че другите независими променливи запазват константно ниво.

Множествена корелация – измерва връзката между една зависима променлива (следствие) и множество фактори и причини. Тази корелация следователно измерва целокупното влияние на всички фактори, включени в дадено проучване.

Множествената и частична корелация са значително по-сложни в техническо изпълнение, но в замяна на това те дават много по-ценна информация за зависимостите между явленията причини и следствия.

1.3. Коефициент на корелация – същност и оценка

Съществено предимство на коефициента на корелация е това, че той посредством едно единствено число дава представа за направлението и силата на връзката между изучаваните явления.

Коефициентът на корелация има числена стойност и знак (+ или -) и се означава най-често с ***r.***

Числената стойност на ***r*** характеризира силата на корелационната зависимост, а ***знакът*** пред числото показва направлението на връзката – при знак (+) е налице права корелационна зависимост, а при знак (-) е обратна.

Числената стойност на коефициента r може да варира от 0 до 1. При ***r=0*** липсва корелационна зависимост. При ***r=1*** зависимостта е функционална. Междинните значения на ***r*** говорят за по-силна или по-слаба връзка. Силата на връзката се оценява по 3– или по-често по 5-степенна скала:

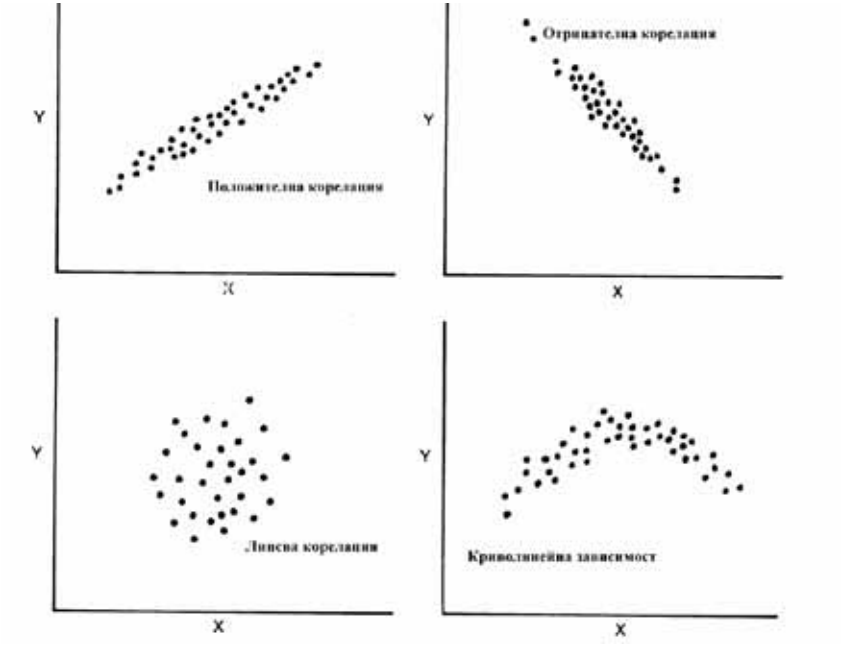
Степени на корелационна зависимост	Стойност на r при 5-степени	Стойност на r при 3-степени
слаба	до 0,3	Слаба – под 0.3
умерена	от 0,31 до 0,5	
значителна	от 0,51 до 0,7	Умерена – 0.3I до 0.7
голяма	от 0,71 до 0,9	
изключително голяма	над 0,9	Голяма – над 0.7

1.4. Диаграма на разсейване и изчисляване на *r*

Корелационният анализ е приложим както при качествени, така и при количествени променливи и независимо от характера на разпределението на променливите – алтернативно разпределение, нормално, асиметрични разпределения и т. н.

Изборът на конкретен метод за изчисляване на ***r*** се определя от целта, характера и обема на изследването, наличието или отсъствието на изчислителна техника и най-вече от начина на представяне на данните на измерителните скали.

При количествени променливи полезно средство за избор на подходящ метод е представянето на промените в независимата променлива и съответстващите им изменения в зависимата променлива на т.нар. ***диаграма на разсейване (фиг. 12.1).***



Фиг. 12.1. Диаграма на разсейването при праволинейни и криволинейни връзки

Диаграмата на разсейването позволява лесно да се определи посоката на връзката (положителна или отрицателна), вида на зависимостта (праволинейна или криволинейна) и да се придобие първоначална представа за силата на връзката или за отсъствие на зависимост между променливите. На абсцисата се поставя независимата променлива ***x***, а на ординатата – зависимата променлива ***y***. За всяка двойка значения на ***x*** и ***y*** се отбелязва пресечната точка. Групирането на точките определя вида на корелационната връзка или липсата на такава**.**

Методите за изчисляване на r могат да се групират най-общо в зависимост от характера на наблюдаваните променливи:

– изчисляване на r при качествени алтернативни променливи, т. е. при четирикратни таблици;

– ***изчисляване на r при качествени променливи с повече от две разновидности, т. е. при многократни таблици;***

– ***изчисляване на r при количествени променливи***, т. е. когато изходните данни за независимите и зависими променливи са представени чрез интервална или пропорционална скала (коефициент на корелация на Пирсън);

– ***рангова корелация*** (коефициент на Спирман);

1.5. Изчисляване на *r* при качествени алтернативни променливи

При качествени алтернативни признаци данните се представят във вид на четирикратна таблица, в която с ***x*** е означена променливата, характеризираща явлението причина, а с ***y*** – явлението следствие. Всяка една от тези променливи има само по две разновидности. Възможни са четири съчетания (комбинации) между разновидностите им, затова и таблицата се нарича четирикратна или таблица 2 x 2:

	y_1	y_2	**Общо**
x_1	a	b	$a + b$
x_2	c	d	$c + d$
Общо	$a + c$	$b + d$	$a+b+c+d$

a е съчетание на x_1y_1
b е съчетание на x_1y_2
c е съчетание на x_2y_1
d е съчетание на x_2y_2

Коефициентът на корелация при качествени алтернативни признаци (C) може да се изчисли ***чрез формулата на Чупров*** по следния начин:

$$C = \sqrt{\frac{\chi^2_{ем}}{x}}, \qquad (12.1)$$

където:

n – брой на всички случаи в таблицата

$\chi^2_{ем}$ – изчислената стойност от емпиричните данни

Когато от четирикратната таблица има изчислена стойност на χ^2, то с тази формула лесно може да се определи степента на взаимодействието. Този начин обикновено се използва, когато няма сигурни данни за формата на разпределението, т.е. при всички видове разпределения на променливите.

При същата ситуация може да се използва и модифицирана формула на ***коефициента на корелация на Пирсон (r),*** който в литературата по-често се отбелязва със символа ***φ*** и се нарича фикоефициент:

$$r(\varphi) = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \qquad (12.2.)$$

Тъй като в знаменателя има квадратен корен, то извлеченото от корена число ще има знак плюс и минус. За оформяне на крайния знак пред коефициента r знаменателят се взема с положителен знак и следователно посоката на връзката ще зависи от разликата в числителя, но често този знак се поставя чрез логиката на изследователския проблем.

След изчисляването на коефициента ***r*** се прави извод за силата на връзката (съгласно 3-степенната или 5-степенната скала) и за посоката на зависимостта (съгласно знака на r).

Освен посоченият по-горе начин, за изучаване на причинно-следствената връзка при качествени алтернативни признаци, когато данните са представени в четирикратни таблици (наричани още таблици на контингенция), може да се изчисли и друг ***специален критерий за факторно влияние, наречен odds ratio (OR***). Преводът на това понятие на български е доста труден – ***съотношение на шансовете (закономерно или случайно)***.

OR се използва много широко в съвременните епидемиологични проучвания на социалнозначимите заболявания. Предимството му е в това, че се изчислява на базата на абсолютните числа в четирикратната таблица и дава добра представа за риска за настъпване

на едно или друго неблагоприятно явление (заболяване, умиране и др.) при експонирани (изложени на рисков фактор) и при неекспонирани лица.

Използвайки посочените символи в клетките на четирикратната таблица:

$$(12.3) \qquad OR = \frac{ad}{bc}$$

Ако OR е равно на единица, то двете групи (експонирани и неекспонирани) имат еднаква вероятност за настъпване на неблагоприятното явление. ***Ако OR е по-голям от единица***, то групата на експонираните лица има по-висок риск за заболяване или друго неблагоприятно влияние. ***При OR по-малък от единица*** проучваният фактор има протективно (защитно) действие.

Пример: За да се установи има ли връзка между страничните явления и метода на лечение е проведено проучване върху появяването на странични явления (***y***) при две групи болни (***x***): лекувани само с антибиотици (x_1) и лекувани с антибиотици и витамини (x_2). Данните от проучването са представени в таблица (***табл. 12.1***):

Табл. 12.1. Резултати от лечението по два метода

Групи болни (*x*) / странични явления (*y*)	**Със странични явления (y_1)**	**Без странични явления (y_2)**	**ОБЩО**
Лекувани с антибиотици и витамини – (x_1)	**a** **9**	**b** **57**	**a + b** **66**
Лекувани само с антибиотици – (x_2)	**c** **16**	**d** **29**	**c + d** **45**
Всичко	**a + c** **25**	**b + d** **86**	**a+b+c+d** **111**

Изчислената стойност на χ^2_{em}=7,74. При К=(2-1).(2-1)=1 теоретичната стойност на $\chi^2_{0,05}$ е 3,84.

Следователно, $\chi^2_{em} > \chi^2_{0,05}$ или 7,74 > 3,84, т.е. има закономерна връзка между страничните явления и лечението.

След като с χ^2 е установена значима връзка може да се провери степента (силата) и посоката на корелацията:

Чрез формулата на Чупров:

$$(12.4) \qquad C = \sqrt{\frac{\chi^2_{em}}{n}} = \sqrt{\frac{7,74}{111}} = 0,26$$

Знакът на коефициента на корелация се определи от специалистите в проучваната област на логическа основа.

Чрез формулата на Пирсон:

$$(12.5) \quad r(\varphi) = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} = \frac{9.29 - 16.57}{\sqrt{(16+29)(9+57)(16+9)(29+57)}} = -0,26$$

***Чрез формулата за OR*:**

$$(12.6) \qquad OR = \frac{ad}{bc} = \frac{29.9}{16.57} = 0,29$$

Тъй като стойността на $OR < 1$, това показва, че лечението с антибиотици и витамини има протективен ефект.

Стойностите на ***C*** и ***r*** (***φ***) показват, че зависимостта е значима обратна и слаба – колкото повече пациенти се лекуват с антибиотици и витамини, толкова по-малко странични явления.

1.6. Изчисляване на *r* при качествени променливи с повече от две разновидности

Когато зависимата и независимата променлива са описателни и имат повече от две разновидности, корелацията се установява чрез коефициента (*C*) на Пирсон или чрез коефициента (*V*) на Крамер. За изчисляването им предварително трябва да е изчислен χ^2 и когато се установи, че връзката е значима се определя посоката и степента ѝ чрез съответните формули за ***C*** и ***V***.

(12.7)
$$C = \sqrt{\frac{\chi^2_{ем}}{\chi^2_{ем} + n}}, \qquad V = \sqrt{\frac{\chi^2_{ем}}{n(r-1)}}$$

където:

n – брой на всички случаи в таблицата

Стойността на коефициента *C* зависи от броя разновидности на променливите ***x*** и ***y*** и никога не достига 1. При *к* = 2 максималната стойност на *C* е 0,707; при к = 3 е 0,816 и т.н. (определя се при *r* = *c* по формулата: $\sqrt{(r-1)/r}$, а когато $r \neq c$ във формулата участвува символът с по-малкото число). Поради тази причина степента на връзката изчислена чрез *C* не може да се сравнява с други коефициенти и групировки на ***x*** и ***y***.

Във формулата на Крамер (V) с ***r*** отразява брой редове в таблицата, а ***c*** – брой колони. *Във формулата винаги участва този символ, който има по-малко число,* т. е. ако броят на редовете и колоните е еднакъв е без значение дали ще е *r* или *c*; ако броят на редовете е по-голям – участва ***c***, а ако е по-малък – участва ***r***.

Пример: Проведено е проучване сред 300 ученици (***табл. 12.2***) относно здравните познания (***x***) и заболяемостта (***y***). Независимата променлива ***x*** има 3 разновидности: x_1 – слаби знания; x_2 – добри знания и x_3 – отлични знания. Зависимата променлива ***y*** има също 3 разновидности: y_1 – ниска заболяемост; y_2 – средна заболяемост и y_3 – висока заболяемост.

Табл. 12.2. Зависимост между нивото на здравни знания и заболяемостта при ученици

x/y	y_1	y_2	y_3	**Общо**
x_1	10	30	60	**100**
x_2	30	40	30	**100**
x_3	60	30	10	**100**
Всичко	**100**	**100**	**100**	**300**

Изчислената стойност на χ^2 е 82,52 (за изчислителната процедура виж раздел 12.5.2). Теоретичната стойност на χ^2_t при доверителна вероятност p = 0,99 и степени на свобода K = (3 – 1) (3 – 1) = 4 е 13,3, т.е. $\chi^2_{ем} > \chi^2_t$ – 82,52 > 13,3.

От този резултат следва, че заболяемостта на учениците (***y***) зависи от здравните им знания (***x***). За да се установи силата (степента) на тази зависимост ще използваме коефициентите на Пирсон (*C*) и Крамер (*V*).

(12.8)
$$C = \sqrt{\frac{\chi^2_{ем}}{\chi^2_{ем} + n}} = \sqrt{\frac{82,52}{82,52 + 300}} = 0,47$$

$$V = \sqrt{\frac{\chi^2_{ем}}{n(r-1)}} = \sqrt{\frac{82,52}{300.(3-1)}} = 0,37$$

Тези стойности показват, че зависимостта на ***y*** от ***x*** е умерена и трябва да се тълкува с отрицателен знак – при високи (отлични) знания (x_3) има ниска заболяемост (y_1).

1.7. Коефициент на корелация на Пирсон при количествени променливи величини

Изчисляването на ***r*** при количествени явления е доста по-сложно в сравнение с посочените примери и обикновено се използва изчислителна техника. При неголям брой случаи обаче е възможно изчисляването да бъде извършено ръчно или с помощта на обикновен калкулатор по формулата:

(12.9)
$$r = \sqrt{\frac{a\sum y + b\sum xy - n\bar{y}^2}{\sum y^2 - n\bar{y}^2}}$$

Изходните данни се представят във вариационни редове, като на всяко значение на ***x*** (независима променлива) съответства определено значение на ***y*** (зависима променлива). След това се попълват следващите три колонки от таблицата, съдържащи квадратите

и сумата от квадратите на ***x*** и ***y***, произведението и сумата от произведенията на ***x*** и ***y***.

Изчислението преминава през следните стъпки:

1. Изчислява се средната аритметична на зависимата променлива:

$$(12.10) \qquad \bar{y} = \frac{\sum y}{n}$$

*2. Определя се коефициентът **в** по формулата:*

$$(12.11) \qquad b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \sum x \sum y}$$

*3. Определя се коефициентът **а** по формулата:*

$$(12.14) \qquad a = \frac{\sum y - b\sum x}{n}$$

4. Получените стойности се заместват във формулата за коефициента на корелация на Пирсон (в литературата съществуват и други формули).

5. Прави се извод за силата (по една от двете скали) и посоката (според знака) на корелационната зависимост.

6. Проверява се значимостта на корелационната зависимост.

Корелационният коефициент е значим и връзката обективно съществува, когато стойността му е по-голяма от табличната стойност ($r_{ем} > r_т$) при *к = n – 2* и доверителна вероятност 0,95 или 0,99.

Проверката може да стане и чрез *t* критерия на Стюдент, но процедурата при малък брой случаи е доста сложна, поради което изследователите трябва да се стремят към по-голям брой случаи.

Пример: Представени са данни за връзката между количествените променливи ръст и тегло в извадка от 13 момчета на възраст от 6 до 43 месеца. Да се определи посоката и силата на корелационния коефициент (***табл. 12.3)***.

Табл. 12.3. Зависимост между променливите ръст и тегло

Изследвани лица	Ръст в см (*x*)	Тегло в кг (*y*)	*x.y*	x^2	y^2
1 (6 мес.)	66,9	7,1	475,0	4475,6	50,4
2 (7 мес.)	68,5	7,2	493,2	4692,2	51,8
3 (12 мес.)	72,0	7,8	561,6	5184,0	60,8
4 (16мес.)	77,0	8,3	639,1	5929,0	68,9
5 (18мес.)	79,0	8,9	703,1	6241,0	79,2
6 (22мес.)	82,1	9,2	755,3	6740,4	84,6
7 (24мес.)	82,7	9,5	785,6	6839,3	90,2
8 (26мес.)	84,2	10,4	875,7	7089,6	108,2
9 (30мес.)	86,0	11,0	946,0	7396,0	121,0
10 (32мес.)	86,5	10,8	934,2	7482,2	116,6
11 (34мес.)	89,5	11,4	1020,3	8010,2	130,0
12 (35мес.)	89,7	11,8	1058,5	8046,1	139,2
13 (43мес.)	95,0	13,0	1235,0	9025,0	169,0
n=13	$\sum x = 1059,1$	$\sum y = 126,4$	$\sum x.y = 10483$	$\sum x^2 = 87150,8$	$\sum y^2 = 1270,1$

$$(12.15) \qquad \bar{y} = \frac{\sum y}{n} = \frac{126,4}{13} = 9,72$$

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \sum x \sum y} = \frac{13.10483 - 1059,1.126,4}{13.87150,8 - 1059,1} = 0,21$$

$$a = \frac{\sum y - b\sum x}{n} = \frac{126,4 - 0,21.1059,1}{13} = -7,38$$

$$r = \sqrt{\frac{a\sum y + b\sum xy - n\bar{y}^2}{\sum y^2 - n\bar{y}^2}} =$$

$$= \sqrt{\frac{-7,38.126,4 + 0,21.104826 - 13.9,72^2}{1270,1 - 13.9,72^2}} = 0,98$$

Следователно, при *r* = + 0,98 корелацията е права и изключително голяма. С увеличаване на ръста се увеличава и теглото.

1.8. Рангов коефициент на корелация на Спирман

Ранговият коефициент на корелация на Спирман ***(ρ)*** се използва при категорийни променливи, представени на ординална скала. Използва се също при количествени променливи или при една количествена и една качествена променлива. В основата на методиката лежи ранжирането, т.е. превръщането, както на количествените, така и на качествените променливи в рангове (присвояването на рангови номера).

Прилага се следната формула:

(12.17) $$\rho = 1 - \frac{6\sum d^2}{n(n^2-1)}$$, където:

n – брой изследвани случаи

d – разлика между ранговете на ***x*** и ***y***

Стойностите на ранговия коефициент за корелация (ρ) са от –1 до +1. Степента и посоката на връзката се определя по скалата в зависимост от знака пред ρ – еднопосочна при знак (+) и разнопосочна при знак (–).

За определяне значимостта на връзката се използва ***табл. 12.4*** при $\boldsymbol{к = n - 2 < 8}$ и доверителна вероятност 0,95 и 0,99. Когато $\rho_{ем} > \rho_{т}$ корелационният коефициент е значим и връзката е реална.

Табл. 12.4. Теоретични стойности на ранговия коефициент на корелация

Степен на свобода $к = n - 2$	Доверителна вероятност 0,95	Доверителна вероятност 0,99
3	1,000	-
4	0,886	1,000
5	0,750	0,893
6	0,714	0,833

Пример: Проведено е проучване за установяване на връзката между социално-икономическия статус и тежестта на дадено заболяване при 8 лица ***(табл. 12.5).***

Табл.12.5. Връзка между социално-икономическия статус и тежестта на заболяване

Изследвани лица	Ранг на *x*	Ранг на *y*	Разлика м/у *x* и *y* (*d*)	(d^2)
1	**6**	**5**	**1**	**1**
2	7	8	-1	1
3	2	4	-2	4
4	3	3	0	0
5	5	7	-2	4
6	4	1	3	9
7	1	2	-1	1
8	8	6	2	4
Общо	$\sum d^2 = 24$			

За да се определи силата и посоката на връзката между променливите ***x*** и ***y***, се преминава през следната последователност на изчисленията:

1. В колона 1 са изследваните лица (с номера от 1 до 8).
2. В колона 2 и 3 са представени ранговете на стойностите, характеризиращи социално-икономическия статус и заболяването: най-ниският социално-икономически статус е с ранг 1, най-високият – с ранг 8; по същия начин – най-ниската стойност на показателя за тежестта на заболяването е с ранг 1, а най-високата стойност – с ранг 8).
3. В колона 4 е разликата между ранговете (колона 2 – колона 3).
4. В колона 5 – квадрата на разликата и сумата е представена в последния ред.
5. Изчислява се коефициента за рангова корелация по формулата:

(12.18) $$\rho = 1 - \frac{6\sum d^2}{n\left(n^2-1\right)} = 1 - \frac{6.24}{8\left(8^2-1\right)} = 0,71$$

Надеждността на ρ се определя като изчислената стойност се сравнява с теоретичната стойност. При $\boldsymbol{k} = 8 - 2 = 6$ теоретич-

ната стойност на ρ е 0,71 при P = 95% и 0,83 при P = 99%. Изчислената стойност 0.72 е по-голяма от 0,71. Следователно, между социално-икономическия статус и тежестта на заболяването има значима голяма права връзка, подкрепена с вероятност P = 95%),.

1.9. Коефициент на детерминация

Коефициентът на детерминация представлява *мярка за степента, в която варирането (промените) на променливата x се свързва с варирането на втората променлива y.* Означава се с $\boldsymbol{K_{DET}}$ и се изчислява по формулата:

$$(\boldsymbol{K_{DET}}) = 100\ \boldsymbol{r^2}$$

Например, ако $\boldsymbol{r}$ = 0,9, то $\boldsymbol{K_{DET}} = 100\ .\ 0{,}9^2 = 100\ ,\ 0{,}81 = 81\%$, т. е. 81% от промените в *y* се дължат на промени в *x*.

2. РЕГРЕСИОНЕН АНАЛИЗ

2.1. Същност на регресионния анализ

В предходния раздел бе обсъждан въпросът за измерване на връзката между две количествени променливи, което е най-често срещаната ситуация при анализа на причинната обусловеност на явленията.

Променливата, която е резултат от въздействието е ***зависима променлива****, отбелязва се с* ***y*** *и в диаграмата на разсейването стойностите ѝ се нанасят върху вертикалната ос на координатната система. Променливата, която оказва влияние върху зависимата, се нарича* ***независима променлива****, отбелязва се с* ***x*** *и се изобразява върху хоризонталната ос.*

Примерите за такива зависимости са безкрайни: нерационално хранене и нивото на серумния холестерол; затлъстяване и артериална хипертония; възраст и нивото на артериално налягане; степен на физическо натоварване честота на пулса; смъртност от рак на белия дроб и тютюнопушене (разглеждано на ниво на човешките общности) и т. н.

Графичното представяне на двете променливи и изчисляването на коефициента на корелация дават представа за наличието или отсъствието на взаимовръзка между количествените променливи, но твърде често изследователите се нуждаят от по-точно описание на зависимостите. Например, може да се зададат въпроси: „По какъв начин възрастта влияе върху нивото на кръвното налягане? Каква е точно връзката между теглото и ръста и можем ли да я използваме за оценка на теглото на даден индивид с определен ръст? Каква е точно връзката между тютюнопушенето и смъртността от рак на белия дроб и каква промяна в смъртността можем да очакваме при определено снижение на честотата на тютюнопушенето в популацията?“

Отговорите на тези и други подобни въпроси изискват задълбочено изследване на връзките между зависимите и независими променливи чрез съответно моделиране. Методите, използвани за тази цел, се наричат ***регресионни анализи*** и най-елементарният от тях е ***обикновената линейна регресия.***

Същността на моделирането чрез обикновена линейна регресия е да се възпроизведе права линия, която най-добре съответствува на диаграмата на разсейване, т.е. такава права линия, при която сумата от квадратите на вертикалните разстояния от всяка точка до линията е най-малка. Много важно е правилното определяне на независимата и зависима променлива, защото регресията на ***y*** по отношение на ***x*** е коренно различна от тази на ***x*** по отношение на ***y.***

За да опишем която и да е права линия, трябва да знаем две стойности:

– ***b, наклонът на линията,*** т. е. колко стръмно нараства линията (положителен наклон) или намалява (отрицателен наклон). Той измерва с колко нараства (или намалява) ***y*** за всяка единица промяна в ***x***. Тази стойност винаги има ясна практическа интерпретация – напр., средно нарастване на кръвното налягане при нарастване на възрастта с една година.

– ***a, пресечната точка на линията,*** т. е. откъде започва линията. Това е стойността на ***y*** при ***x*** равно на нула. Понякога това има практически смисъл – например честотата на пулса при физическо натоварване равно на нула, т.е. в покой. В други случаи ***a*** няма реален смисъл – напр., теглото на индивида при нулев ръст. И в двата случая обаче функцията на ***a*** е да покаже началната точка на линията на връзката.

Всичко това се обединява в уравнението на правата линия: $y = a + b\,x$ при положителна или $y = a - b\,x$ при отрицателна корелация.

С други думи, височината на линията (***y***) във всяка точка на ***x***, се определя от началната пресечна стойност (***a***) плюс количеството (***b***), чрез което линията нараства или намалява при всяка промяна на ***x.***

Следователно, описанието на връзката се свежда до намиране на стойностите на пресичането (***a***) и наклона (***b***), които определят линията на регресия.

При определяне на линията, която най-добре описва дадена диаграма на разсейване, интуитивно прекарваме линия през средата на точките на разсейване, която минава най-близко по отношение на всички точки. Този подход е известен като намиране на линията, която дава най-доброто прилягане към точките или ***метод на най-малките квадрати.***

Какво означава точно понятието ***„най-малките квадрати“*** и защо то е най-добрият начин за обобщаване на мрежа от точки чрез линия?

Отклонението на всяка индивидуална точка от линията се измерва чрез разстоянието от тази точка по вертикал до пресичането с линията. Колкото по-голямо е това разстояние, толкова по-неточно е прилягането на линията към съответната точка. Цялостната мярка за близостта на линията се получава чрез сумиране на тези разстояния за всички точки. Обаче, ако линията минава през средата на точките на разсейване, то сумата от получените положителни и отрицателни разстояния ще бъде равна на нула. За да се избегне този проблем, подобно на подхода при определяне на стандартното

отклонение, прибягваме до повдигане на тези разлики на квадрат и след това ги сумираме. Когато тази сума от квадратите е възможно най-малка, тогава линията описва връзката най-добре. Оттук идва и названието ***„Метод на най-малките квадрати“ (Least Square Differences – LSD).***

Стойностите на ***a*** и ***b***, които определят линията на най-малките квадрати, могат да бъдат изчислени доста лесно ръчно, с калкулатор или с помощта на съответен програмен продукт.

Коефициентът b е ***коефициент на регресия*** и неговото смислово значение е да ***измерва количествено с колко се променя y (зависимата променлива) при промяна на x с единица***. Положителният знак пред коефициента ***b*** показва, че с нарастване (намаляване) на ***x*** с единица ***y*** нараства (намалява) със стойността на ***b***. Обратно – отрицателният знак показва, че с нарастване (намаляване) на ***x*** с единица ***y*** намалява (нараства) точно със стойността на ***b***.

Обикновената линейна регресия има важно практическо значение. Моделирането на зависимостта между ***x*** и ***y*** позволява да бъдат прогнозирани стойностите на ***y*** при съответни нива на ***x*** чрез последователно интерполиране или екстраполиране със стойността на коефициента ***b***. Тези подходи се прилагат широко при проучване тенденциите на здравните явления.

2.2. Предназначение на регресионния анализ

Познавателните функции на регресионния анализ се определят от три основни задачи, които се решават чрез него:

1. Да се установи съществува ли връзка между изучаваните явления.

2. Да се конструира моделът на връзката или взаимодействието между явленията чрез подходящ математически израз. За тази цел са необходими измерители на явленията, които задължително трябва да бъдат представени на силните скали на измерване, т. е. на интервалната или пропорционалната скала.

Най-общо регресионният модел има вида $y_i = f(x_1, x_2 ..... x_k, e_i)$, където y_i – променливата, която представлява следствието (зависимата променлива); $x_1, x_2, ... x_k$ – независимите променливи; e_i – случайният компонент в модела, който отразява влиянието на случайните фактори върху връзката.

3. Да се определят количествените съотношения между явленията в модела, т. е. да се измерят факторните влияния и конкретните ефекти от влиянието на един или няколко фактора върху дадено следствие (резултат). Това става чрез ***регресионните коефициенти (b_i).***

Решаването на тези три задачи има значение не само при разкриване на връзките и техните закономерности, но и за управление на явленията от действителността, тъй като може да се прогнозират стойностите на зависимите променливи величини, при различни значения на факторите.

2.3. Видове регресионни модели

Видовете регресионни модели се определят от няколко критерия.

1. Според броя на включените в анализа фактори:

- ***единични или еднофакторни*** – изследва се връзката между едно следствие (y) и една причина (x) и се представят: $y = f(x_1, e)$.
- ***множествени или многофакторни*** – изследват връзките между едно следствие (зависима променлива) и две или повече причини (фактори) – $x_1, x_2 ..... x_n$. Представят се с израза: $y = f(x_1, x_2 ..... x_i, e)$.

2. *Според формата на връзката (типа на модела):*

- ***линейни модели*** – представят се с уравнението на права линия в равнината или линеен полином в пространството. Еднофакторният линеен модел се представя с израза

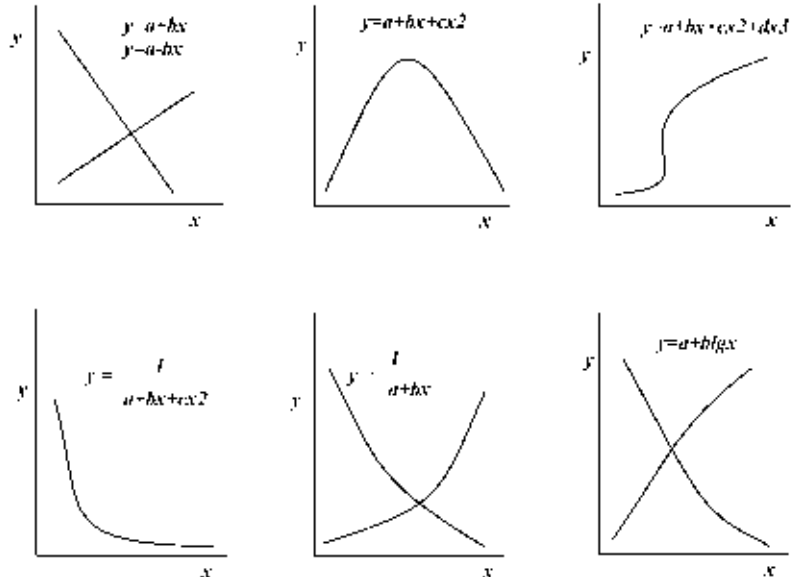
$y = a + в_i x + e$, където:

- y и x – изследваните явления;
- $в_1 ... в_i$ – параметри на модела (регресионни коефициенти) и се представят в същите мерни единици както резултатът ***y;***
- a – свободен член в уравнението, несвързан пряко с някоя величина, няма определен съдържателен смисъл, защото се формира от влиянието на много и разнопосочно действуващи причини (в стойността му е акумулирано влиянието на фактори, невключени в модела);
- e – случаен компонент (отчита влиянието на случайни причини).

- ***нелинейни модели*** – представят се с някакво уравнение на крива в равнината или с нелинеен полином в пространството.

Връзката между ***y*** и ***x*** напр. може да се представи с парабола от втора степен $y = a + в_1 x + в_2 x^2 + e$

На ***фиг. 12.2*** са представени някои видове регресионни линии и съответните им уравнения.



Фиг. 12.2. Избрани регресионни линии и съответните им уравнения

2.4. Проверка на адекватността и линейността на регресионния модел

Под адекватен модел се разбира ***приближението на модела до действителността.*** Съществуват различни методи за изследване и проверка на адекватността на модела. Както при всяка проверка, така и тук най-сигурния критерий за адекватността на модела е практиката.

Проверката на адекватността на модела се опира на ***проверката на хипотези.*** H_0 ще гласи, че между изучаваните явления обективна връзка не съществува или че компонентите на проверявания модел не могат да я представят, макар че тя обективно съществува. H_1 гласи, че връзката обективно съществува и моделът съдържа нейните основни компоненти**.** В този случай се използва ***F***-тестът – отношение между две независими и самостоятелни оценки (стойности) на общата дисперсия на зависимата променлива ***y*** при съответни степени на свобода. ***При $F_{ем}>F_m$ се отхвърля H_0 в полза на H_1, $p < 0.05$.*** Процедурата на извършване на такава проверка е известна.

Друг подход е ***техниката на конкуриращите се модели*.** По принцип тук се изследват два или повече конкуриращи се модела на дадената връзка. Например проверката на линеен с параболичен модел – $y = a + в_1 x + в_2 x^2$ и $y = a + в_1 x$.

Основно изискване при тази методика е броят на параметрите в двата модела да бъде различен. Ако проверката покаже поне два модела, за които H_0 се отхвърля, това означава, че тези модели се конкурират и трябва да се избере един от тях. Тук отново се използва ***F-тестът*** (дисперсионен анализ) чрез две независими и самостоятелни оценки (стойности) на общата дисперсия на моделираното явление ***y***.

Когато $F_{ем} > F_m$ – H_0 трябва да бъде отхвърлена, което означава, че двата модела не са равностойни. По-добър е моделът с по-малка случайна девиация, защото осигурява по-малки стохастични грешки.

Когато H_0 не може да бъде отхвърлена, двата модела се третират като еднакво подходящи и ще се използва онзи от тях, който е по-лесно приложим и изисква по-малко разходи**.**

Линейността на модела се определя също с принципите на ***дисперсионния анализ***. Сравненията тук могат да се извършат ***между линейния модел и кривата, минаваща през средните стойности на y за всяка стойност на x*** и когато се сравнява линейния модел с някакъв известен нелинеен модел на връзката. Последното може да се провери и с методиката на конкуриращите се модели.

При търсене на подходящи модели на връзките между явленията, добре е да се започне с изследване на най-простите графични образи на връзките между ***y*** и ***x***. Това става като с данните за ***y*** и ***x*** се построява графика, за да се получи емпиричната крива на връзката. С помощта на каталог, който съдържа типовете линии на регресията, част от които са показани на ***фиг. 12.2,*** се определя с кой (кои) математически израз на връзката да се работи и да се проверява адекватността на модела.

Винаги съществува риск избраният модел да не е достатъчно адекватен. Ето защо е крайно необходимо да се проверява адекватността на модела. Така изследователите ще бъдат гарантирани в по-голяма степен от неприятни изненади и рискове за неверни изводи и заключения.

Необходимо е резултатите и изводите да се тълкуват внимателно, защото при всяка проверка на хипотези сме ограничени от наличната информация, нерядко съдържаща грешки и отклонения. Ако при дадени условия проверяваният модел е адекватен, това не значи, че той е общовалиден при всички условия на място и време, при които се проявява дадената връзка. Необходимо е да се правят многократни изследвания и проверки.

2.5. Проверка на значимостта на параметрите на регрегресионния модел

Друг въпрос, който трябва да се реши когато се правят изводи от резултатите на регресионния анализ е въпросът за значимостта на оценките (стойностите) на регресионните коефициенти, тъй като съществува риск да се получат резултати, които не отразяват реалните съотношения и стойностите на параметрите могат да бъдат плод на чисто случайни фактори.

Значимостта на параметрите на регресионния модел се определя чрез теорията за проверка на хипотези, а методиката е чрез t – теста-отношение на стойността на параметъра (в) и репрезентативната му грешка.

H_0 ще гласи, че регресионните коефициенти са статистически незначими и не отразяват действителните съотношения на връзката.

Формулирането на H_1 може да стане по 3 начина в зависимост от знака на *в*:

– когато *в* е с отрицателен знак – връзката между ***y*** и ***x*** е разнопосочна, H_1 е лявостранна – напр., регресионният коефициент ***в*** значимо намалява стойността на зависимата променлива;

– когато *в* е с положителен знак – връзката между ***y*** и ***x*** е еднопосочна – H_1 е дясностранна – регресионният коефициент *в* значимо увеличава стойността на ***y***;

– когато няма сигурност за посоката на взаимодействието независимо от знака на *в* – H_1 е двустранна – регресионният коефициент *в* е значим и различен от нула.

Всичко това е валидно, когато разпределението е нормално или близко до нормалното или когато броят на изследваните единици е достатъчно голям.

Теоретичната стойност на ***t*** се определя от таблицата на Стюдент (***Приложение 1***). Решението се взема като се сравняват изчислената (t_{em}) и табличната стойност.

Когато $t_{em} > t$ ***H_0 се отхвърля*** и това означава, че регресионният коефициент ***b*** е статистически значим, проявил се е от действието на закономерни фактори, или връзката е действителна, обективна.

Обратното, когато $t_{em} < t$ – получените резултати са статистически незначими, т. е. те се проявяват под действието само на случайни фактори.

2.6. Множествена линейна регресия

Множествената линейна регресия се представя с някакво уравнение или линеен полином в пространството. Аналитичният вид на този модел се представя по следния начин:

$y = a + в_1x_1 + в_2x_2 + \ldots\ldots\ldots + в_ix_i$, където:

y – е следствието или зависимата променлива

x_i – факторите (причините) или независимите променливи

a и $в_i$ – параметрите на модела.

Същността на множествената регресия е изследване на корелационната връзка между повече от две явления. За тази цел е необходимо да се провери дали връзката реално съществува, да се избере адекватен модел на връзката, да се установи общият ефект от едновременното влияние на изследваните фактори и да се измери влиянието на всеки от факторите върху следствието (зависимата променлива).

Горният модел е приложим само тогава, когато между независимите променливи не съществува взаимодействие.

Както при единичната линейна регресия, така и при множествената линейна регресия, изчисляването на параметрите се осъществява по ***метода на най-малките квадрати.*** Разликата е в това, че при многомерната линейна регресия оценките се извършват в многомерното пространство, а всеки регресионен коефициент измерва влиянието на даден фактор върху следствието при контролирано (задържано на постоянно ниво) влияние на факторите, включени в модела на изследваната връзка. Така параметърът ***(в)***,

който стои пред факторната променлива, измерва влиянието на тази променлива ***(x)*** върху следствието ***(y)*** при постоянно равнище, т. е. при контролирано влияние на останалите фактори, участващи в модела на връзката.

Когато информацията за анализа е осигурена от представителни извадки, необходимо е да се определят стохастичните (репрезентативните) грешки и интервали на доверителност на регресионните коефициенти и на очакваните (теоретичните) стойности на ***y***. Когато границите на доверителните интервали са значително широки, резултатите от анализа не са достатъчно информативни, а причините за това могат да се търсят в недостатъчен обем на извадката, неадекватен модел, слаба зависимост на *y* от факторите x_1, x_2 и т. н.

При множествената линейна регресия могат да се проверяват няколко хипотези:

1. Хипотези за значимостта на параметрите на модела (по отделно за всеки параметър или общо) – решението се взема като се сравнят $t_{ем}$ ***и*** t_m.

2. Хипотези за наличие на регресионна връзка, адекватност на модела и относно взаимодействието между факторните променливи – решението се взема като се сравнява $F_{ем}$ ***и*** F_m ***(дисперсионен анализ).***

Процедурата е еднаква с тази при единичната линейна регресия.

2.7. Нелинейна регресия – единична и множествена

Нелинеен регресионен модел е този, при който регресионната връзка се представя с уравнение на някакъв тип крива линия в равнината или нелинеен полином в пространството. Съществуват различни типове и форми на нелинейни регресионни модели ***(фиг. 12.2)***, но условието за тяхното приложение е параметрите да бъдат свързани линейно със зависимата променлива ***y.***

А. Видове нелинейни единични и множествени модели

1. Според начина, по който са свързани параметрите със зависимата променлива y:

- ***нелинейни само по отношение на формата*** – параметрите на модела са свързани линейно (права връзка) с ***y***, но формата на модела не е линейна – този модел може да се представи като парабола от втора, трета и т. н. степен или с хипербола.

$y = a + в_1 x + в_2 x^2$ ***(параболичен модел)***

$y = 1 / a + в_1 x + в_2 x^2$. ***(хиперболичен модел)***

За определяне на параметрите на тези модели се прилага методът на най-малките квадрати, както при линейните модели.

- ***нелинейни по отношение на параметрите (обикновено са нелинейни и по формата на връзката)***. При тях променливата ***x*** може да се намира в степенния показател на параметрите или в някаква комбинация на взаимодействие с ***y.***

$y = a . в_1^x$ ***(експоненциален модел)***

$y = a /1 + в_1 . в_2^x$ ***(логистичен модел)***

2. Според възможността за трансформиране в линейни модели:

- ***вътрешно линейни модели*** – тези нелинейни модели, които могат да се трансформират в линейни чрез прости математически преобразувания.

  Напр., $y = exp(a + в_1 x)$. Този модел е нелинеен по отношение на параметрите, но е вътрешнолинеен, защото може да се трансформира в линеен, като се използва логаритмична трансформация $lg\ y = a + в_1 x$ и може да се приложи метода на най-малките квадрати. Тук особеното е, че изчислените параметри се отнасят не за оригиналния модел, а за трансформирания и затова горният израз трябва да се антилогаритмува, за да се получат оригиналните параметри ***a*** и $в_1$ и оригиналните променливи ***y*** и ***x***.

- ***вътрешно нелинейни модели*** – не могат да се превръщат в линейни чрез прости математически трансформации.

В статистическата литература съществуват различни методи за изчисляване на параметрите на вътрешно нелинейните модели – метод на крайните разлики, метод на многостепенно оценяване, метод на пряко търсене, метод на разложението по реда на Тейлър и др. Тези методи са трудоемки и не винаги водят до желаните резултати.

Б. Изисквания към множествената регресия

Основните изисквания при множествената регресия са изискванията, които налага методът на най-малките квадрати. За да се приложи множествена регресия е необходимо ***многомерно разпределение на y и x_i, равенство на дисперсиите и променливите да са случайни величини.*** Всичко това е разгледано подробно при единичната линейна регресия.

Многофакторните регресионни модели изискват голяма по обем извадка (според някои автори 7-8 пъти по-голяма от броя на променливите величини, включени в анализа). При малка извадка може да се създаде лъжлива представа за степента на връзка между променливите, тя може да се покаже дори като пълна функционална, а в действителност това да не е така.

Едно от най-важните изисквания при множествената регресия е независимост на променливите величини x_i. Когато между тях съществува връзка се говори за мултиколинеарност (ако връзката е само между две променливи – колинеарност). При наличие на мултиколинеарност (колинеарност) регресионният анализ е неприложим. Когато се установи ***r*** между независимите променливи величини по-голяма от 0.8, това е признак за колинеарност и мултиколинеарност. Ето защо, от регресионния модел трябва да се изключат корелиращите независими променливи величини.

За успешното използване на многофакторния регресионен анализ е необходимо променливите да бъдат представени на ***силните скали на измерване.*** В някои случаи, когато някоя от факторните променливи величини е представена на слаба скала на измерване и

се смята, че тя оказва силно влияние върху зависимата променлива ***y*** се препоръчва включването ѝ в анализа чрез въвеждане на фиктивна променлива в модела. Разновидностите на фиктивната променлива ***x*** се представят вместо с наименованията си с числа. Ако разновидностите са две – 0 и 1 и т. н.

В. Методи за включване на независимите променливи (факторите) в множествения регресионен модел

Друг важен въпрос при множествената регресия е избирането на независимите променливи x_i (факторите, причините). Съществува грешна представа, че включването на голям брой променливи осигурява по-добро опознаване и измерване на взаимодействията между явленията. Това в много случаи се оказва измамливо, защото много често не се отчита взаимната връзка или зависимостта между факторите. Когато между два фактора съществува линейна зависимост, в модела е достатъчно да се включи само единият от тях, който ще осигури цялата възможна информация за влиянието на двата фактора върху ***y***. Това налага от множеството възможни фактори да се включат в модела малка част без да се губи информация.

Съществуват различни методи за избор и включване на факторите в модела на многофакторната регресионна връзка:

1. Метод на последователното изключване. При него се конструира общото регресионно уравнение, в което се включват всички възможни фактори. Например:

$$y = a + в_1x_1 + в_2x_2 + в_3x_3 + в_4x_4$$

Изчисляват се параметрите на това уравнение. Определя се частният ***F***– критерий за всеки фактор като въведен последен в модела. За горното уравнение например са получени следните резултати: за връзката между ***y*** и x_1, $F_{ем}$ = 10.57, за връзката между **y** и x_2, $F_{ем}$ = 7.44, между ***y*** и x_3, $F_{ем}$= 2.80 и между **y** и x_4, $F_{ем}$ = 4.69. Теоретичната характеристика на ***F*** = 3.46 – с нея ще се сравняват емпиричните. Когато минималните емпирични стойности са по-големи от теоретичните, в модела ще останат всички

фактори и обратно. В примера има една емпирична стойност (2.80) по-малка от теоретичната (3.46) което показва, че трябва да се започне избиране на факторите отначало. Изборът започва от фактора, който има най-голяма стойност на $F_{ем}$ – за него се проверява хипотезата за влиянието му и т. н.

2. Метод на предварителното изключване. Представлява разновидност на предходния метод и при него най-напред се определят единичните регресионни уравнения и корелационна матрица на единичните корелационни коефициенти. Избира се факторът с най-голям корелационен коефициент. Определя се $F_{ем}$ и се проверява хипотезата за влиянието на този фактор. Ако нулевата хипотеза се отхвърли, този фактор остава в модела. Определят се корелационните коефициенти на единичната частна корелация между y и всеки от останалите фактори при елиминирано влияние на вече включения в модела фактор (x_1). Отново в модела се включва факторът с най-голям корелационен коефициент, проверява се H_0 и така продължава проверката, докато се изчерпят всички фактори.

3. Стъпков регресионен анализ. Представлява усъвършенстван вид на метода на изключването. Новото е, че на всеки етап от проверката става отново проверка на вече включените в предходните етапи фактори. По същество се прилага последователният F – критерий. При този анализ е възможно да отпадне фактор, който е бил включен в предходните етапи.

Провеждането на стъпковия регресионен анализ се свежда до следното:

Построява се корелационна матрица на единичната корелация между y и x_i. В първото уравнение ще се включи факторът с най-висока стойност на корелационния коефициент. С F – теста се проверява H_0 за значимостта на фактора. Ако се отхвърли H_0 – факторът остава в модела. Определят се частните корелационни коефициенти на y с останалите фактори при елиминирано влияние на включения вече в модела фактор. Включва се в модела факторът с max корелационен коефициент. Изследва се влиянието на втория фактор, а след това се

изследва влиянието на първия фактор (включен в предходния етап), за да се види, дали влиянието му върху y остава значимо. Така продължава изборът на следващите фактори. По същество тук на всеки етап се проверява значимостта не само на последния от включените фактори, но и на всички фактори, включени по-рано и се запазват в модела само онези, които имат значимо влияние.

Например, ако моделът на връзката може да се представи с израза $y = a + в_1x_1 + в_2x_2 + в_3x_3 + в_4x_4 + в_5x_5$, оценяването на факторите при съответни корелационни коефициенти ще изглежда по следния начин:

Първа стъпка. Максимален корелационен коефициент е между y и x_3, $r = 0.743$; в модела се включва x_3; установява се $y = f(x_3)$. $F_{ем} > F_m$, H_0 се отхвърля; x_3 остава в модела.

Втора стъпка. Максимален корелационен коефициент между y и x_4. $x_3 = 0.604$ (точката между x_4 и x_3 показва, че се търси връзка между y и x_4 при задържане влиянието на x_3); включва се x_4 в модела; установява се $y = f(x_3, x_4)$;

$F_{ем} > F_m$; H_0 се отхвърля; x_4 остава в модела.

Трета стъпка. Максимален корелационен коефициент между y и x_2. $x_3x_4 = 0.34$; включва се x_2 в модела; установява се $y = f(x_3x_4, x_2)$; $F_{ем} > F_m$; H_0 се отхвърля; x_2 остава в модела.

Четвърта стъпка. Максимален коефициент на корелация между y и $x_5, x_2, x_3, x_4 = 0.20$; включва се x_5 в модела; установява се $y = f(x_3, x_4, x_2, x_5)$; $F_{ем} < F_m$; H_0 се приема; x_5 не се включва в модела.

Пета стъпка. Максимален корелационен коефициент между y и x_1. $x_2x_3x_4x_5 = 0.19$; включва се в модела x_1; установява се $y = f(x_3x_4x_2x_5x_1)$; $F_{ем} < F_m$; H_0 се приема; x_1 отпада от модела.

От този анализ се вижда, че x_1 и x_5 отпадат от модела и видът на модела ще бъде: $y = a + в_2x_2 + в_3x_3 + в_4x_4$

При стъпковия регресионен анализ винаги се използва последователният F – критерий, при който се отчита реда на включване на факторите в модела.