§9. Регресионен и дисперсионен анализ

1. Линеен регресионен анализ. Този вид статистически анализ е предназначен да даване количествен израз на ефектите на дадена група метрични величини X_1, X_2, ..., X_p, които условно се наричат ***независими*** (***independent***) върху друга величина Y, която условно се нарича ***зависима*** (***dependent***). Независимите величини се наричат понякога и ***фактори***. Предметният контекст на регресионния анализ предполага каузални връзки между факторите и зависимата величини. Основната идея се състои в следното. Въз основа на взаимното популационно разпределение на всичките величини се търси естествена функционална връзка от вида

(9.1) $\quad y = f\left(x_1, x_2, \ldots, x_p\right),$

която по статистически обоснован начин дава прост израз на ефектите на отделните независими величини върху зависимата. Уравнението (9.1) се нарича ***уравнение на регресия*** на Y върху X_1, X_2, ..., X_p и се получава на база усредняване стойностите на Y при фиксирани стойности на X_1, X_2, ..., X_p. Когато взаимното популационно разпределение на всичките величини е ***нормално***, уравнението на регресия (9.1) се оказва ***линейно***

(9.2) $\quad y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p,$

където b_0, b_1, b_2, ..., b_p са коефициентите на уравнението. Тези коефициенти лесно се интерпретират по знак. Ако например $b_1 > 0$, то нарастването на x_1 води до нарастване на y и ако $b_1 < 0$, то нарастването на x_1 води до намаляване на y. По-големите по абсолютна стойност коефициенти са свързани с по-голяма промяна при зависимата величина. Съпоставката по абсолютна стойност на тези коефициенти като критерий доколко е голям ефектът на отделните независими величини трябва да отчита и други обстоятелства, свързани с дисперсиите на величините. По тази причина е по-удобно всичките величини да бъдат приведени към z-стойности (нормално стандартно разпределение) вместо (9.2), при което се получава уравнението

(9.3) $\quad \dfrac{y - \bar{y}}{s_y} = \beta_1 \dfrac{x_1 - \bar{x}_1}{s_1} + \beta_2 \dfrac{x_2 - \bar{x}_2}{s_2} + \cdots + \beta_p \dfrac{x_p - \bar{x}_p}{s_p},$

в което свободният коефициент е равен на нула, а коефициентите β_1, β_2, ..., β_p се наричат ***стандартизирани коефициенти на регресия***. Стандартизираните коефициенти се интерпретират по знак както преди но вече са съпоставими и по абсолютна стойност.

За всеки от коефициентите на регресия се пресмята значимост, като фактически се проверява нулева хипотеза, че съответният популационен коефициент е равен на нула. При тези хипотези проверяващата статистика е има разпределение на Стюдънт $t\left(n - p - 1\right)$, където n е обемът на извадката. Ако някой от коефициентите се получи с пренебрежима значимост, то съответната независима величина може да бъде изключена от анализа без съществена загуба на информация. Освен това за целия модел се пресмята величина R^2 (квадрата на множествения коефициент на корелация), която показва каква пропорция от изменчивостта на зависимата величина се обяснява посредством изменението на независимите величини. В тази връзка се проверява и нулева хипотеза, че линейният регресионен модел не обяснява по същество изменчивостта при зависимата величина, като проверяващата статистика е разпределена $F\left(p, n - p - 1\right)$.

Да разгледаме следния пример. Изследва се група от $n = 61$ девойки от специализирано висше учебно заведение по различни психични показатели. Зависимата величина представлява нивото на адаптация към средата, формирана чрез факторен анализ въз основа на шест основни скали за адаптация, а в качеството на независими се вземат следните седем величини $(p = 7)$, всяка от които също е формирана чрез факторен анализ от няколко основни скали, които няма да описваме във всички подробности. Следващата таблица съдържа кратко описание на използваните независими величини – фактори.

Таблица 9.1.

ФАКТОРИ	**описание**
Положителна перцепция на средата	задоволеност, защитеност, висока нормативна база, строгост на реда
Отрицателни фактори на средата	изолираност, натовареност, противоречивост
Негативни емоционални състояния	безпокойство, неувереност, потиснатост, самота, стеснителност
Позитивни емоционални състояния	дружелюбност, активност, вътрешен интерес
Личностна проблематичност	невротичност, занижена самооценка, липса на общителност, стеснителност, емоционална лабилност
Импулсивност на поведението	спонтанна агресивност, раздразнителност, реактивна агресивност, откритост
Социална подкрепа	организационна, вертикална, хоризонтална

След провеждане на регресионния анализ се получават следните резултати.

Таблица 9.2.

Фактори	**R^2=0.697; F(7,53)=17.424; p=0.000**		
	Beta	**t(53)**	**p**
Положителна перцепция на средата	**0.199**	**1.737**	**0.088**
Отрицателни фактори на средата	**-0.399**	**4.340**	**0.000**
Негативни емоционални състояния	**-0.143**	**1.487**	**0.143**
Позитивни емоционални състояния	**0.385**	**4.051**	**0.000**
Личностна проблематичност	**-0.017**	**0.170**	**0.866**
Импулсивност на поведението	**0.022**	**0.247**	**0.806**
Социална подкрепа	**0.171**	**1.205**	**0.233**

Тук статистически значими в рамките на традиционния критерий $p < 0.05$ се оказаха само коефициентът на фактора "отрицателните фактори на средата" $[beta = -0.399\,;\, p = 0.000]$, който има отрицателен ефект върху адаптацията понеже знакът на коефициента е минус и коефициентът на фактора "позитивни емоционални състояния" $[beta = 0.385\,;\, p = 0.000]$, който фактор има положителен ефект понеже знакът на съответния коефициент е плюс. За останалите коефициенти не разполагаме с достатъчно основания да приемем, че техните популационни стойности фактически са

различни от нула, освен евентуално за фактора "положителна перцепция на средата", за който имаме [$beta = 0.199$; $p = 0.088$]. Същият коментар с по-ниска степен на увереност може да бъде направен за фактора "социална подкрепа" и за фактора "негативни емоционални състояния".

Моделът като цяло обяснява $R^2 100\% \approx 70\%$ от изменчивостта на зависимата величина "адаптация", което е един много добър резултат.

Постигнатият резултат може лесно да бъде коментиран и в рамките на предметната област, понеже женската природа разбира адаптацията преди всичко като контрол над средата в повече детайли и свързаната с това проява на положителни емоции като средство за компенсация и контрол. Липсата на ефект на "личностната проблематичност" не влияе върху адаптацията се обяснява от факта, че изследваната група от девойки е преодоляла много сериозна селекция на входа, която е допуснала индивиди с еднородно високи личностни характеристики.

Следващият пример е свързан с анализ на семестриалния успех на група от $n = 135$ обучаеми в специализирано висше учебно заведение. Целта на анализа е установяване степента на прогностична валидност на бал образуващите фактори върху средния годишен успех от семестриалните изпити, който представлява зависимата величина. В качеството на независими величини се разглеждат оценката по математика от дипломата, оценката по физика от дипломата, оценката по български език от дипломата, оценката от конкурсния изпит по математика и средния успех от дипломата.

Резултатите от анализа са приведени в следващата таблица.

Таблица 9.3.

Фактори	**R^2=0.372; F(5,129)=15.309; p=0.000**		
	Beta	**t(129)**	**p**
оценка по математика от дипломата	**-0.021**	**-0.252**	**0.801**
оценка по физика от дипломата	**0.055**	**0.644**	**0.521**
оценка по български език от дипломата	**0.137**	**1.109**	**0.270**
оценка от конкурсния изпит по математика	**0.429**	**5.957**	**0.000**
среден успех от дипломата	**0.294**	**2.205**	**0.029**

Анализът показва, че водещо значение има оценката по математика от конкурсния изпит [$beta = 0.429$; $p = 0.000$] и средния успех от дипломата [$beta = 0.294$; $p = 0.029$]. Другите бал образуващи фактори имат подчертано незначим ефект, при което особено впечатление прави ефектът от оценката по математика [$beta = -0.021$; $p = 0.801$] на фона на водещото значение на оценката от конкурсния

изпит. От всичките приведени дипломни оценки най-важна роля има оценката по български език $[beta = 0.137; p = 0.270]$, който ефект си струва да бъде отбелязан въпреки неговата формална статистическа незначимост.

Моделът като цяло обяснява $R^2 100\% \approx 37\%$ от изменчивостта на зависимата величина "среден годишен успех от семестриалните изпити", което сравнително малко и означава, че доброто обяснение на зависимата величина изисква включване на други независими величини, които в дадения модел не присъстват.

Полученият резултат позволява различни интересни интерпретации, като преди всичко доказва ефективността от проведения конкурсен изпит.

2. Дисперсионен анализ. В предишните раздели вече беше разгледан случая на еднофакторен дисперсионен анализ както и на дисперсионен анализ за повтарящи се измервания, които представляват частни случаи на дисперсионен анализ. Тук ще разгледаме примери на многофакторни дисперсионни анализи.

Основната цел на дисперсионния анализ е същата както при регресионния анализ – да се изследва значимостта на определена група фактори (независими величини) върху дадена зависима величина, само че тук факторите са номинални величини.

Да разгледаме пример на двуфакторен дисперсионен анализ, при който зависимата величина представлява началната заплата на работещи в държавни учреждения с два фактора, първият от които е полът $(F1)$, а вторият $(F2)$ принадлежността на лицето към основната група население или към малцинствена група. Данните са взети от съпровождащите файлове на **SPSS**. Извадката съдържа $n = 216$ наблюдения на лица със средно образование, подредени по различните категории както следва.

Таблица 9.4.

	основно население	**малцинство**	общо по редове
жени	**128**	**30**	158
мъже	**36**	**22**	58
общо по стълбове	164	52	216

В този случай се говори за небалансиран дизайн на експеримента, понеже броят на наблюденията в основните клетки е различен.

Тук се проверяват три нулеви хипотези.

H_0^{F1} : *първият фактор има незначим ефект*

H_0^{F2} : *вторият фактор има незначим ефект*

$H_0^{F1\times F2}$: *двата фактора имат незначимо взаимодействие*

Всяка от тези нулеви хипотези има проверяваща статистика, пресметната въз основа на дисперсии, която следва някакво F-разпределение, откъдето идва и наименованието дисперсионен анализ. Иначе по същество дисперсионният анализ представлява специфична форма на сравнителен анализ.

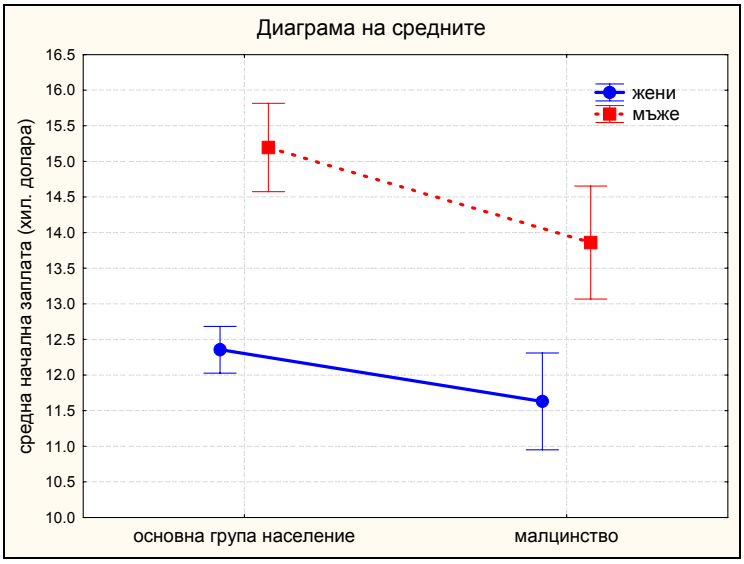
Следващата таблица съдържа резултатите от анализа.

Таблица 9.5. (лица със средно образование)

	F(1,212)	**p**
пол	63.136	0.000
малцинствена принадлежност	10.427	0.001
взаимодействие (пол* малцинствена принадлежност)	0.908	0.341

В случая на двуфакторен дисперсионен анализ, резултатите много добре се илюстрират от една диаграма, показваща разположението на средните по категориите на факторите.

Диаграма 9.1. (лица със средно образование)



Ефектът на пола [$F(1,212) = 63.136$; $p = 0.000$] се изразява в това, че мъжете и от двете групи население започват с по-висока начална заплата, което се вижда от диаграмата, понеже линията за мъжете лежи по-високо от тази на жените. Ефектът на принадлежност към малцинствена група [$F(1,212) = 10.427$; $p = 0.001$] се изразява в това, че както жените така и мъжете в групата от малцинствата започват с по-ниска начална заплата. Между двата фактора няма значимо взаимодействие

[$F(1,212) = 0.908$; $p = 0.342$], което графично се изразява в това, че линиите на двата пола вървят почти успоредно.

Настоящият пример е подбран с оглед лесно тълкуване на ефектите. В общия случай нещата могат да се получат трудни за тълкуване, особено когато е налице взаимодействие при двата фактора.

При наличие на значими ефекти следва да се проведе и Post-Hoc анализ. Следващата таблица съдържа резултатите от този анализ.

Таблица 9.6. Post-Hoc анализ за ефектите (лица със средно образование)

			{1}	{2}	{3}	{4}
{1}	жени	малцинство		0.443123	0.000008	0.040707
{2}	жени	основна група	0.443123		0.000008	0.000531
{3}	мъже	малцинство	0.000008	0.000008		0.087976
{4}	мъже	основна група	0.040707	0.000531	0.087976	

Същите ефекти се повтарят и при извадката на лица с висше образование.

Таблица 9.7. (лица с висше образование)

	F(1,143)	**p**
пол	19.791	0.000
малцинствена принадлежност	3.926	0.049
взаимодействие (пол* малцинствена принадлежност)	2.258	0.135

Диаграма 9.2. (лица с висше образование)

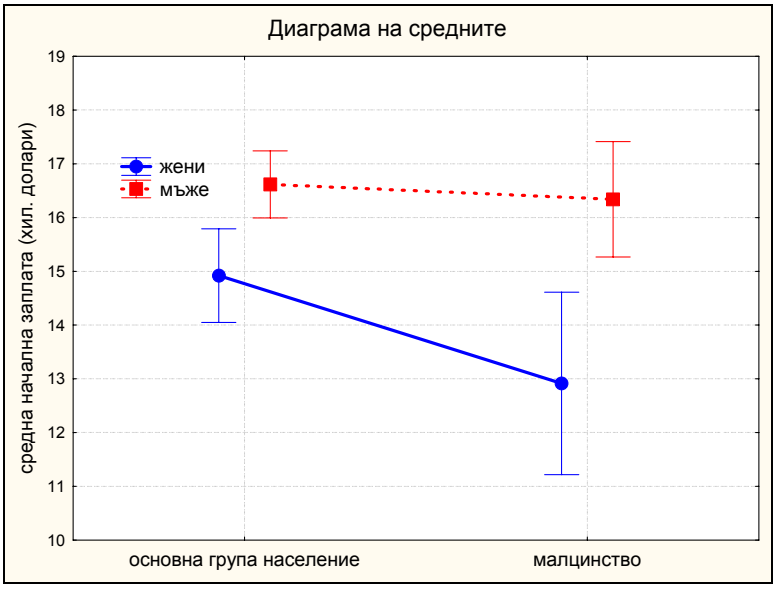


Таблица 9.8. Post-Hoc анализ за ефектите (лица с висше образование)

			{1}	{2}	{3}	{4}
{1}	жени	малцинство		0.349489	0.032822	0.250159
{2}	жени	основна група	0.349489		0.012316	0.024677
{3}	мъже	малцинство	0.032822	0.012316		0.984193
{4}	мъже	основна група	0.250159	0.024677	0.984193	

Описаните два двуфакторни анализа могат да бъдат обединени в един трифакторен анализ. Резултатите от него са приведени в следващата таблица.

Таблица 9.9. (за всички лица)

	F(1,355)	p
пол $(F1)$	69.298	0.000
образование $(F2)$	40.070	0.000
малцинствена принадлежност $(F3)$	12.575	0.000
взаимодействие $(F1*F2)$	0.002	0.966
взаимодействие $(F1*F3)$	0.840	0.360
взаимодействие $(F2*F3)$	0.033	0.857
взаимодействие $(F1*F2*F3)$	3.648	0.057

При този трифакторен анализ се проверяват общо седем нулеви хипотези, съответни на редовете в таблица 9.9. Първите три хипотези са за главните ефекти на трите фактора, следващите три хипотези са за двойните взаимодействия и една последна хипотеза за тройно взаимодействие. Анализът показва значими главни ефекти и на трите фактора.

Таблица 9.10. Post-Hoc анализ за ефектите (за всички лица)

				{1}	{2}	{3}	{4}	{5}	{6}	{7}	{8}
{1}	жени	средно	основна група		0.918	0.000	0.999	0.000	0.345	0.000	0.000
{2}	жени	средно	малцинство	0.918		0.000	0.909	0.000	0.023	0.000	0.000
{3}	жени	висше	основна група	0.000	0.000		0.491	1.000	0.776	0.024	0.337
{4}	жени	висше	малцинство	0.999	0.909	0.491		0.317	0.982	0.006	0.016
{5}	мъже	средно	основна група	0.000	0.000	1.000	0.317		0.509	0.131	0.624
{6}	мъже	средно	малцинство	0.345	0.023	0.776	0.982	0.509		0.001	0.007
{7}	мъже	висше	основна група	0.000	0.000	0.024	0.006	0.131	0.001		1.000
{8}	мъже	висше	малцинство	0.000	0.000	0.337	0.016	0.624	0.007	1.000	

Печалните изводи от направеният в таблица 9.9 анализ оставяме на читателя.