Никола Иванов Чолаков е роден през 1944 година в София. Завършва висше образование по математика в СУ "Св. Климент Охридски"). В УНСС е на работа от 1977 година, доктор по икономика и доцент в кат. "Труд и социална защита" – УНСС.

Специализирал в Московския държавен университет "Ломоносов" (1978 г.) и в Международната организация по труда – Женева (1980г.).

Консултант на ООН по демография (1984 и 1988 г.), Международен експерт на ООН (2006), пълномощен министър в постоянното представителство на Република България в ООН в Ню Йорк (1997 г.) и пълномощен министър в посолството на Република България в САЩ (1998 – 2001 година).

Основни публикации: "Демографските процеси и населението на България" (1972 г.), "Прогноза за населението на НРБ по окръзи (градове и села) през периода 1970 – 2000 година" (1975 г.), "Демографическая ситуация, принципы, проблемы и меры демографической политики НР Болгарии" (1985 г.), "Population ageing and its social and economic consequences", Economic Studies No. 3, United Nations, Geneva, 1992, "Методологически и методически проблеми при статистическото изучаване на смъртността" (1992 г.), "Математически методи и модели в животозастраховането и пенсионното осигуряване (2003 г.), "Статистическата информация за естественото движение на населението и миграцията в България и някои насоки за нейното усъвършенстване", сб. "Научни трудове 2003" № 2, Университетско издателство "Стопанство", УНСС, Migration Statistics, Population Accounts and Life Tables – the Methodology used in Bulgaria, сп. "Migracijske i etnicke teme" № 2-3, Загреб, 2003; Mortality and Life Expectancies in EU Acceding Countries – Long-term Outlook, сп. "Migracijske i etnicke teme" № 1-2, Загреб, 2005 и други.

ЛОГИКА И МЕХАНИЗЪМ НА МЕТОДА НА НАЙ-МАЛКИТЕ КВАДРАТИ

доц. д-р Никола Чолаков

Методът на най-малките квадрати е един от най-разпространените методи за анализ на статистически данни и заема особено място в развитието на науката. През 1801 г. астрономът Зах публикува данни за движението на малка планета, всъщност астероид, наблюдаван на 1 януари с. г. от Джузепе Пиаци в обсерваторията в Палермо. Данните се оказали оскъдни, тъй като Пиаци успял да засече само девет градуса от траекторията преди астероидът да се скрие зад слънцето. В своя труд Зах дава няколко варианта на прогноза за движението на това небесно тяло, вкл. един на Карл Фридрих Гаус, значително различаващ се от другите. Когато се появява отново на 7 декември с.г. астероидът се оказва много близо до мястото, предсказано от Гаус. По-късно[1] се изяснява, че грубнакът на прогнозата на движението на Гаус е всъщност методът на най-малките квадрати. Междувременно и Лежандър публикува своя труд[2] върху същия метод. Има данни, че Гаус е намерил метода на най-малките квадрати още през 1795 г. и с известно основание неговото име се свързва

[1] Gauss, K. F. (1809) Werke 4, *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientum*. Göttingen.
[2] Legendre. A.M. (1806). *Nouvelles méthodes pour la détermination des orbites des comètes*. Paris.

като изобретател на този инструментариум[1]. Особен принос в развитието има и Лаплас[2] с осмислянето на цялото построение през призмата на вероятностите и статистиката. По същото време и Гаус дава много по-пълно изложение на метода[3]. Основните положения в съвременната теория на метода на най-малките квадрати са развити от Марков (1900), Ейткин (1935), Нейман (1938), Колмогоров (1946), Рао (1945, 1968), Парзън (1961), Линник (1961) и др.

Към днешна дата методът на най-малките квадрати може да се открие във впечатляващо широк спектър научни направления и практически приложения: от астрономията, метеорологията, геодезията, оптиката до биологията, зоологията, ботаниката, фармацевтиката, медицината и клиничните изследвания, практическите икономически проучвания и иконометрията, социологията, че дори и в спорта. В това обширно поле статистиката изпълнява свързваща функция, нещо като общ знаменател, между тези иначе отдалечени предметни области и това до голяма степен се дължи и на метода на най-малките квадрати, често означаван и възприеман като принцип на най-малките квадрати. Тази поразителна жизненост на метода винаги е заслужавала специално внимание.

Методът на най-малките квадрати обикновено се свързва с линейните статистически модели, при които най-ярко личи неговата ефикасност. В следващата част се спираме на тази тема за да стане ясно, че линейността на тези модели не представлява съществено ограничение и че нелинейните зависимости по принцип може да се привеждат към линейни, което и обяснява широкото разпространение на метода на най-малките квадрати. По-нататък се разглеждат теоретичните основи на метода, за което линейните статистически модели се оказват подходящ материал, и постепенно се преминава към различни негови приложения, някои от които отиват доста далече от теоретичната статистика и представляват определен практически интерес в икономиката и социологията.

1. СТАТИСТИЧЕСКО ИЗУЧАВАНЕ НА ЗАВИСИМОСТИ

Значението на метода на най-малките квадрати за обосновка и теоретично осмисляне на регресионния анализ е добре известно[4]. В сравнително по-обща форма сценарият на този аналитичен апарат за статистическо изучаване на зависимости е следният. Основният проблем е по установяване доколко величината на един показател y се определя от няколко други $x_1, x_2, \ldots x_K$. Често показателите $x_1, x_2, \ldots x_K$ се възприемат като фактори за резултата y, а понякога дори и като причини, което е малко пресилено. Може би затова в англо-саксонската литература са възприети по-неутрални термини като регресори за

[1] Nievergelt, Y. (2001) *A tutorial history of least squares with applications to astronomy and geodesy*, в *Numerical Anslysis: Historical Develoments in the 20th Century* (ред. Claude Brezinski and Luc Wuytack), Amsterdam: Elsevier, стр. 77-112. Интересен поглед върху историята на метода се намира и у Plackett, R. (1949). A Historical Note on the Method of Least Squares, *Biometrica* 36, стр. 458-460. По-пълно изложение на развитието на този дял от статистиката с приложение в икономиката може да се намери у: Stigler, S. (1986). *History of Statistics*. Cambridge, MA: Harvard University Press; Epstein, R. J. (1986). *A History of Econometrics*. Amsterdam: Elsevier.

[2] В по-късните издания на известния труд *Théorie Analytique des Probabilités*.

[3] Gauss, K.F. (1821/1823). *Theoria combinationis observationum erroribus minimis obnoxiae*, Göttingen: Heinrich Dieterich.

[4] Съйкова, И. (1981). *Статистически анализ на връзки и зависимости*. София: "Наука и изкуство", Димитров, А. (1999). *Въведение в иконометрията*, В. Търново: "Свети Евтимий Патриарх Търновски". У нас са добре известни трудовете на Дрейпер, Смит . (1973) *Прикладной регресионный анализ*, Москва: "Статистика", Болч, Хуань. (1979). *Моногомерные статистические методы для экономики*. Москва: "Статистика". Темата е развита доста подробно у: Маленво (1975), Рао (1968), Кендалл и Стюарт (1973), Theil (1971), Себер (1976).

$x_1, x_2, \ldots x_K$ и регресанд за y. Срещат се още и термините обясняващи и обяснявани променливи или признаци. В иконометрията обикновено те се разглеждат като екзогенни и ендогенни показатели. Всъщност истинските фактори и причини обикновено са дълбоко скрити зад регресорите, като последните само отбелязват някои количествени страни на тяхното действие. Аналогично и регресандът отразява на числовата скала само отделен фрагмент от ефекта на факторите. Независимо от терминологията обаче, трудно може да се очаква, че отнапред подбраните регресори биха давали изчерпателно обяснение за величината на регресанда, предвид което търсената връзка по-скоро би трябвало да се очаква като

$$y = f(x_1, x_2, \ldots x_K) + \varepsilon, \tag{1.1}$$

където допълнителният член ε отговаря на сумарното действие на неотчетените в явната форма $f(\ldots)$ регресори.

Така например потреблението на дадена група стоки може да се опита да се свърже с дохода на домакинствата и с цените на тези стоки и да се потърси обяснение доколко тези два регресора определят размера или нивото на потреблението. При такъв модел всички други фактори като размер и структура на домакинствата, каква част от доходите са парични и каква – натурални, стабилност на източниците на доход, сезонност, потреблението на други стоки, спестявания, модни влияния и др. се полагат като несъществени поотделно, а сумарният им ефект се отнася към остатъчната компонента ε. Една от задачите при регресионния анализ като традиционен метод за статистическо изучаване на зависимости е да се определи въз основа на подходящ емпиричен материал доколко такъв модел отговаря на действителността и дали всъщност не са изтървани главните фактори за потреблението, несправедливо пропуснати в основната част на регресионния модел.

За да отговаря моделът (1.1) на действителността, би трябвало остатъчната компонента ε да е сравнително малка по големина, при това случайна величина, да варира разнопосочно, да не проявява някаква тенденция, която да затъмни главната част f, да наклони везните в една посока и да се компрометира идеята, че тъкмо регресорите обясняват величината на регресанда. В противен случай би трябвало да се дешифрира какво се крие зад случайната компонента, да се изведат на преден план възможни допълнителни регресори и да се разшири обясняващата компонента f до степен, че остатъкът ε действително да има поведение на случайна величина и в нея да не се крие някакво обяснение за регресанда. Може да се окаже и обратното, че някои от първоначално подбраните регресори не оправдават очакванията и би трябвало да се изведат от главната част на модела и ефектът им да се причисли към случайната компонента.

Обичайна практика, макар и само като първо приближение при изследването на зависимости, пропорции и съотношения, е да се предполага, че формата на зависимостта f е линейна. Логиката в това почива на факта, че в развитието на функцията f в ред на Тейлор фокусът би могъл да бъде предимно върху главната част – линейния член на подобно представяне, а всички следващи – квадратични, кубични и т.н., да се отнесат към остатъчната компонента ε. Често използваните статистически данни подкрепят подобна хипотеза.

Така моделът (1.1) би могло да се представи в следната по-конкретна форма:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_K x_K + \varepsilon. \tag{1.2}$$

Тук $\beta_0, \beta_1, ... \beta_K$ представляват регресионни параметри, чиято величина би трябвало да се оцени въз основа на статистическия материал. Параметрите $\beta_1, \beta_2, ... \beta_K$ изпълняват свързваща функция в модела (1.2), като шарнирни елементи, трансформиращи движението от отделните регресори към регресанда и съчетаващи мерните единици на съответните променливи. Например, доходите в примера с изучаването на потреблението на домакинствата се очаква да бъдат в парични единици, докато размерът на потреблението би могло да бъде в натурално изражение. Така всяка единица увеличение на доходите би водила до увеличение на потреблението в размер равен на величината на съответния регресионния параметър, ако се абстрахираме от стохастичната компонента ε, която всъщност би могла за замъгли подобна релация.

Параметърът β_0, представляващ свободен член в (1.2), изпълнява малко по-друга предимно техническа роля. Първо, с него се елиминират евентуални разлики в началото на измерителните скали на регресорите и регресанда. Второ, разчиства се пътя за хипотезата, че математическото очакване на стохастичната компонента е нула. Това не е ограничително условие, тъй като ако се предположи, че свободният член отсъства, т.е. $\beta_0 = 0$, би трябвало математическото очакване на стохастичната компонента да се оцени по величина въз основа на статистическите данни, и ако се окаже че не е нула, би могло тази стойност да се обособи като свободен член в регресионния модел, но с обратен знак.

Статистическите данни, необходими за проиграване на регресионни модели, обикновено се състоят от многомерни статистически редове, отделните случаи на които представляват стойности за регресанда, съответни на регресорите. При изучаването на потреблението това би могло да бъде информация от текущото наблюдение на икономиката на семействата и домакинствата, широко разпространено в развитите страни. Отделните случаи в подобен статистически ред биха представлявали данни примерно за месечното потребление на съответната група стоки, както и за месечния размер на доходите на домакинствата и средните цени на тези стоки за съответния месец. Ако се означи с N обемът на такъв статистически ред (броят на домакинствата, чиито данни образуват отделните случаи), емпиричната информация, необходима за регресионния анализ, с отчитане на връзката от модела (1.2), би могла да получи следната матрично-векторна форма:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ ... \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & x_{31} & ... & x_{K1} \\ 1 & x_{12} & x_{22} & x_{32} & ... & x_{K2} \\ 1 & x_{13} & x_{23} & x_{33} & ... & x_{K3} \\ ... & ... & ... & ... & ... & ... \\ 1 & x_{1N} & x_{2N} & x_{3N} & ... & x_{KN} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ ... \\ \beta_K \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ ... \\ \varepsilon_N \end{bmatrix}. \tag{1.3}$$

Много е удобно за последното да се използва съкратен запис

$$y = X\beta + \varepsilon. \tag{1.4}$$

В него векторът y е с размерност $N \times 1$. т.е. N реда и един стълб и съдържа данните за регресанда в съответния ред на отделните случаи. Матрицата X с размери $N \times (K+1)$ е за данните за регресорите по колони като първия й стълб по конструкция се състои само от 1, съответстващ на параметъра β_0, а редовете отговарят на отделните случаи. Следва векторът на регресионните параметри с размер $(K+1) \times 1$ и накрая е векторът от отделните случаи за стохастичната компонента с размер $N \times 1$, отразяващи особеностите и спецификата на отделните единици, които водят до отклоняване в някаква степен на регресанда от онази величина, която се предписва от действието на регресорите. Векторът y и матрицата X представят статистическия материал, необходим за изучаването на зависимости като (1.4), докато векторите β и ε са условни построения с неизвестни величини и подлежат на определяне в хода на анализа.

За размерите на векторите и матриците в (1.4) може да се приеме, че N е повече от $K+1$ (на практика е много повече), така че броят на уравненията надхвърля броя на неизвестните по величина параметри. В противен случай при $N = K+1$ би се оказало, че стохастичната компонента ε е неидентифицируема, неоткриваема, неразпознаваема, вън от обхвата и възможностите на регресионния анализ и метода на най-малките квадрати. При $N < K+1$ може и част от регресионните параметри се окажат неидентифицируеми. Видно е, че обемът на използваните статистически редове е от съществено значение за успеха на приложението на методи като регресионния анализ.

Проблемът по идентификацията на елементите на регресионния модел може да възникне и при $N > K+1$, ако матрицата X има дефект в ранга, т.е. $rank(X) < K+1$. Това представлява екстремният вариант на мултиколинераност, т.е. линейна зависимост между самите регресори, между колоните в матрицата X. Последното би означавало, че някои от регресорите дублират влиянието си върху регресанда, че моделът (1.4) е пренаситен от обясняващи променливи и част от тях като своебразни плагиати би трябвало да се отстранят и индивидуалният им принос, доколкото такъв въобще има, да се отнесе към стохастичната компонента.

В практическите икономически приложения на регресионния анализ и иконометричното моделиране по-скоро може да се срещне смекчен вариант на мултиколинеарност, когато между регресорите е налице по-тясна корелация, но не и пълна линейна зависимост. Подобна ситуация не би създала сериозен проблем от математическа гледна точка, не би имало особени спънки за приложението на метода на най-малките квадрати[1], но би пострадало икономическото съдържание на резултатите с това, че липсва яснота, категоричност и убедителност за регресорите като обясняващи променливи за величината на регресанда. Както обикновено се случва при известна мултиколинеарност, оценките на регресионните параметри може да се окажат с висока дисперсия, тяхната величина би била нестабилна и съответните изводи несигурни и ненадеждни.

Икономистът обаче често е поставен пред свършени факти, в пасивната позиция на метеоролога, да анализира статистическа информация, каквато е достъпна и налична, извън

[1] Ако не се взима предвид възможната недостатъчна обусловеност (ill-conditioned) на системата нормални уравнения, което би създало изчислителни трудности дори за компютрите и понякога би породило гротескна загуба на точност в сметките, вж. Rust, B., W. Burrus (1972). *Mathematical Programming and the Numerical Solution of Linear Equations*. Modern Analytic and Computational Methods in Science and Mathematics, No 38. New York: Elsevier.

неговия контрол. При формата на модела (1.4) единственото нещо, което може да се промени в архитектурата на данните, е да се включи или не първият стълб състоящ се само от 1, съответно на хипотезата за стойността на параметъра $\beta_0 \neq 0$ или $\beta_0 = 0$. Неслучайно такъв стълб от единици се разглежда като изкуствен регресор, фиктивна или инструментална променлива. С помощта на подобни инструменти обаче може да се открие по-ефикасен модел и дори по-подходяща програма на самото статистическо наблюдение, от което се очакват данни за изследването, и да се придаде по-голяма солидност и убедителност в резултатите от съответното приложение на метода на най-малките квадрати. При лабораторни или полеви експерименти изследователят обикновено разполага с доста повече варианти. При емпиричните социологически и маркетингови проучвания също има по-широки възможности за избор на един или друг дизайн на наблюдението и съответно различна и по-подходяща структура на матрицата X. В иконометрията и особено в макроикономическото моделиране на развитието изборът е доста стеснен, а често съвсем отсъства.

Горното е свързано с един съществен момент в регресионния анализ – поведението на случайната компонента и нейните N на брой скрити присъствия в статистическите данни. При хипотезата, че векторът ε в (1.4) има нулево математическо очакване, следва въпросът за ковариацията и корелацията между неговите N на брой координати, между отделните случаи в статистическите редове. Това е един твърде деликатен и често уязвим момент при регресионния анализ и въобще при всяко приложение на метода на най-малките квадрати. Ако данните идват от анкетно наблюдение в рамките на кратък период и представляват сравнително хомогенна моментна съвкупност, може да се очаква, че векторът ε има некорелирани координати с еднаква дисперсия σ^2, макар и с неизвестна големина. При такава хипотеза ковариационната матрица на стохастичната компонента в (1.4) би имала сравнително проста структура

$$E(\varepsilon\varepsilon^T) = \sigma^2 I, \tag{1.5}$$

където I представлява единичната матрица (с размер $N \times N$) с единици по главния диагонал и нули извън него (с T е означено транспонирането и E представлява знакът за математическо очакване).

За съжаление в редица случаи подобна хипотеза може да се окаже неоснователна. Ако данните се отнасят за по-дълъг срок, цяла година или повече, и особено, ако някои единици са многократно наблюдавани, би трябвало да се очаква някаква корелация между отделните случаи и ковариационната матрица в (1.5) едва ли би била диагонална. В иконометрията това е по-скоро типичният случай, особено при макроикономическото моделиране. При него се използва информация от по-продължителни хронологични статистически редове, предимно от административни източници и статистическата отчетност. Такива редове съдържат серии от случаи, представляващи годишни данни и отнасящи се за една и съща икономика, сектор или регион и подобна многократна повтаряемост би могло да породи корелация между отделните случаи.

Отклонения от хипотезата в (1.5) обикновено възникнат по две направления. Първо, възможно е да се наруши хомоскедастичността, т.е. диагоналните елементи в ковариационната матрица да се окажат с различна големина и извеждането на общ множител σ^2 да загуби ясен смисъл. Известно е например, че с увеличаване на доходите на

домакинствата се поражда по-голямо разнообразие в потреблението им[1], което *ceteris paribus* би могло да се отрази върху елементите на ковариационната матрица и да раздвижи големината им. Този феномен на хетероскедастичност може още повече да се усложни, ако се открие зависимост на ковариационната матрица от регресорите, което би поставило въобще под съмнение линейността на регресионния модел.

Второто направление в ущърб на хипотезата е свързано с възникване на ненулеви елементи вън от диагонала на ковариационната матрица. Това би било свидетелство за автокорелация в стохастичната компонента (корелация между нейните координати), и, както вече се привлече вниманието, подобна ситуация е доста характерна за приложението на регресионния анализ при статистическо изучаване на развитието по данни от динамични редове.

В случай на автокорелация може да възникне проблем и с ранга на ковариационната матрица и в това отношение да се появи дефект. В (1.5) подобен проблем не би могло да има тъй като единичната матрица е с пълен ранг.

2. ОБЩА ПОСТАНОВКА НА ЛИНЕЙНИТЕ СТАТИСТИЧЕСКИ МОДЕЛИ[2]

По-удобно би било на модела (1.3) да се придаде следната малко по-прибрана форма без да се определя в кои колони в матрицата на регресорите се намират фиктивни и инструментални построения (допуска се да има няколко и различни комбинации от такива):

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \dots \\ y_N \end{bmatrix} = \begin{bmatrix} x_{11} & x_{21} & x_{31} & \dots & x_{K1} \\ x_{12} & x_{22} & x_{32} & \dots & x_{K2} \\ x_{13} & x_{23} & x_{33} & \dots & x_{K3} \\ \dots & \dots & \dots & \dots & \dots \\ x_{1N} & x_{2N} & x_{3N} & \dots & x_{KN} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_K \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \dots \\ \varepsilon_N \end{bmatrix}. \tag{2.1}$$

В по-лаконичен запис последното е аналогично на (1.4)

$$y = X\beta + \varepsilon. \tag{2.2}$$

като размерностите тук са малко по-други: векторите y и ε са $N \times 1$, а матрицата X и векторът β са съответно $N \times K$ и $K \times 1$. Предполага се, че използваните статистически редове за y и X са с достатъчен обем, така че $N > K$. Засега ще се възприеме, че матрицата X има пълен ранг $rank(X) = K$.

Заслужава да се отбележи, че наименованието на модела се определя от линейността спрямо параметрите β, докато какво съдържание имат регресорите в матрицата X, какви фактори се крият зад тях и в каква форма се проявяват може да е от първостепенен интерес за икономическия анализ, но няма никакво значение от гледна точка на статистическата

[1] Маленво (1975, стр. 306) обръща внимание на този факт като се позовава на изследванията на Прейс и Хаутекър (Prais, Houthakker 1955). Вж също Джонсън (1980, стр. 212).

[2] При изложението в тази и следващата част се следва Рао (1968), Кендалл, Стюарт (1973), Theil (1971), Себер Дж. (1976).

теория. Така се постига максимална универсалност на модела и широк периметър на приложение.

За математическото очакване на стохастичната компонента се предполага $E(\varepsilon) = 0$ и за ковариационната матрица засега аналогично на (1.5)

$$E(\varepsilon\varepsilon^T) = \sigma^2 I. \tag{2.3}$$

По-късно ще бъдат разгледани и други възможности относно поведението на стохастичната компонента.

Основната идея при метода на най-малките квадрати е, като се изхожда от данните в съответните статистически редове, за параметрите β да се подберат такива стойности, че да се постига минимум в сбора от отклоненията на квадрат на резултативния показател от математическото му очакване по отделните случаи в статистическия ред

$$(y - X\beta)^T (y - X\beta) \Rightarrow Min. \tag{2.4}$$

Основанието за такова търсене идва от хипотезата, че стохастичната компонента не би трябвало да оказва някакво систематично въздействие върху резултативния показател, а неговата величина да се определя главно от съчетанието на обясняващите променливи в X чрез математическото очакване $X\beta$, представляващо линейна комбинация на регресорите. Квадратичната форма (2.4), на която се търси минимумът, представлява сбор от квадратите на величини, което дава и наименованието на метода.

Намирането на този минимум е равносилно на нулирането на първата производна в лявата част на (2.4) относно β, което води до системата нормални уравнения

$$X^T X\beta = X^T y \tag{2.5}$$

Решението на това уравнение е единствено и се получава като

$$\hat{\beta} = (X^T X)^{-1} X^T y. \tag{2.6}$$

Така величината на минимума по (2.4) се получава равна на

$$(y - X\hat{\beta})^T (y - X\hat{\beta}). \tag{2.7}$$

Втората производна на лявата част в (2.4) е матрицата на Хесе и е положително дефинитна[1] (равна на $2X^T X$), което показва, че там се достига минимум. Последното може и директно да се установи от

[1] В смисъл на неотрицателно дефинитна (за разлика от строго положителна дефинитност). Това трябва да се има предвид и при по-нататъшното изложение. Когато например се говори за минималност на дисперсията, има се предвид минималност на ковариационната матрица, т.е. всяка друга би я превишавала, като разликата би представлявала положително дефинитна матрица.

$$\varepsilon^T\varepsilon = (y - X\beta)^T(y - X\beta) = (y - X\hat{\beta})^T(y - X\hat{\beta}) + (\hat{\beta} - \beta)^T X^T X(\hat{\beta} - \beta) \qquad (2.8)$$

В това разглеждане условието матрицата от регресорите да има пълен ранг е съществено, макар изцяло от техническо естество. Така се осигурява строга положителна дефинитност на матрицата $X^T X$, която е също с пълен ранг K, и съществуването на обратната матрица $(X^T X)^{-1}$. В т. 9 и по-нататък ще се покаже как може да се преодолеят подобни ограничения.

Заслужава да се отбележи, че (2.6) представлява линейна форма относно ε и y, което значително опростява нейното изследване. Освен това, там не фигурира неизвестната величина на σ^2, което освобождава метода от необходимостта предварително да се намери нейната стойност и така се облекчават разглежданията.

Получената по метода на най-малките квадрати количествена оценка $\hat{\beta}$ за стойността на регресионните параметри, макар и случайна величина (зависи от ε чрез y), притежава редица предимства и ценни статистически качества, между които:

1. В дясната част на формула (2.6) не фигурират неизвестни величини (като σ^2 например), което я прави съвсем практична.

2. $\hat{\beta}$ представлява неизместена оценка за регресионните параметри, т.е. $E(\hat{\beta}) = \beta$, което означава, че не е обременена с някаква систематична грешка, която да натежава в едната посока.

3. Дисперсията на оценката, т.е. нейната ковариационна матрица, е

$$\sigma^2(X^T X)^{-1} \qquad (2.9)$$

Във връзка с последното заслужава да се подчертае, че всяка друга линейна неизместена оценка на регресионните параметри би имала по-висока дисперсия, свидетелстваща за по-голямо разсейване и така би отстъпвала по точност на (2.6). По тази причина често за оценките по (2.6) се използва абревиатурата BLUE (Best Linear Unbiased Estimator), т.е. най-добра линейна неизместена оценка.

В приложенията често се налага вместо самите регресионни параметри да се намери оценка за някаква тяхна линейна комбинация или трансформация във формата на $\theta = C^T\beta$. Такава (представляваща BLUE) може също да се получи по метода на най-малките квадрати и се оказва равна на $\hat{\theta} = C^T\hat{\beta}$, също с минимална дисперсия сред всички линейни неизместени оценки, равна на

$$\sigma^2 C^T(X^T X)^{-1} C. \qquad (2.10)$$

Така че всяка линейна трансформация на регресионните параметри се оказва идентифицируема и притежава съответната BLUE, при условие, че матрицата X е с пълен ранг.

За да се разграничат елементите на цялото построение, често оценките по (2.6) се означават като регресионни коефициенти за разлика от регресионните параметри, чиято величина е неизвестна и се оценява по метода на най-малките квадрати. Поради наличието

на стохастичен елемент в (2.1) регресионните коефициенти се оказват случайни величини, които би трябвало да са близки до стойността на регресионните параметри, да се колебаят около тях с по възможност минимално разсейване. С метода на най-малките квадрати тази цел се постига. Това в общи линии е съдържанието на известната теорема на Гаус-Марков (Theil 1971, стр. 119; Шеффе 1980, стр. 25).

Въз връзка с това понякога се възприема като хипотеза, че стохастичната компонента има многомерно (съответно на обема на статистическите редове) нормално разпределение (на Гаус-Лаплас) с нулево математическо очакване и ковариационна матрица $\sigma^2 I$. В такива случаи и самата $\hat{\beta}$ би имала такова разпределение, но с математическо очакване $X\beta$ и ковариационна матрица $\sigma^2(X^TX)^{-1}$ (и съответно $C^T\hat{\beta}$ би имала математическо очакване $C^TX\beta$ и ковариационна матрица $\sigma^2C^T(X^TX)^{-1}C$). Разгледаните дотук свойства на оценките по метода на най-малките квадрати обаче са налице и без подобно предположение.

Нещо повече, при някои допълнителни, но не съвсем ограничителни условия, оценката (2.6) се оказва и състоятелна, т.е. с нарастване на обема на статистическите редове величината и се фокусира и се концентрира и в крайна сметка клони (съгласно закона за големите числа) към стойността на регресионните параметри. При това макар и асимптотично, разпределението й става нормално (по силата на централните гранични теореми), което широко се използва при регресионния анализ и е солидна опора при иконометричното моделиране.

Формула (2.6) дава оценка с минимална дисперсия сред възможните линейни неизместени оценки. В общия случай е възможно да съществуват неизместени, но нелинейни оценки с още по ниска дисперсия. Такива обаче трудно се откриват и тази проблематика предстои да се разработва.

При положение, че стохастичната компонента има многомерно нормално разпределение обаче, ситуацията се избистря и придобива много по-контрастен образ. Освен нормално разпределение на оценката при всякакъв обем на статистическите редове, оказва се, че тя има минимална дисперсия въобще, не само сред линейните оценки. Освен това дисперсията й достига минимума, определен от неравенството на Рао-Крамер. Формула (2.6) дава и една достатъчна оценка по Фишер, т.е. в нея се съдържа цялата информация за регресионните параметри и в този смисъл оценката е изчерпателна, и заедно с това отговаря на принципа на максималното правдоподобие. Всичко това в равна степен важи и за линейните трансформации на регресионните параметри $\theta = C^T\beta$, каквато и да е матрицата C, и придава значителна тежест на понятието, скрито зад абревиатурата BLUE.

В противоположната посока се откриват опити да се намерят практични, но изместени оценки. Такива са например асимптотично неизместените оценки, при които с нарастване обема на статистическите редове изместеността се стопява. Друг вариант са т.нар. хребетни оценки от вида на $b = (X^TX + kI)^{-1}X^Ty$ при подходящ избор на величината на k. Идеята при тях е с допускане на малка изместеност да се постигне значително намаление при дисперсията, така че общото квадратично отклонение от величината на регресионните параметри да се окаже дори по-малко от това, получено по (2.6).

3. ОЦЕНКА НА СТОХАСТИЧНАТА КОМПОНЕНТА

Оценката на регресионните параметри чрез формула (2.6) е изходен пункт за намиране и на останалите неизвестни величини в модела (2.1). Оценките за стохастичната компонента се получават, като се сравнят фактическите данни y с това, което се предвижда по модела $\hat{y} = X\hat{\beta}$

$$\hat{\varepsilon} = y - \hat{y} = (I - X(X^T X)^{-1} X^T) y \tag{3.1}$$

Така оценената стохастична компонента се оказва вектор ортогонален на регресорите в X, което насочва към една съдържателна геометрична интерпретация на метода на най-малките квадрати[1]: фактическите стойности на регресанда като вектор се явяват хипотенуза в правоъгълен триъгълник с катети $\hat{y}$ и $\hat{\varepsilon}$, предвид което

$$y^T y = \hat{y}^T \hat{y} + \hat{\varepsilon}^T \hat{\varepsilon} = (X\hat{\beta})^T (X\hat{\beta}) + (y - X\hat{\beta})^T (y - X\hat{\beta}). \tag{3.2}$$

Матрицата $P = X(X^T X)^{-1} X^T$ е симетрична и идемпотентна[2] и с нея се осъществява проектирането на вектора y в пространството от всички линейни комбинации на регресорите, което като резултат дава "сянката" $Py = X\hat{\beta}$ на фактическите данни в пространството на регресорите. С матрицата $I - P$ (която е също симетрична и идемпотентна) пък се осъществява проекцията на y в ортогоналното допълнение на регресорите и така се образува $\hat{\varepsilon}$. Това показва, че $\hat{\varepsilon}$ е проекция и на стохастичната компонента ε в същото ортогонално допълнение, осъществена пак от $I - P$

$$\hat{\varepsilon} = (I - P)\varepsilon. \tag{3.3}$$

Последното се използва, за да се получи

$$E(\hat{\varepsilon}^T \hat{\varepsilon}) = (N - K)\sigma^2, \tag{3.4}$$

така че

$$\hat{\sigma}^2 = \frac{1}{N-K} \hat{\varepsilon}^T \hat{\varepsilon} = \frac{1}{N-K} (y - X\hat{\beta})^T (y - X\hat{\beta}) \tag{3.5}$$

се явява неизместена оценка за дисперсията σ^2 на стохастичната компонента, която обикновено се нарича остатъчна дисперсия. Остатъчността идва от това, че общата вариация на регресанда се намалява с включването на регресорите като обясняващи математическото й очакване, а което остава необяснено, се отнася към случайната компонента. Това представлява статистическият смисъл на (3.2).

---

[1] Дадена за пръв път у Колмогоров (1946).

[2] За симетричните матрици е валидно $A^T = A$, за идемпотентните $A^2 = A$.

Величината $N-K$ представлява степените на свобода на остатъчната дисперсия. Колкото N е повече от K, толкова повече възможности дават статистическите данни за оценката на дисперсията на стохастичната компонента. Обратно, при $N \le K$ такива въобще няма, и, както вече бе обърнато внимание, стохастичната компонента става напълно неидентифицируема и (3.5) губи всякакъв смисъл.

За дисперсията на (3.5) се получава $\frac{2\sigma^4}{N-K}$, което разкрива в каква степен стохастичната компонента влияе върху величината на регресанда. Същевременно показва, че с нарастване на обема N на статистическите редове, оценката (3.5) става все по-точна.

Като се има предвид (2.9), може да се отбележи, че формула (3.5) дава възможност и за неизместена оценка на ковариационната матрица на $\hat{\beta}$:

$$\hat{\sigma}^2 (X^T X)^{-1}. \tag{3.6}$$

Освен че е винаги неизместена, оценката $\hat{\sigma}^2$ на дисперсията на стохастичната компонента в повечето ситуации се оказва и състоятелна. Минималност на дисперсията обаче тук не е така универсално свойство, както е при оценката на регресионните параметри. Същото се отнася и за повечето от останалите качества, разгледани по-горе във връзка с (2.6). Причината за това, е че $\hat{\sigma}^2$ не е линейна оценка, а квадратична относно регресанда, което усложнява проблема. Така хипотезата за нормално разпределение на стохастичната компонента тук е по-наложителна. При такова положение (3.5) дава неизместена оценка с минимална дисперсия без никакви условности, а и повечето от останалите свойства също се оказват налице. Най-съществено за статистическия анализ с помощта на модела е, че при това положение разпределението на

$$\frac{N-K}{\sigma^2}\hat{\sigma}^2 = \frac{\hat{\varepsilon}^T \hat{\varepsilon}}{\sigma^2} \tag{3.7}$$

би било хи-квадрат със степени на свобода $N-K$ и освен това оценките $\hat{\beta}$ и $\hat{\sigma}^2$ се оказват независими една от друга случайни величини.

При хипотезата за нормално разпределение на стохастичната компонента може още да се отбележи, че

$$\frac{(y - X\hat{\beta})^T (y - X\hat{\beta})}{\sigma^2} \text{ и } \frac{(\beta - \hat{\beta})^T X^T X (\beta - \hat{\beta})}{\sigma^2} \tag{3.8}$$

биха били независими с хи-квадрат разпределение и със степени на свобода съответно $N-K$ и K. По такъв начин (2.8) се явява илюстрация на теоремата на Кокрън (Кендалл, Стюарт 1966).

Едновременно с това и (3.2) представлява сбор от независими хи-квадрат разпределени величини, макар и не всичките централни. Относителният дял на обяснената част $\hat{y}^T\hat{y}$ в общата вариация $y^T y$

$$R^2 = \frac{\hat{y}^T\hat{y}}{y^T y} \tag{3.9}$$

е показателен за това доколко моделът (2.1) отговаря на действителността и представлява коефициента на детерминация на регресанда от регресорите. Колкото по-близо до единица е величината на R^2, толкова по-адекватен е моделът и съответно стохастичната компонента има по-малък ефект върху резултативната величина на регресанда. Квадратен корен от коефициента на детерминация е известен като коефициент на корелация.

4. ОБХВАЩАНЕ НА ДОПЪЛНИТЕЛНА ИНФОРМАЦИЯ

При специфицирането на конкретен статистически линеен модел в хода на практическите приложения неизменно се използва натрупаният вече опит и познанието за разглежданата зависимост преди още да се привлекат съответни статистически данни. Най-вече какъв показател да се възприеме за регресанд и какви като обясняващи величината на регресанда променливи. При формулирането на стартовата позиция за статистическия анализ често се предвижда методът на най-малките квадрати да се прилага с отчитане на определени особености, при по-специални условия с известни изисквания или ограничения за параметрите на линейния модел. Такива може да са продиктувани от икономическата природа на изучваното явление като например функциите на Коб-Дъглас, при които сборът от степенните показатели, явяващи се регресионни параметри за разходите на труд и капитал, трябва да дава единица за да се осигури независимост на възвръщаемостта от мащаба. Друг пример често се среща при приложенията на регресионния анализ, когато някой от регресорите е под съмнение дали мястото му е в обясняващата компонента, т.е. в матрицата X. Последното е равносилно на хипотезата съответният регресионен параметър да е нула.

Минимизирането на квадратичната форма, представляваща сбора от квадратите на отклоненията при ограничения за регресионните параметри от вида $C^T\beta = \theta$, може да се получи примерно с помощта на множителите на Лагранж и резултатът от това би бил нова оценка за регресионните параметри по следната формула:

$$\hat{\beta}_C = \hat{\beta} + (X^T X)^{-1} C (C^T (X^T X)^{-1} C)^{-1} (\theta - C^T \hat{\beta}), \tag{4.1}$$

в която фигурира първоначалното $\hat{\beta}$ и се вижда каква модификация настъпва с отчитане на априорното изискване (второто събираемо в дясната част на равенството). Непосредствено може да се провери, че така получената оценка отговаря на предварителните условия: $C^T\hat{\beta}_C = \theta$.

Оценката (4.1) е неизместена с минимална дисперсия (BLUE), изцяло в духа на теоремата на Гаус-Марков и нейната ковариационна матрица, показваща лабилността й, е равна на

$$\sigma^2((X^TX)^{-1} - (X^TX)^{-1}C(C^T(X^TX)^{-1}C)^{-1}C^T(X^TX)^{-1}). \quad (4.2)$$

Формула (4.1) показва, че новата оценка е с линейна форма относно разликата $\theta - C^T\hat{\beta}$, така че ако безусловната оценка $\hat{\beta}$ по една случайност отговаря на допълнителните условия, ще има съвпадение $\hat{\beta}_C = \hat{\beta}$.

Минимумът (2.7) в сбора от квадратите с отчитане на ограниченията би бил

$$(y - X\hat{\beta}_C)^T(y - X\hat{\beta}_C) = y^T(I-P)y + (\theta - C^T\hat{\beta})^T(C^T(X^TX)^{-1}C)^{-1}(\theta - C^T\hat{\beta}), \quad (4.3)$$

където $P = X(X^TX)^{-1}X^T$ е матрицата на проектирането на регресанда в пространството на регресорите, разгледано във връзка с (3.2) и (3.3). Вижда се, че с отчитането на допълнителните ограничения се стига до известно увеличение на безусловния минимум $y^T(I-P)y$ с величината на

$$(\theta - C^T\hat{\beta})^T(C^T(X^TX)^{-1}C)^{-1}(\theta - C^T\hat{\beta}), \quad (4.4)$$

представляваща положително дефинитна квадратична форма.

Среща се следната доста нагледна интерпретация на горните формули. Без ограничения за регресионните параметри минимумът на сбора на квадратите на отделните случаи за стохастичната компонента (Residual Sum of Squares) се получава съгласно (2.7) като

$$RSS = (y - X\hat{\beta})^T(y - X\hat{\beta}) = y^T(I-P)y. \quad (4.5)$$

С въвеждане на допълнителни ограничения за регресионните параметри този сбор се повишава и съгласно (4.3) получава нова стойност RSS_C. Размерът на прираста $RSS_C - RSS$ се дава с (4.4).

Рангът на матрицата C^T играе скрита, но съществена роля в тези разглеждания. Нека размерът на матрицата е $M \times K$, така че векторът θ би трябвало да бъде с размерност $M \times 1$. Може да се покаже, че всички съдържателни варианти линейни ограничения на регресионните параметри може да се сведат до $rank(C) = M < K$, в резултат на което матрицата $C^T(X^TX)^{-1}C$ има обратна, формули (4.1) и (4.3) са валидни и се гарантира положителната дефинитност на (4.4). Условието за максимален ранг обаче има само технически характер и може лесно да се преодолее, както е показано в т. 9 и по-нататък.

Понякога априорната информация за зависимостите може да има формата на ограничения от вида $C^T\beta \le \theta$, а не непременно равенства. Макар че тази задача пред метода на най-малките квадрати е доста по-сложна, може все пак да се трансформира в

линеен модел от вида на (2.1), но с много повече регресионни параметри (Себер 1976). Последното би могло да стесни и ограничи практическото приложение на такива варианти, тъй като за решаването на подобна задача ще са необходими статистически данни в значителен обем. Предварителни ограничения за регресионните параметри с още по-сложна структура пък може да се окажат несъвместими с линейността на модела.

5. СЕЛЕКЦИЯ НА РЕГРЕСОРИТЕ В ЛИНЕЙНИТЕ СТАТИСТИЧЕСКИ МОДЕЛИ

При статистическия анализ с линейните статистически модели ограничения върху регресионните параметри от вида на $C^T\beta = \theta$ често имат хипотетичен характер и съответната задача е да се установи доколко фактическите данни подкрепят такива положения. Решението на подобни задачи обаче е по-чувствително към капризите на стохастичната компонента в линейните модели и затова често се предполага, че последната има нормално разпределение или поне близко до него, така че да се разчита на ефекта на законите за големите числа и централните гранични теореми.

При такива условия дясната част на (4.3) съответства на сбор от две независими квадратични форми с хи-квадрат разпределение (по теоремата на Кокрън) и следователно показателят

$$F_{M,N-K} = \frac{\frac{1}{M}(\theta - C^T\hat{\beta})^T (C^T(X^TX)^{-1}C)^{-1}(\theta - C^T\hat{\beta})}{\hat{\sigma}^2} \tag{5.1}$$

би имал разпределение на Фишер-Снедекор (известно още и като *F*-разпределение) със степени на свобода M и $N-K$. Така, ако при конкретните практически приложения се окаже, че величината на (5.1) навлиза във върховите перцентили на съответното *F*-разпределение, такъв факт би означавал, че използваните данни опровергават хипотезата $C^T\beta = \theta$. С други думи, второто събираемо в дясната част на (4.3) достатъчно натежава относително първото, така че увеличението в сбора от квадратите на остатъците в сравнение с първоначалния им размер е статистически значимо. Същото би могло да се изрази и другояче, като се каже, че ако хипотетичното ограничение за регресионните параметри отговаряше на действителността, представена от конкретния статистически материал, не би трябвало да се получават толкова високи и малко вероятни стойности за показателя (5.1).

На (5.1) може да се придаде по-друга форма като се използва (4.1)

$$F_{M,N-K} = \frac{\frac{1}{M}(\hat{\beta} - \hat{\beta}_C)^T X^T X(\hat{\beta} - \hat{\beta}_C)}{\hat{\sigma}^2} \tag{5.2}$$

или

$$F_{M,N-K} = \frac{\frac{1}{M}(RSS_C - RSS)}{\frac{1}{N-K}RSS}. \tag{5.3}$$

Тази техника често се използва при регресионния анализ за установяване на значимостта на отделните регресори като обясняващи величината на регресанда. Като първа крачка обикновено се подлага на тест хипотезата, че всички регресионни параметри са нула, т.е че регресорите като цяло не оказват влияние върху регресанда. При такова положение (5.2) би придобило следния вид:

$$F_{K,N-K} = \frac{\frac{1}{K}(X\hat{\beta})^T(X\hat{\beta})}{\hat{\sigma}^2}. \tag{5.4}$$

Същото може да се представи и с коефициентите на детерминация (3.9)

$$F_{K,N-K} = \frac{\frac{1}{K}R^2}{\frac{1}{N-K}(1-R^2)}, \tag{5.5}$$

с което да се установява дали величината на коефициентите на детерминация е достатъчно близко до единица, свидетелстващо за тясна корелационна зависимост, и така се оправдава конкретният избор за регресорите.

При удачен избор на обясняващи променливи хипотеза, че регресорите като цяло не отговарят на очакванията, обикновено не издържа пред статистическите данни, освен ако някаква мултиколинеарност сред регресорите или някое друго нарушение на условията за прилагане на метода не компрометира цялата конструкция. При статистическото изучаване на потреблението например в зависимост от доходите на домакинствата, цените на стоките и услугите и други фактори обикновено се оказва, че величината на (5.4) е изключително висока. В такива случаи и коефициентът на детерминация показва висока стойност, доближаваща 1 и свидетелстваща за сполучливо първо попадение. Последното означава, че така подбраните регресори като цяло оправдават присъствието си в регресионния модел, което обикновено се постига на основата на познанието за икономиката на домакинствата преди още да са използвани някакви статистически данни. В конкретната ситуация обаче възниква въпросът дали всичките използвани регресори са действително обясняващи променливи или част от тях би могло да се изключат от модела като несъществени и "паразитиращи" в обясняващата компонента. При такава постановка линейният статистически модел би могъл да се представи по-подробно по следния начин:

$$y = \left[X_1 \;\vdots\; X_2\right]\left[\begin{array}{c}\beta_1 \\ \cdots \\ \beta_2\end{array}\right] + \varepsilon, \tag{5.6}$$

където регресорите са разделени на две части, в две групи колони, като на съмнение подлежи втората група X_2 и за нея се поставя въпросът дали допринася съществено за обяснението на размера на потреблението, или просто трябва да отпадне. Съответно на това и регресионните параметри може да се разглеждат разделени на две групи, като по такъв начин моделът (5.6) би изглеждал така

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon. \tag{5.7}$$

Нека матриците X_1 и X_2 имат съответно $K-M$ и M колони. Със същите размери е разделен на две части и векторът на регресионните параметри, но по редове. Отпадането на втората група регресори е равносилно на установяването, че $\beta_2 = 0$. Това е линейно ограничение за регресионните коефициенти във формата $C^T\beta = \theta$, където

$$C^T = \left[O \mid I \right]. \tag{5.8}$$

Матрицата C^T е с размери $M \times K$. Левият й участък O с размери $M \times (K-M)$ е запълнен само с нули, а десният I представлява единична квадратна матрица с размерност $M \times M$, от което следва, че $rank(C) = M$. Векторът θ също се състои само от нули и е с размери $M \times 1$.

С помощта на разгледания по-горе критерий с използване примерно на (5.3) може да се установи дали статистическите данни подкрепят хипотезата за несъщественост на втората група регресори и ако това се потвърди, следва да се приеме нова по-адекватна на действителността форма на модела

$$y = X_1\beta_1 + \varepsilon, \tag{5.9}$$

след отпадане на излишното $X_2\beta_2$ в (5.7).

Тази процедура би могло да се повтори, като се подложат на тест някои от регресорите в първата група X_1 и да се продължи с отсяването на несъществените регресори. Така би могло да се намери най-удачната форма на регресионния модел, при която фигурират само съществени регресори и тяхната обясняваща функция изпъква най-добре. Това представлява т.нар. стъпков или итерационен метод за последователно елиминиране на излишните регресори без да пострада обясняващата част на регресионните модели. Нещо повече, по такъв начин обясняващата функция на модела би изпъкнала повече, би се получила по-контрастна картина на изследваната зависимост.

Възможно е итерациите да се извършват и в обратната посока. Започва се с регресори, които не предизвикват съмнение като обясняващи променливи, като за всеки случай може да се приложи и критерият, свързан с (5.4). После се пристъпва към последователно разширяване на модела с включване на допълнителни обясняващи променливи като нови регресори, като на всяка крачка се проверява дали е постигнат съществен прогрес, а не се внася просто шум в обясняващата компонента на модела.

Съществуват и други подходи за селекция на съществените обясняващи променливи като например метода на изчерпателно апробиране на всички комбинации от възможни регресори. Съответните алгоритми, както и на разгледаните итерационни методи, са налице при съвременния статистически софтуер като специализирани статистически процедури.

Съблазнително изглежда при селекцията на регресорите те да се подредят по значимост и да се отмери по някакъв начин приносът на всеки от тях като фактор за величината на регресанда, при това сборът от приносите да се равнява на общия ефект от обясняващата компонента. За съжаление линейният статистически модел не дава такава възможност освен в някои много специални и изключителни случаи. Причината за това се долавя в методиката на статистическия анализ, при процедурите по проверка на хипотези, където става сравняване на множество от регресори с някоя негова част, така че няма възможност за

определяне на относителния дял на отделните фактори за обяснение величината на регресанда. Просто няма как един регресор да се явява част от друг. Във връзка с това следва да се подчертае, че и методът на изчерпателно апробиране на всички комбинации от възможни регресори страда от подобна невъзможност за сравнение.

Препятствието в тази насока се крие във факта, че ефектът на отделните регресори се филтрира, така да се каже, през регресионните параметри, чиято големина в крайна сметка детерминира и приноса им за величината на регресанда. Освен това обаче и вариацията на регресорите определя тяхната относителна тежест. Незначителна на пръв поглед величина на регресионните параметри може да съответства на значима, дори доминираща обясняваща променлива, ако нейната вариация е достатъчно голяма.

Приносът на отделните регресори при обясняване величината на регресанда може да се определи непосредствено като относителен дял от общото влияние само при ортогонален дизайн на модела, т.е. когато регресорите са ортогонални един на друг. При такова положение матриците $X^T X$ и $(X^T X)^{-1}$ стават диагонални и координатите на регресионните коефициенти се оказват некорелирани помежду си (съгласно (2.9)), откривайки по такъв начин възможност за определяне на индивидуалната тежест на отделните регресори в обясняващата компонента на модела.

6. КОРЕЛАЦИОНЕН АНАЛИЗ

Проблемът по определянето на тежестта на отделните регресори при обясняване величината на регресанда е в предмета на корелационния анализ. Ефектът от отделните регресори може да бъде директен върху големината на регресанда и паралелно с това – косвен, чрез зависимостите с останалите регресори и оттам върху резултата. Както вече се привлече вниманието, само при ортогонален дизайн последните зависимости не съществуват и така непосредствено се открива значимостта на отделните регресори и лесно се определя относителният им дял в обясняващата компонента.

Подобна плетеница от взаимозависимости може да се разгледа при модела (5.6) и съкратения му вариант (5.9). За простота на изложението ще се счита, че втората група регресори X_2 се състои само от един, т.е. $M = 1$. При хипотезата, че този регресор не допринася съществено за обяснение на величината на регресанда и е статистически незначим, моделът се превръща в (5.9). За проблема може да се мисли и в обратен ред. Моделът (5.9) дава определено обяснение как варира регресанда в зависимост от първата група регресори X_1 със съответен коефициент на детерминация R_1^2. Добавянето на нов регресор в обясняващата компонента би довело до определено увеличение на величината на коефициента на детерминация до R^2. Прирастът $R^2 - R_1^2$ може обаче да се окаже недостатъчен и съответните показатели (5.2) и (5.3) да разкриват статистически незначим ефект и да опровергават включването на допълнителен регресор. Във връзка с това на критерия (5.3) би могло да се придаде следната форма:

$$F_{1,\,N-K} = \frac{(R^2 - R_1^2)}{\frac{1}{N-K}(1-R^2)}. \qquad (6.1)$$

Последното е всъщност квадрат на разпределението на Стюдънт (известно още и като *t*-разпределение) поради степените на свобода в числителя, равни на единица $M = 1$. Разпределението на Стюдънт е възникнало исторически по-рано от *F*-разпределението и (6.1) често се среща като

$$t_{N-K} = \sqrt{(N-K)\frac{R^2 - R_1^2}{1-R^2}}. \tag{6.2}$$

Логично е прирастът $R^2 - R_1^2$ да се съпостави с онази част от вариацията на регресанда, която остава необяснена от първата група регресори и която би трябвало да бъде дообяснена с включването на допълнителни регресори

$$r_1^2 = \frac{R^2 - R_1^2}{1-R_1^2}. \tag{6.3}$$

Това представлява частният коефициент на детерминация, показващ доколко регресорът X_2 е свързан с регресанда y, независимо от връзките му с първата група регресори X_1. Последното често се възприема като пряка, директна, нетна корелационна зависимост. Към подобна интерпретация обаче би трябвало да се подхожда по-предпазливо като се има предвид, че релацията е фактически опосредствана от всички останали фактори, които не фигурират дори в разширения модел (5.6) и по такъв начин разглежданата зависимост попада под влиянието на околната среда. Така че нетните корелационни зависимости са и брутни, макар и в неявна форма.

На показателя (6.1) може да се придаде друга форма, като се използва частния коефициент на детерминация (6.3)

$$F_{1,\,N-K} = (N-K)\frac{r_1^2}{(1-r_1^2)}. \tag{6.4}$$

И съответно на (6.2) би се получило познатото

$$t_{N-K} = \sqrt{N-K}\,\frac{r_1}{\sqrt{1-r_1^2}}. \tag{6.5}$$

Специален вариант представлява единичният коефициент на детерминация, който се получава при отсъствие на групата X_1, т.е., когато се разглежда само един регресор като обясняваща променлива. Това отговаря на модела на единичната регресия и често се възприема като брутна релация, тъй като върху нея индиректно влияят необхванатите в схемата фактори, макар и прикрити в стохастичната компонента.

Извличането на квадратен корен от коефициентите на детерминация дава коефициентите на корелация и има по-ясен смисъл само при единичните частни коефициенти от вида на (6.3), както и при единичните брутни. Причината за това е, че само при тези случаи може да се определи посоката на корелационната зависимост, позитивна

или негативна, и така да се получи знакът на тези показатели – плюс или минус. Типичен пример е (6.5) като квадратен корен на (6.4).

При $M > 1$ се получават частни множествени коефициенти на детерминация по формули аналогични на (6.3). Въобще при множествените коефициенти на детерминация, било частни или общи, въпросът за съответните коефициенти на корелация има по-особен смисъл. Причината за това е, че регресорите в общи линии действат разнопосочно, поради което не може да се определи и знакът на корелационната зависимост, не може да се каже дали зависимостта е позитивна или негативна. Всъщност множествените коефициенти на детерминация и корелация са единични такива между резултативния показател и неговата проекция върху регресорите, получена по метода на най-малките квадрати.

А и самото понятие за позитивна и негативна корелационна зависимост е доста условно, тъй като със смяна на измерителните скали и мерните единици на използваните показатели позитивното може да се трансформира в негативно или обратното. Освен това и критериите и процедурите за проверка на статистическата значимост (свързани с показатели от вида на (5.5) и (6.1)) показват, че значение имат само коефициентите на детерминация, т.е. квадратите на коефициентите на корелация.

Един друг поглед върху частните коефициенти на детерминация/корелация може да се получи по следния начин. Най-напред се разглежда регресионният модел (5.9), в резултат на което обяснената величина на регресанда би се получила като

$$\hat{y} = X_1(X_1^T X_1)^{-1} X_1^T y, \tag{6.6}$$

като за необяснената част (съгласно (3.1)) остава

$$y - \hat{y} = (I - X_1(X_1^T X_1)^{-1} X_1^T)y. \tag{6.7}$$

и предстои тези величини да се отнесат към втората група регресори и да се види доколко включването на X_2 може да подобри обяснението за величината на регресанда.

Коефициентът на детерминация съответно на (6.6) би бил

$$R_1^2 = \frac{y^T X_1(X_1^T X_1)^{-1} X_1^T y}{y^T y}. \tag{6.8}$$

За да се прецени доколко вторият регресор е статистически значим, т.е. да се подложи на проверка съответната хипотеза, логично е най-напред да се определи доколко вариацията в X_2 е свързана с X_1 и като такава не би могло да допринесе за подобрение при разширяване на модела (5.9). Съответният регресионен модел би разкрил каква част от вариацията на втория регресор зависи от първата група

$$\hat{X}_2 = X_1(X_1^T X_1)^{-1} X_1^T X_2, \tag{6.9}$$

като необяснената част

$$X_2 - \hat{X}_2 = (I - X_1(X_1^T X_1)^{-1} X_1^T)X_2, \tag{6.10}$$

бидейки некорелирана с първата група регресори, крие самостоятелното въздействие на X_2 върху величаната на регресанда y, т.е. тъкмо онова, което допълнителният регресор би дал съществено в повече при разширяването на първоначалния модел (5.9) до (5.7).

Остава да се види по какъв начин необяснената част (6.7) се отнася към (6.10) за да се разкрие самостоятелното въздействие на X_2 върху величината на y, освободено от сянката на останалите регресори в X_1. Този регресионен модел би дал коефициент на детерминация, съвпадащ с частния коефициент между X_2 и y, при изключено действие на първата група регресори X_1.

7. ИНТЕРВАЛНО ОЦЕНЯВАНЕ И ДОВЕРИТЕЛНИ ОБЛАСТИ

Точковите оценки (point estimators) по метода на най-малките квадрати, свързани с формули (2.6) и (3.5), изглеждат като твърде амбициозна задача, тъй като получените величини, макар и близки до неизвестните β и σ^2 (благодарение на законите за големите числа и централните гранични теореми), едва ли биха съвпадали с тях. Нещо повече, дори и ако стохастичната компонента би имала нормално разпределение, всякакво съвпадение би било невероятно, направо изключено. Алтернативата на този подход е свързана с получаване на интервални оценки (confidence intervals), с които се цели да се покриват неизвестните стойности, но като диапазон. С други думи оценките да не звучат като "равно на", а по-скоро – "в границите на" или "от – до". Границите на доверителните интервали са също случайни величини, тъй като се пресмятат по данните за регресорите и регресанда, но са целенасочено изместени в двете посоки, така че да образуват диапазон, включващ неизвестната величина на оценявания параметър. Последното разбира се не е съвсем сигурно и съществува известна вероятност търсената величина да се "изплъзне" от доверителните области, да се окаже извън калкулираните граници. Това е т. нар. грешка от първи род, чиято вероятност обикновено се означава като α. Тази величина обаче може да се контролира и на практика се избира да бъде в граници под 10%, най-често 5%.

Доверителни области за регресионните параметри, както и на техни линейни трансформации от вида на $\theta = C^T\beta$, се получават от (5.1) по следния начин. При определено ниво за грешката от първи род α се определя съответния перцентил на *F*-разпределението $F_{M,N-K}(\alpha)$, така че вдясно от тази величина да попадат точно α дял от възможностите. При такова положение неравенството

$$\frac{\frac{1}{M}(\theta - C^T\hat{\beta})^T(C^T(X^TX)^{-1}C)(\theta - C^T\hat{\beta})}{\hat{\sigma}^2} \le F_{M,N-K}(\alpha) \qquad (7.1)$$

би определяло търсената доверителна област на приемливите стойности за θ. Допълнителната вероятност $1-\alpha$ се възприема като коефициент на доверие и фактически показва в каква степен доверителната област би покривала оценяваната величина.

Геометричното място на приемливите стойности, отговарящи на неравенството (7.1), представлява елипсоид в пространството на всички възможности за θ. Върху размера на доверителния елипсоид влияят редица условия, вкл. наличието на едни или други степени

на свобода, и преди всичко резултатите, получени по метода на най-малките квадрати: оценките $\hat{\beta}$ и $\hat{\sigma}^2$. Така например по-малка величина за остатъчната дисперсия би довела до по-стегнат елипсоид, а с това и до по-конкретни, по-точни и по-съдържателни изводи, базирани на доверителната област. Последното обаче, в общи линии, е извън контрола на статистиката и единственият най-близък механизъм за директно въздействие и управление се оказва само вероятността за грешка от първи род. Нейната величина може да се намалява, с което да става все по-правдоподобно доверителният елипсоид да обхване неизвестната величина на θ, но в тази посока има и подводни камъни. Намаляването на риска от грешка от първи род всъщност води до увеличаване на обема на елипсоида (7.1) и това е в сериозен ущърб на информативността на доверителните области. При $\alpha = 0$ например доверителният елипсоид се оказва разтегнат до безкрай и фактически изпълва пространството на всички възможности за θ, което означава, че всъщност никакъв резултат не е получен, пропиляна е цялата информация от използваните данни. Последното представлява т. нар. грешка от втори род, макар и в екстремния си вариант на абсолютна баналност. Излиза, че колкото повече се подсигурява срещу грешка от първи род, толкова повече се излага на грешка от втори род, т.е. на риска да се олекотят и разводнят постигнатите резултати, макар и да се повишава тяхната достоверност. Тази реципрочност е характерна за противоречието между двата рода грешки и представлява съществен момент в теорията на Джърси Нейман и Егон Пирсън за доверителните интервали и статистическата проверка на хипотези.

Взаимоотношението между двата рода грешки се открива и при проверката на хипотези от разгледания вече вид на $C^T\beta = \theta$. В случай, че величината на (5.1) надхвърли критичната граница (попада примерно в последните $\alpha = 5\%$ на разпределението), следва хипотезата да се отклони, макар че има само частична гаранция за правилността на такова заключение. Всъщност гаранцията е "само" $1-\alpha$-процентова. В действителност хипотезата би могло да бъде съвсем справедлива и въпреки това съществува риск, макар и съвсем случайно, т.е. с вероятност α, величината на (5.1) да се окаже прекалено висока и погрешно да се отклони хипотезата. Това представлява грешката от първи род при статистическата проверка на хипотези. Критичната за хипотезата област в случая е външността на доверителния елипсоид и α е вероятността величината на (5.1) да попадне в нея, въпреки действителността на хипотезата.

Грешката от втори род би се получила, ако се възприеме невалидна хипотеза. В теорията на статистическата проверка на хипотези се изяснява каква да бъде критичната област, така че да се минимизира рискът от допускане на грешка от втори род при зададен лимит за α. Особено предимство на метода на най-малките квадрати се състои в това, че критичната област (външността на доверителния елипсоид) в общи линии се явява равномерно най-мощна и в такъв смисъл процедурата, свързана с (5.1), е оптимална. Впрочем техниката на доверителните области и статистическата проверка на хипотези изглеждат като две лица на един и същ статистически метод.

В редица приложения, най-вече в експерименталната статистика – полевите опити, социологическите проучвания и т.н., степените на свобода също може да се направляват чрез намирането на по-подходящ проект и план на статистическото наблюдение с оглед на постигането на по-достоверни и едновременно с това и по-съдържателни резултати. Началото на това направление в статистиката е поставено от Роналд Фишер. По такъв начин може реципрочността на двата рода грешки да се измести на друга орбита и да се омекоти ефектът от омагьосания им кръг.

Методът на доверителните интервали и статистическата проверка на хипотези се прилага и за коефициентите на детерминация. За целта може да се използва (5.5) вместо (5.1). За частните коефициенти на детерминация доверителни интервали и съответните критични области се получават от (6.4). Доверителните интервали за единичните частни коефициенти на корелация например са тясно свързани със съответните за регресионните параметри. Статистическа проверка на хипотези за незначимост на отделен регресор е равносилна на установяване на частна некорелираност между регресора и регресанда, в който случай съответните доверителни интервали би трябвало да включват нулата. Тази кореспонденция между разгледаните инструменти на линейния статистически анализ може да се представи със следната таблица:

Линейни хипотези за регресионните параметри	**Коефициенти на детерминация и корелация**	**Доверителни области**
1	2	3
Касаещи всички регресионни параметри	Множествен коефициент на детерминация	Доверителни елипсоиди за съвкупността от регресионните параметри, Доверителни интервали за множествения коефициент на детерминация
Касаещи част от регресионни параметри	Множествени частни коефициенти на детерминация	Доверителни елипсоиди за тази част от регресионни параметри и съответните множествени частни коефициенти на детерминация
Касаещи един от регресионни параметри	Единични частни коефициенти на детерминация и корелация	Доверителни интервали за регресионния параметър и съответните частни коефициенти на детерминация и корелация

Продължение

1	2	3
Касаещи регресионния параметър в модела на единична регресия	Единични коефициенти на детерминация и корелация	Доверителни интервали за регресионния параметър, единичните коефициенти на детерминация и корелация

Доверителни интервали за стохастичната компонента, по-точно за нейната дисперсия σ^2, се получават от (3.7). Съгласно това остатъчната сума от квадратите има хи-квадрат разпределение и 95%-ов доверителен интервал примерно би се получил, като се използват крайните 2,.5%-ви перцентили на хи-квадрат разпределение със степени на свобода $N-K$

$$\frac{(N-K)\hat{\sigma}^2}{\chi^2_{0.975}} < \sigma^2 < \frac{(N-K)\hat{\sigma}^2}{\chi^2_{0.025}}. \tag{7.2}$$

Техниката на доверителните интервали, както и статистическата проверка на хипотези, са доста чувствителни по отношение разпределението на стохастичната компонента. Методологията е най-изчистена, ако стохастичната компонента има нормално раз-

пределение. Доста съществено е например оценките за регресионните параметри $\hat{\beta}$ и за остатъчната дисперсия $\hat{\sigma}^2$ да бъдат независими, за да се получат съответните F-разпределения (и t-разпределения), на които се базира целият метод. Последното се осигурява експлицитно при нормално разпределение на стохастичната компонента.

8. ЛИНЕЙНИТЕ СТАТИСТИЧЕСКИ МОДЕЛИ КАТО МЕТОД ЗА ПРОГНОЗИРАНЕ

Една от най-важните задачи на статистиката е да се очертае бъдещето развитие на процесите, като се изхожда от фактическия материал за тях, от наблюдението върху тях в настоящето и близкото минало. Линейните статистически модели позволяват подобно екстраполиране на констатираните тенденции, което ги прави подходящ инструмент при иконометричните изследвания и прогностичните разчети. Прогнозирането обаче би могло да бъде както във времето, така и в пространството. Последното се проявява при репрезентативните проучвания, където информацията от извадката се екстраполира за цялата генерална съвкупност с отчитане на възможната стохастична грешка. Би могло например да се мисли за възможността по данни от постоянно действащите наблюдения върху икономиката на домакинствата да се прецени каква промяна може да настъпи в разходите на населението в генералната съвкупност при един или друг размер на домакинствата, при определено движение на цените и доходите и други хипотези, които не са се появили в извадките.

Прогнозирането при линейните статистически модели е в известен смисъл условно построение и изглежда по следния начин. По данни за регресорите и регресанда, примерно за някакъв базисен период, се определят елементите на модела чрез регресионните коефициенти, остатъчната дисперсия и т.н. След това се пристъпва към прогнозирането, като се преценява, при различни други възможности за регресорите, какво може да се очаква за съответната величина на регресанда, как тя би кореспондирала на различни хипотези за обясняващите променливи. В това отношение изникват два варианта: да се прогнозира математическото очакване на регресанда, т.е като средна стойност, или дори самата му величина. Първият вариант се обхваща от разглежданията в т. 2. чрез релацията $\theta = C^T \beta$, като в случая C^T представлява матрицата от бъдещите стойности за регресорите и θ е математическото очакване за регресанда като резултативна на тези регресори величина. Там бе посочено, че $\hat{\theta} = C^T \hat{\beta}$ представлява линейна неизместена оценка за тази неизвестна величина, при това с минимална дисперсия (BLUE). Ковариационната матрица на тази оценка се дава с (2.10).

Доверителните области за подобна прогноза се получават по (7.1) и биха показвали в какви граници може да се очаква търсената величина.

Вторият вариант представлява малко по-особена задача, тъй като трябва да се прогнозира не някаква неизвестна константа, а случайна величина, и така се излиза от рамките на обичайните линейни статистически модели (представляващи т.нар. обикновен метод на най-малките квадрати – OLS, Ordinary Method of Least Squares). Моделът за тази случайна величина по-скоро би изглеждал по следния начин:

$$u = C^T \beta + \eta, \tag{8.1}$$

където u и η са бъдещите стойности за регресанда и стохастичната компонента. Проблемът е това "бъдеще" да се изясни на основата на данни за "миналото" на изучваното явление, представено с адекватен линеен модел от вида на (2.2). Логиката на метода на най-малките квадрати може да помогне за канализирането на решението на подобна задача с оглед намирането на най-подходяща оценка $\hat{u}$ за величината на u.

В (8.1) математическото очакване на регресанда u би трябвало да е равно на $\theta = C^T\beta$, съдържащо бъдещите стойности за регресорите. Стохастичната компонента η се очаква да има нулево математическо очакване, дисперсия σ^2, съвпадаща с тази на своето "минало" ε в (2.2), и да бъде некорелирана с него. Последното означава, че основните тенденции в развитието би трябвало да се съдържат в обясняващата компонента на (2.2), доколкото този модел се предполага да бъде адекватен на действителността.

За търсената оценка $\hat{u}$ на бъдещата величината на регресанда е логично да се очаква условна неизместеност в следния смисъл:

$$E(\hat{u} \mid u) = u. \tag{8.2}$$

Още повече, в резултат на (8.2), подобна връзка би се получила и между безусловните математически очаквания $E(\hat{u}) = E(u) = C^T\beta$. Като отделни случаи обаче величините на $\hat{u}$ и u може да се разминават, да се окажат твърде далечни една от друга и големината на подобна дисперсия би се определяла от ковариационната матрица на разликата $\hat{u} - u$ – колкото по-минимална е тя, толкова по-точна и надеждна би била и прогнозата.

Линейните оценки на u и на нейното математическо очакване $C^T\beta$ може да се представят в най-обща форма като $\hat{u} = (A + C^T(X^TX)^{-1}X^T)y$ при всякаква матрица с подходящи размери A. Неизместеността на оценките означава обаче, че трябва все пак да е изпълнено условието $AX = 0$. От това се получава

$$E(\hat{u} - u)(\hat{u} - u)^T = \sigma^2(AA^T + C^T(X^TX)^{-1}C). \tag{8.3}$$

Последното би било минимум, само ако $A = 0$. Така се получава и ковариационната матрица за оценката на математическото очакване (2.10).

От (8.3) може да се получи и ковариационната матрица на прогнозата $\hat{u}$ за величината на (8.1)

$$\sigma^2(AA^T + C^T(X^TX)C + I). \tag{8.4}$$

Последното е минимум пак при $A = 0$, което означава, че прогнозата за бъдещата величина на регресанда съвпада с тази за математическото й очакване $\hat{u} = C^T\hat{\beta}$. Разликата е само в ковариационните им матрици. При прогнозиране на математическото очакване ковариационната матрица на отклоненията $\hat{u} - C^T\beta$ се получава равна на (2.10). Това означава, че върху точността на такива прогнози влияе само стохастиката в миналото ε, при съответна адекватност на модела (2.2). При прогнозиране на самата величина на

регресанда, отклоненията $\hat{u}-u$ придобиват ковариационната матрица (8.4) с допълнителен член $\sigma^2 I$ в сравнение с (2.10), отразяващ ефекта от η, т.е. стохастиката в бъдещето, и правещ подобна прогноза по-проблематична. Това отговаря на положението, че прогнозите за величини, представляващи отделни случаи, би трябвало да бъдат доста по-деликатна материя в сравнение с тези за средните стойности.

Аналогията между двата вида прогнози и фактът, че прогнозата за математическото очакване е BLUE, позволява за прогнозата на отделните стойности на регресанда да се възприеме, че е BLUP (Best Linear Unbiased Predictor), като по такъв начин се подчертава и съществената разлика между двете (Theil 1971).

9. РАЗШИРЕНИЕ НА ЛИНЕЙНИТЕ СТАТИСТИЧЕСКИ МОДЕЛИ И ОБОБЩЕН МЕТОД НА НАЙ-МАЛКИТЕ КВАДРАТИ

Линейните статистически модели притежават вътрешен заряд за развитие и в други посоки. Не представлява особена трудност методът на най-малките квадрати да се разпростре и върху модели с автокорелация и хетероскедастичност на стохастичната компонента, стига съответната ковариационната матрица да е известна с точност до един общ множител. Така вместо (1.5) и (2.3) може да се разгледа

$$E(\varepsilon\varepsilon^T)=\sigma^2 Q, \tag{9.1}$$

където матрицата Q е напълно определена и само величината на σ^2 остава неизвестна и подлежи на оценка заедно с регресионните параметри на модела (2.2). Идеята за такова разширение принадлежи на Ейткин (Aitkin 1935) и по метода на най-малките квадрати за съответните оценки се получава

$$\hat{\beta}=(X^T Q^{-1} X)^{-1} X^T Q^{-1} y \tag{9.2}$$

и

$$\hat{\sigma}^2=\tfrac{1}{N-K}(y-X\hat{\beta})^T Q^{-1}(y-X\hat{\beta}). \tag{9.3}$$

Както може да се забележи, минимизирането на най-малките квадрати в случая означава намирането на такива стойности за β, така че да се постигне минимум за величината на $(y-X\beta)^T Q^{-1}(y-X\beta)$. С това се постига и минималност на дисперсията на оценките (9.2) в пълно съответствие с теоремата на Гаус-Марков. В резултат на това за ковариационната матрица на (9.2) се получава

$$\sigma^2(X^T Q^{-1} X)^{-1}. \tag{9.4}$$

Съвсем праволинейно е разширението на модела и в посока оценяване на линейни комбинации на регресионните параметри във формата на $\theta=C^T\beta$ и резултатът е аналогичен на този от т. 2:

$$\hat{\theta} = C^T \hat{\beta} \tag{9.5}$$

с ковариационна матрица

$$\sigma^2 C^T (X^T Q^{-1} X)^{-1} C. \tag{9.6}$$

В този вариант методът на най-малките квадрати често се означава като GLS (Generalized Least Squares) за разлика от първоначалния OLS. Допълнителни хипотези от вида на многомерно нормално разпределение на стохастичната компонента също се разпространяват без трудности в модела GLS. В този случай оценките (9.2) и (9.3) са също независими случайни величини, което отваря пътя за приложение на методите от т. 7. Доверителни области например за линейни комбинации на регресионните параметри във формата на $\theta = C^T \beta$ се получават чрез неравенството

$$\frac{\frac{1}{M}(\theta - C^T \hat{\beta})^T (C^T (X^T Q^{-1} X)^{-1} C)(\theta - C^T \hat{\beta})}{\hat{\sigma}^2} \leq F_{M, N-K}(\alpha), \tag{9.7}$$

което е аналогично на (7.1), като в знаменателя оценката е от (9.3).

В редица приложения ковариационната матрица Q може да не е напълно позната и да се окаже обременена с неизвестни величини. По-нататък ще се разгледат някои по-типични примери, а засега заслужава да се подчертае, че (2.6) продължава да бъде неизместена оценка за регресионните параметри в модела GLS, но отстъпва на (9.2) по големината на дисперсията си. Този факт често се използва при наличие на неизвестни елементи на матрицата Q. Като първо приближение се предполага модел OLS и с помощта на (3.1) се оценява величината на стохастичната компонента, след което от получения резултат се прави опит да се оцени самата матрица Q, така че да отпаднат съответните неизвестни, и се преминава към модела GLS.

В горните формули фигурира обратната матрица Q^{-1} и това предизвиква въпроса за ранга на ковариационната матрица Q. В случай, че $rank(Q) < N$ обратната матрица не би съществувала, голяма част от предходните формули биха били невалидни и това би подронило устоите на метода GLS. Такъв дефект в ранга би означавал, че между отделните случаи на стохастичната компонента в модела (2.2) има не само корелация, а и линейни зависимости и съответно дублиране на информация. С други думи, информационната стойност на съответните статистически редове би била по-ниска от обикновеното и съответно би настъпила редукция в степените на свобода във формули като (9.3). Изход от подобна ситуация е трансформирането на използваните показатели така, че да се свие размерността на модела в пределите на информационния си ресурс и новата ковариационна матрица на стохастичната компонента да получи максимален ранг.

Подобни въпроси имат предимно технически характер и затова решението само ще бъде скицирано[1]. Вместо обратна матрица се въвежда т. нар. обобщено обратна (generalized inverse)[2], накратко – *g*-обратна. Всяка матрица, дори и да не е квадратна и независимо от

[1] За подробности вж Рао (1968).

[2] Често се нарича и псевдообратна, условно обратна или g-обратна. Ползата от такава конструкция е разкрита от Мур (Moore 1920) и развита по-нататък от Пенроуз (Penrose 1955) и други.

ранга си, има *g*-обратна. Ако се следва определението на Пенроуз (Penrose 1955) *g*-обратната е дори единствена. Така например *g*-обратната на матрица X с пълен ранг (като матрицата на плана в (2.2)) се получава $X^+ = (X^T X)^{-1} X^T$, което подсказва, че тази концепция би трябвало да е генетично свързана с метода на най-малките квадрати.

С помощта на *g*-обратните матрици лесно се преодолява проблемът с ранга на ковариационната матрица на стохастичната компонента в случай, че $n = rank(Q) < N$. Така например формула (9.3) би придобила следната форма:

$$\hat{\sigma}^2 = \tfrac{1}{n-K}(y - X\hat{\beta})^T Q^+ (y - X\hat{\beta}), \tag{9.8}$$

като съществото тук е не само появата на *g*-обратната на матрица Q^+, но и редукцията на степените на свобода от $N-K$ до $n-K$.

Хипотезите за многомерно нормално разпределение на стохастичната компонента също се засягат от дефекти в ранга на ковариационната матрица. При $n = rank(Q) < N$ това разпределение се концентрира в *n*-мерно подпространство, което се отразява както на процедурите по статистическа проверка на хипотези, така и на методите за получаване на доверителни области. Например формула (9.7) става

$$\frac{\frac{1}{M}(\theta - C^T\hat{\beta})^T (C^T (X^T Q^+ X)^{-1} C)(\theta - C^T\hat{\beta})}{\hat{\sigma}^2} \le F_{M,n-K}(\alpha), \tag{9.9}$$

като в знаменателя се намира (9.8) и степените на свобода на *F*-разпределението са съответно редуцирани.

Привличането на *g*-обратни матрици създава и други улеснения. При линейна зависимост между регресорите, т.е. екстремна мултиколинеарност, матрицата на плана X в (2.2) също би показала дефект в ранга, в резултат на което и $X^T X$ би имала непълен ранг, не би съществувала обратната матрица $(X^T X)^{-1}$ и (2.6) би било лишено от смисъл. Както се обърна внимание в т. 1, при такова положение може да се окаже, че регресионните параметри дори не са идентифицируеми.

Решението на такъв кръг проблеми в общи линии представлява следното (Рао 1968). Системата нормални уравнение (2.5) в този случай би имала безброй решения, поради което и оценката за параметрите би била неопределена. Някои линейни трансформации на регресионните параметри обаче като $C^T\beta$ би могло да се окажат разпознаваеми в рамките на модела и биха имали еднозначна оценка, представляваща BLUE. Необходимото и достатъчно условие за това е колоните на матрицата C^T, разглеждани като вектори, да се явяват линейни комбинации на колоните на X^T (или на $X^T X$, което е равносилно). Последното означава, че матрицата C^T може да се представи като $C^T = L^T X$ с помощта на подходящо линейно преобразуване (подходяща матрица) L. В случай, че матрицата на плана X има максимален ранг, такова преобразование може да се намери като $L^T = C^T (X^T X)^{-1} X^T$. При това положение всяка линейна трансформация на регресионните параметри, вкл. и самите тях, се

оказва идентифицируема, има оценка BLUE и така се получава основният сценарий OLS за линейните статистически модели.

10. ДИСПЕРСИОНЕН АНАЛИЗ

Дисперсионният анализ като концепция е предложен от Роналд Фишер през 1920 г., представлява един от най-важните методи на емпиричните проучвания и придава специфичния облик на статистиката през XX-тия век. Методът съществува в много варианти, като най-простият е еднофакторният дисперсионен анализ. Неговият сценарий се състои в следното. Основният въпрос е доколко даден резултат, представен с числов показател, се определя от някои признаци, условия, предпоставки, фактори и т.н. Например доходът на домакинствата може да се определя от броя лица в тях и особено от икономически активните и техния трудов статус, тяхното образование и занятие, към коя социална група би могло да се причисли домакинството, местоживеенето и редица други. Данните, подходящи за подобен анализ, би трябвало да представляват статистически ред за доходите, в който са обособени отделни групи (класове, категории) в зависимост от горните признаци. Универсалността на дисперсионния анализ се дължи главно на това, че за разлика от регресионния анализ, определящите признаци може да не са величини, а би могло да имат каквато и да е природа.

$$
\begin{array}{l}
\\
\textit{Първа група} \\
\\
\\
\\
\textit{Втора група} \\
\\
\\
\\
\ldots \\
\\
\\
\\
\textit{Последна група} \\
\\
\\
\\
\end{array}
\left[\begin{array}{c}
y_{11} \\ y_{12} \\ \ldots \\ y_{1N_1} \\ \hline
y_{21} \\ y_{22} \\ \ldots \\ y_{2N_2} \\ \hline
\ldots \\ \ldots \\ \ldots \\ \ldots \\ \hline
y_{K1} \\ y_{K2} \\ \ldots \\ y_{KN_K}
\end{array}\right]
=
\left[\begin{array}{ccccc}
1 & 1 & 0 & \ldots & 0 \\
1 & 1 & 0 & \ldots & 0 \\
\ldots & \ldots & \ldots & \ldots & \ldots \\
1 & 1 & 0 & \ldots & 0 \\ \hline
1 & 0 & 1 & \ldots & 0 \\
1 & 0 & 1 & \ldots & 0 \\
\ldots & \ldots & \ldots & \ldots & \ldots \\
1 & 0 & 1 & 0 & 0 \\ \hline
1 & \ldots & \ldots & \ldots & \ldots \\
1 & \ldots & \ldots & \ldots & \ldots \\
1 & \ldots & \ldots & \ldots & \ldots \\
1 & \ldots & \ldots & \ldots & \ldots \\ \hline
1 & 0 & 0 & \ldots & 1 \\
1 & 0 & 0 & \ldots & 1 \\
\ldots & \ldots & \ldots & \ldots & 1 \\
1 & 0 & 0 & \ldots & 1
\end{array}\right]
\left[\begin{array}{c}
\beta_0 \\ \beta_1 \\ \beta_2 \\ \ldots \\ \beta_K
\end{array}\right]
+
\left[\begin{array}{c}
\varepsilon_{11} \\ \varepsilon_{12} \\ \ldots \\ \varepsilon_{1N_1} \\ \hline
\varepsilon_{21} \\ \varepsilon_{22} \\ \ldots \\ \varepsilon_{2N_2} \\ \hline
\ldots \\ \ldots \\ \ldots \\ \ldots \\ \hline
\varepsilon_{K1} \\ \varepsilon_{K2} \\ \ldots \\ \varepsilon_{KN_K}
\end{array}\right]. \qquad (10.1)
$$

Моделът на разглежданата зависимост за така групирани данни би могло да се представи като (10.1), т.е. във формата на (2.2). Данните за резултативния показател би трябвало да са

разделени на K групи с размери съответно $N_1, N_2, \ldots N_K$, като $N_1 + N_2 + \ldots + N_K = N$ дава пълния им обем. Матрицата на плана X е подбрана целенасочено да се състои само от нули и единици в съответна последователност, така че за отделните случаи да се получи

$$y_{kn} = \beta_0 + \beta_k + \varepsilon_{kn}. \tag{10.2}$$

Последното означава, че математическото очакване на резултативната величина се определя от едно еднакво ниво β_0, общо за всичките случаи и групи, плюс специфично за отделните групи отклонение β_k (при $k = 1, 2, \ldots K$). Интригата в данните, доколкото такава има, е дали може да се възприеме, че

$$\beta_1 = \beta_2 = \ldots \beta_K, \tag{10.3}$$

т.е. че няма различие между групите в математическото очакване на резултативния показател, в репликата, каквато терминология се е наложила в дисперсионния анализ. Примерно, че размерът на доходите като средна величина не кореспондира с факторите и признаците, по които са групирани домакинствата, отклоненията при отделните случаи не може да се обяснят с тях, а имат само случайна природа и се генерират от стохастичната компонента ε.

За стохастичната компонента се предполага нулево математическо очакване и стандартното положение (2.3), както и некорелираност на отделните случаи, така че случайният ефект да има еднакво поведение за всички групи.

Матрицата на плана в (10.1) има непълен ранг равен на K (а не $K+1$, колкото са колоните в нея), така че за приложението на метода на най-малките квадрати се налага привличането на *g*-обратните матрици. Тъй като

$$X^T X = \begin{bmatrix} N & N_1 & N_2 & \ldots & N_K \\ N_1 & N_1 & 0 & \ldots & 0 \\ N_2 & 0 & N_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ N_K & 0 & 0 & \ldots & N_K \end{bmatrix}, \tag{10.4}$$

за *g*-обратната се получава

$$(X^T X)^+ = \begin{bmatrix} 0 & 0 & 0 & \ldots & 0 \\ 0 & N_1^{-1} & 0 & \ldots & 0 \\ 0 & 0 & N_2^{-1} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & N_K^{-1} \end{bmatrix}. \tag{10.5}$$

Така оценките на параметрите на модела стават

$$\hat{\beta} = \begin{bmatrix} 0 \\ \frac{1}{N_1}\sum_{n=1}^{N_1} y_{1n} \\ \frac{1}{N_2}\sum_{n=1}^{N_2} y_{2n} \\ \cdots \\ \frac{1}{N_n}\sum_{n=1}^{N_K} y_{Kn} \end{bmatrix}, \tag{10.6}$$

т.е. вътрешногруповите средни $\overline{y}_k, (k = 1, 2, \ldots K)$ се явяват оптималните стойности по смисъла на теоремата на Гаус-Марков.

Прави впечатление още, че общият параметър β_0 остава неидентифицируем с неопределена величина. Всъщност той се оказва скрит в останалите параметри като обща част, тъй като в модела не може да се различи съчетанието $\beta_0, \beta_1, \beta_2, \ldots \beta_K$ от $0, \beta_1 - \beta_0, \beta_2 - \beta_0, \ldots \beta_K - \beta_0$ или кое и да е друго отместване в тяхната величина. Това не е проблем за анализа тук, тъй като интересната хипотеза е (10.3), каквато и да е величината на отделните параметри.

Може да се направи извод, че първата колона в матрицата на плана (състояща се само от 1) в (10.1) би могла да отпадне, като по такъв начин се избегне привличането на *g*-обратни матрици. При еднофакторния дисперсионен анализ това е така и представлява известно улеснение, но при другите варианти на метода често се оказва, че отбягването от *g*-обратни матрици излишно усложнява и всъщност утежнява техниката и дори затъмнява логиката на статистическия извод.

Минимумът на сбора от квадратите на отклоненията (2.7), при този модел се оказва равен на

$$RSS = (y - X\hat{\beta})^T (y - X\hat{\beta}) = \sum_{k=1}^{K}\sum_{n=1}^{N_k} (y_{kn} - \overline{y}_k)^2 \tag{10.7}$$

със степени на свобода $N - K$ и е показателен за сумарния ефект на стохастичната компонента върху резултативния показател. Неговата величина не може да се свърже с класифициращите признаци, тъй като девиациите в (10.7) са в рамките на отделните групи, където единиците не се различават по тези признаци.

Хипотезата (10.3) може да се моделира във формата на $C^T\beta = \theta$ по различни начини, например

$$\begin{bmatrix} 1 & -1 & 0 & \dots & 0 \\ 0 & 1 & -1 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ -1 & 0 & 0 & \dots & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \dots \\ \beta_K \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \dots \\ 0 \end{bmatrix}. \tag{10.8}$$

Тук матрицата C^T също има непълен ранг, равен на $K-1$, и (10.8) всъщност налага само $K-1$ независими ограничения.

Формула (4.1) с използване на *g*-обратни матрици, където се налага, дава в условията на ограниченията (10.8) оценка, представляваща вектор с еднакви координати, равни на глобалната средна аритметична, обхващаща всичките N случая за резултативния показател. Тази величина може да се представи и като средна претеглена от вътрешногруповите средни

$$\overline{y} = \tfrac{1}{N} \sum_{k=1}^{K} \sum_{n=1}^{N_k} y_{kn} = \tfrac{1}{N} \sum_{k=1}^{K} N_k \overline{y}_k. \tag{10.9}$$

Минимумът в сбора от квадратите с отчитане на ограниченията, съгласно (4.3), се оказва равен на

$$RSS_C = (y - X\hat{\beta}_C)^T (y - X\hat{\beta}_C) = \sum_{k=1}^{K} \sum_{n=1}^{N_k} (y_{kn} - \overline{y})^2, \tag{10.10}$$

и отговаря на дисперсията в статистическия ред за резултативния показател без никакво групиране.

Увеличението, съответно на (4.4), във величината на (10.10) в сравнение с (10.7), поради наложените ограничения (10.8), отговарящи на хипотезата (10.3), има степени на свобода $K-1$. Така показателят (5.3) става

$$F_{K-1,\,N-K} = \frac{\frac{1}{K-1}(RSS_C - RSS)}{\frac{1}{N-K} RSS}. \tag{10.11}$$

и хипотезата (10.3) би трябвало да се отклони, ако величината на (10.11) надхвърля критичните стойности. Това би свидетелствало, че резултативният показател зависи от класификацията, а с това и от признаците, по които е направена групировката.

В каква степен резултативният показател зависи от признаците, залегнали в групирането на данните, се отразява и на коефициента на детерминация (3.9), който при еднофакторния дисперсионен анализ придобива известната форма

$$R^2 = \frac{RSS_C - RSS}{RSS_C}. \tag{10.12}$$

След евентуално установяване на зависимост между резултативния показател и класифициращите признаци обикновено интересът е насочен към диференциране на факторите за резултата, определяне на техния индивидуален принос с евентуално отсяване на несъществените такива. Би могло например при статистическото изучаване на доходите на домакинствата да се търсят главните фактори и да се направи опит поотделно да се претегли тяхното влияние. Подобна задача е характерна за многофакторните модели на дисперсионния анализ, аналогична е на съответната при регресионния анализ и решението й е белязано с подобна проблематика.

При многофакторните модели на дисперсионния анализ може да възникне проблемът по взаимодействието на факторите, аналогично на мултиколинеарността при множествения регресионния анализ, в резултат на което ефектът от отделните фактори да се окаже неопределим. При подобно положение не би било възможно да се диференцира и факторното влияние върху резултативния показател.

За разлика обаче от регресионния анализ при дисперсионния има значително повече възможности за гъвкаво проектиране и планиране на статистическите наблюдения, така че не само да се избегне проблемът по взаимодействието на факторите, но и да стане възможно отмерването на индивидуалното им влияние, а и въобще да се постигне максимална съдържателност и надеждност на статистическите резултати. Тази особеност произтича от непретенциозността на измерителните скали за факторните признаци, което открива широки възможности за многообразни факторни съчетания и комбинации, респ. подходящи матрици на плана, така че по-контрастно да изпъкват различните аспекти на влиянието им върху резултативния показател. За тези възможности се загатна с обсъждането на ранга на матрицата на плана в (10.1). Впрочем, комбинирането на различни изкуствени регресори, състоящи се само от нули и единици, е доста характерно при избора на един или друг модел на дисперсионния анализ още на фаза проектиране и планиране на практическите статистически проучвания.

Това подсказва също, че дисперсионният анализ би се срещал предимно при експерименталната статистика, репрезентативните проучвания, маркетинга, сондажите на общественото мнение, полевите опити, клиничните изследвания и въобще там, където е не само възможно, а често и неизбежно, да се възприема един или друг дизайн на статистическите наблюдения с оглед на диференцирано и детайлизирано изучаване на факторите за определен резултат. Като основоположник на този дял от статистиката се счита Роналд Фишер с неговия труд *The Design of Experiments* (1935). Добре познати у нас са и книгите на Кендалл (1976) и Шеффе (1980) със задълбочени изследвания на дисперсионния анализ. Планирането на статистическите наблюдения е разгледано задълбочено у Джонсън (1981).

11. РЕПРЕЗЕНТАТИВНИ СТАТИСТИЧЕСКИ ПРОУЧВАНИЯ

Особеното при репрезентативните проучвания е, че информацията се извлича от извадки, които често представляват много малка част от съответната генерална съвкупност. Въпреки това, напълно възможно става, започвайки от подобна привидно доста стеснена основа, да се постигне екстраполация на събраната информация и обобщаване в посока към генералната съвкупност. С други думи, да се прогнозира генералната съвкупност, като се изхожда от данните в извадката.

Често срещана задача при статистиката на репрезентативните проучвания е да се оцени някаква средна величина или обща сума в генералната съвкупност. Примерно, да се

определят средните доходи за цялото население или в някоя социална група или регион. Обикновено за генералните съвкупности предварително се знаят много неща, налична е информация от други източници и от предходни статистически наблюдения, което се използва при намирането на най-подходящ модел на извадката, респективно и обем, така че да се постигне максимално точна оценка на подобни средни величини. Понякога се провеждат дори и предварителни пробни наблюдения с цел да се уточнят всички механизми на проучването и да се настроят на вярната вълна.

Като начало ще се разгледа най-простият модел на лотарийния подбор, при който единиците от генералната съвкупност имат *a priori* еднакъв шанс да попаднат в извадката. Подборът се предполага без връщане, т.е. не е възможно повторно анкетиране на една и съща единица. Това обикновено се постига, като се ползват данни от регистрите за населението, адресните регистри, кадастри, градоустройствени планове, телефонните указатели и подобни списъци. От предварителната информация за генералната съвкупност засега ще се използват само данни за нейния размер, означен с N. За обема на извадката ще се използва n. Такъв сценарий подхожда най-вече на експресните социологически проучвания и телефонните интервюта.

Отделните случаи $y_i, i = 1, 2, \ldots n$ в извадката би могло да се представят така $y_i = \mu + \varepsilon_i$. Величината μ би представлявала средната за генералната съвкупност, която обикновено е неизвестна и нейното определяне би било една от задачите на репрезентативното проучване. Стохастичната компонента ε_i отразява спецификата на отделните случаи, изразяваща се в известно отклонение от генералната средна, т.е. математическото им очакване. Следователно, за математическото очакване на отделните случаи в извадката би било валидно $E(y_i) = \mu$, а заедно с това и $E(\varepsilon_i) = 0$.

На получените данни в извадката може да се придаде следната форма:

$$\begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \ldots \\ 1 \end{bmatrix} [\mu] + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \ldots \\ \varepsilon_n \end{bmatrix}, \tag{11.1}$$

съответстваща на (2.1) и (2.2). При тази конструкция обаче няма място за хипотезата (2.3), тъй като отделните случаи в извадките не би могло да бъдат независими. Причината е, че в хода на отбора и попълването на извадките постепенно намаляват свободните места в нея, с което се редуцира и шансът да се попадне под наблюдение за останалите единици от генералната съвкупност. Така че ковариационната матрица за стохастичната компонента се очертава да има следната форма:

$$E(\varepsilon\varepsilon^T) = \sigma^2 \begin{bmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \ldots & \ldots & \ldots & \ldots \\ \rho & \rho & \ldots & 1 \end{bmatrix}. \tag{11.2}$$

Еднаквите коефициенти на корелация ρ за всяка двойка случаи в извадката и въобще симетрията в (11.2) произтичат от лотарийния подбор, от факта на равнопоставеност на всичките случаи и от отсъствието на каквито и да е било предимства на едни единици пред други в генералната съвкупност в хода на попълването на извадката.

Обратната матрица на корелационната в (11.2) е

$$\frac{1}{(1-\rho)(1+(n-1)\rho)}\begin{bmatrix} 1+(n-2)\rho & -\rho & \ldots & -\rho \\ -\rho & 1+(n-2)\rho & \ldots & -\rho \\ \ldots & \ldots & \ldots & \ldots \\ -\rho & -\rho & \ldots & 1+(n-2)\rho \end{bmatrix}, \quad (11.3)$$

така че като BLUE оценка по метода на най-малките квадрати за генералната средна се получава съгласно (9.2) обичайната средна аритметична на случаите в извадката

$$\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i. \quad (11.4)$$

Този резултат може да се използва и за оценка на агрегати за генералната съвкупност. Например общият доход би се получил от оценката за средния доход на лице като $N\hat{\mu}$, съответно на размера на генералната съвкупност N.

За дисперсията на оценката (11.4) съгласно (9.3) се получава

$$E(\hat{\mu}-\mu)^2 = \frac{\sigma^2}{n}(1+(n-1)\rho). \quad (11.5)$$

На пръв поглед тази величина като че не зависи от размера на генералната съвкупност, а само от обема на извадката и коефициента на корелация. Неизвестната стойност на този коефициент може да се получи със съображения за симетрията на случаите в извадката и се оказва, че зависи само от обема на генералната съвкупност, но не и на извадката

$$\rho = -\frac{1}{N-1}. \quad (11.6)$$

Като се използва (11.6), за (11.5) се получава

$$E(\hat{\mu}-\mu)^2 = \frac{\sigma^2}{n}\cdot\frac{N-n}{N-1}. \quad (11.7)$$

Величината на втория множител в дясната част на (11.7) често се свързва с относителния обем на извадката (като относителен дял на попадналите в извадката от цялата изучавана съвкупност) и се представя като корекция, съобразена с размера на генералната съвкупност, като отразяване на мащаба на извадката в сравнение с генералната съвкупност.

Доверителни области за генералната средна μ, в случая интервали, се получават по (9.7), което често се използва за намиране на минималният обем на извадката, така че тези интервали да са достатъчно тесни, в рамките на предварително зададена точност или определен лимит за разходите по статистическото наблюдение. Тази проблематика е изчерпателно разгледана у Кокрен (1976).

Формула (11.7) може да се представи и така

$$E(\hat{\mu}-\mu)^2 = \frac{\sigma^2}{N-1}\left(\frac{N}{n}-1\right) = \frac{N}{N-1}\sigma^2\left(\frac{1}{n}-\frac{1}{N}\right), \tag{11.8}$$

от което се вижда, че размерът на генералната съвкупност по-слабо влияе върху величината на тази дисперсия (с което и върху точността и надеждността на (11.4) като оценка за генералната средна) и далече по-голямо значение имат големината на дисперсията σ^2 и относителният дял на извадката от генералната съвкупност. Колкото по-голяма е дисперсията в генералната съвкупност, толкова повече единици би трябвало като компенсация да се наблюдават, за да се получат достатъчно тесни и съдържателни доверителни интервали.

Като оценка за дисперсията на стохастичната компонента в (11.1) формула (9.3) придобива следния вид:

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{n}(y_i-\bar{y})^2. \tag{11.9}$$

Особеното в тази формула е, че в знаменателя се намира обемът на генералната съвкупност, а не на извадката, каквото би било, ако се отнася за дисперсията в рамките само на извадката. Така се получава, че с нарастване на обема на извадките величината на (11.9) се доближава до дисперсията в генералната съвкупност и едновременно с това (11.7) и (11.8) клонят към нула. Това отговаря на интуитивните представи, че с по-голям обем на извадките се постига повече точност и повече надеждност на получените резултати. Обратно, при малък обем на извадките тази величина е пак показателна за дисперсията в генералната съвкупност, но е обременена и със стохастиката в резултат на случайния подбор на единиците в извадката. Независимо от тези особености обаче и при всякакъв обем на извадките, формула (11.9) дава неизместена оценка за генералната дисперсия σ^2.

В практическите приложения и емпиричната социология обикновено са известни много предварителни факти за генералната съвкупност освен нейният размер. Доходите например доста се различават между градовете и селата и като среден размер, и като разнообразие – по дисперсията си, т.е. диференциацията на доходите. Различия има и между двата пола, отделните социални групи и прослойки, в териториален разрез и др. Подобна информация обикновено се набавя от предходни проучвания, различните административни и статистически регистри и др. Във връзка с това уместно е генералната съвкупност да се разглежда съставена от относително обособени и сравнително хомогенни части, слоеве или страти[1], каквато терминология се е наложила в статистиката. Отделните страти се разглеждат като самостоятелни съвкупности от N_k единици, със специфични средни μ_k и

[1] От лат. *stratum*, мн. ч. *strata*.

дисперсии σ_k^2, $(k = 1, 2, ... K)$, като се предполага, че различието между стратите е съществено и така представлява определено основание за тяхната идентификация. За да образуват стратите пълно покритие на генералната съвкупност, общият им обем трябва да е равен на $N = N_1 + N_2 + ... + N_K$. Такъв дизайн може да се разглежда като непосредствено разширение на лотарийния подбор чрез включване в модела на допълнителни подробности за генералната съвкупност.

За всяка от стратите може да се разглеждат модели като (11.1), като от всяка страта се извлича отделна извадка с обем n_k, съвсем самостоятелно и независимо от другите страти. Сборът на тези извадки от отделните страти има общ обем $n = n_1 + n_2 + ... + n_K$, дава т.нар. стратифицирана (или разслоена) извадка съответно на цялата генерална съвкупност и този начин на "сглобяване" от няколко части представлява същността на т. нар. стратифициран подбор. При такова положение средната величина в генералната съвкупност се явява линейна комбинация на средните в отделните страти

$$\mu = \sum_{k=1}^{K} \frac{N_k}{N} \mu_k \tag{11.10}$$

и съгласно (9.5) нейната BLUE оценка е равна на претеглената средна от извадките в отделните страти

$$\hat{\mu} = \sum_{k=1}^{K} \frac{N_k}{N} \hat{\mu}_k. \tag{11.11}$$

Дисперсията на тази оценка се дава с (9.6), което при този комплекс от модели придобива следната по-конкретна форма:

$$E(\hat{\mu} - \mu)^2 = \frac{1}{N^2} \sum_{k=1}^{K} \frac{N_k^2}{N_k - 1} \sigma_k^2 \left(\frac{N_k}{n_k} - 1 \right). \tag{11.12}$$

Последното разкрива, че минимумът постигнат по метода на най-малките квадрати (по смисъла на теоремата на Гаус-Марков), би могъл по-нататък да се намали още чрез подходящ план за обема на отделните извадки в стратите дори и при зададен общ лимит за n. Като първо приближение може да се разгледа ситуацията, при която е налице еднаквост на дисперсиите в отделните страти $\sigma_k^2 = \sigma^2$, така че общият множител σ^2 би могъл да излезе пред знака за сбора. С помощта на множителите на Лагранж може да се намери минимумът на (11.12) при условие общият обем на извадките от отделните страти да бъде равен на n. Решението на подобна задача показва, че такъв минимум може да се постигне, ако в относителният дял на отделните извадки (спрямо размера на съответната страта) се възпроизвежда общия относителен дял, т.е.

$$n_k = \frac{n}{N} N_k. \tag{11.13}$$

По този начин се постига по-ниска дисперсия на оценките за генералната средна в сравнение с обикновения лотариен подбор, т.е. само чрез преструктуриране, без да се увеличава общия обем на извадките. Така се получават т. нар. пропорционални извадки и често този метод се нарича пропорционален или квотен стратифициран подбор. Пропорционалността идва от това, че относителният обем на извадките спрямо размера на съответната страта е еднакъв и равен на този за обединението им (спрямо размера на генералната съвкупност), като последното всъщност задава квотата за отделните страти. При такъв метод на стратификация във връзка с моделирането на стратите фактически се използва повече предварителна информация за генералната съвкупност отколкото при обикновения лотариен подбор и така се осигурява по-висока представителност.

Съществено ограничение при практическото приложение на метода на пропорционалния подбор е хипотезата за еднаквост на дисперсиите в отделните страти, което често е трудно за верификация. Конкретни данни за тяхната величина обикновено не са известни, по-скоро може да се разчита на информация за тяхното съотношение между отделните страти. Например диференциацията в доходите сред градското население е обикновено два-три пъти по-висока отколкото в селата. Подобни данни или поне резонни хипотези може да се използват при (11.12), като абсолютният размер на дисперсията се изнесе пред знака за сбора и вътре останат само пропорциите между отделните страти. Минимумът на (11.12) при определен лимит за общия обем n на отделните извадки може пак да се получи с използване на множителите на Лагранж и резултатът е, че за обема на извадките в отделните страти трябва да е изпълнено

$$n_k = n\frac{\sigma_k\sqrt{N_k(N_k-1)}}{\sum_{j=1}^{K}\sigma_j\sqrt{N_j(N_j-1)}}. \tag{11.14}$$

Последната формула е известна и като

$$n_k = n\frac{N_k s_k}{\sum_{j=1}^{K}N_j s_j}, \tag{11.15}$$

където всъщност в (11.14) е заместено

$$\sigma_k^2 = \frac{N_k}{N_k-1}s_k^2. \tag{11.16}$$

Това представлява оптималният метод за подбор при стратифицирането и той притежава следното ярко предимство пред пропорционалния. Ако някои страти са по-хомогенни, с по-ниска дисперсия, и дори ако не са многобройни, би могло да послужат като резерв, откъдето да се предостави допълнителна квота за други с по-висока дисперсия и така, при предопределен и фиксиран обем на обединената извадка, да се получи по-голяма точност и надеждност на резултатите. Може още да се направи извод, че колкото по-хомогенни са стратите, толкова по-икономично би могло да бъде и проучването, въпреки че между

стратите би могло да има големи различия както в средните величини, така и при дисперсията, а и самите страти да са значителни по размер.

12. СТАТИСТИЧЕСКИ АНАЛИЗ НА РАЗВИТИЕ

Динамичните статистически редове са характерни с ред особености, като основната тенденция в развитието им често е изтъкана от определени зависимости, разтеглени във времето. Такива са лаговите ефекти, породени от това, че бъдещото развитие в някаква степен е обусловено от настоящото състояние, което от своя страна е резултат от сложилите се тенденции в близкото минало. Намесват се и различни външни фактори, формират се циклични компоненти като сезонни колебания, оставящи отпечатък върху развитието. Приложението на метода на най-малките квадрати и линейните статистически модели за анализ на развитието най-често се сблъсква с подобна проблематика при иконометричното моделиране, тъй като икономическата природа на изучаваните процеси обикновено подсказва, че хипотези от вида за хомоскедастичност и некорелираност на стохастичната компонента в (1.5) и (2.3) не са съвсем реалистични.

При прилагането на регресионни модели на развитието по метода OLS е препоръчително да се подложат на допълнителен анализ получените оценки (3.1) за стохастичната компонента в нейния хронологичен ред. За целта често се използва критерият на Дърбин-Уотсън или метода на фон Нойман (Theil 1971), с което може да се открие автокорелация в стохастичната компонента и това да даде основание да се потърси в посока към GLS по-адекватен вариант на модела. Възможности има и за установяване на хетероскедастичност в случайната компонента, което би наложило по-нататъшно усъвършенстване на използвания модел. Често се забелязва например, че нарастване във величината на обясняващите променливи тясно кореспондира с увеличаване на разсейването в резултативния показател. Както се привлече вниманието в т. 1, по-високи доходи на домакинствата в общи линии разширяват и разнообразяват възможностите за потребление и, освен че се повишава средното ниво, това води и до по-голяма вариация в съответните разходи. Подобен ефект се наблюдава при всяко увеличение на мащаба в икономическите процеси. При такова положение може вместо стандартния регресионен модел $y = \beta_0 + \beta_1 x + \varepsilon$ да се опита примерно

$$\frac{y_t}{x_t} = \beta_1 + \frac{\beta_0}{x_t} + \frac{\varepsilon_t}{x_t}$$

с надежда така да се елиминира ефекта от мащаба върху стохастичната компонента. Това би се постигнало в случая, ако разглежданият процес подсказва, че дисперсията й е пропорционална на големината на регресора на квадрат. Статистическата практика познава множество подобни похвати, с които се цели да се открият най-подходящите трансформации на използваните показатели, така че да се получи стохастична компонента, освободена от хетероскедастичност, и да се разчисти пътят за прилагане на метода на най-малките квадрати. Такъв подход обаче има определени ограничения. Ако самите регресорите са случайни величини, с подобни трансформации не винаги е постижимо преодоляването на хетероскедастичността, а и често се привнася и допълнителна автокорелация, което само усложнява проблема.

Проблемите около хетероскедастичността и автокорелацията може да се третират и съвместно. В редица приложения например може да се очаква, че стохастичната компонента по-скоро е свързана в авторегресионна зависимост от вида

$$\varepsilon_t = \rho\varepsilon_{t-1} + \eta_t, \tag{12.1}$$

с неизвестен по величина коефициент на автокорелация $|\rho| \leq 1$. В този модел се полага, че за първообраза на стохастичната компонента η_t са налице нулево математическо очакване $E(\eta_t) = 0$, хомоскедастичност $E(\eta_t^2) = \omega^2$ и некорелираност и по такъв начин всеки следващ случай за ε в съответния хронологичен статистически ред се оказва чрез (12.1) в определена степен обусловен от предходния. Може да се забележи, че в рамките на (12.1) от $E(\eta_t) = 0$ следва $E(\varepsilon_t) = 0$, каквато и да е величината на ρ.

Моделът (12.1) има формата на (2.2), но някак не се вписва в неговата атмосфера дори и като GLS. Причината за това е, че променливата ε се явява както регресор в дясната част на (12.1), така и с лаг единица и като регресанд в лявата част. Подобно размиване на границите между екзогенните и ендогенните променливи е доста типично за иконометричното моделиране, където трябва да се отразят взаимоотношенията между различните икономически подсистеми и персонажи, всеки от които следва своите собствени интереси и виждания, преследва определени цели със съответна мотивация, има специфично поведение, собствена позиция и съответна инерция в развитието си. Реакциите на една или друга промяна в икономическото пространство рядко се случват наведнаж и в пълен размер, а по-скоро представляват последователност от действия в една посока. Така резултатите от едно или друго решение се появяват след известно време прогресивно, с натрупване на части със съответни лагове, оказват влияние върху останалите участници на пазара и създават новите условия и предпоставки за следващата фаза в развитието. Подобна взаимна зависимост показва, че и връзката между съответните модели по правило би трябвало е двупосочна и в някаква степен разтеглена във времето с лаг в съответните резултати и ефекти. Така се получава обикновено в иконометрията – регресорът в един от моделите се оказва регресанд в съседния модел, а често при участието на показатели с лаг, подобно нещо възниква и в отделните модели по подбие на (12.1).

Моделът (12.1) обаче има сравнително проста структура и позволява да се идентифицират неговите основни елементи. Така например при $|\rho| < 1$ за ковариационната матрица на ε се получава

$$E(\varepsilon\varepsilon^T) = \frac{\omega^2}{1-\rho^2}\begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{N-1} \\ \rho & 1 & \rho & \dots & \rho^{N-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{N-3} \\ \dots & \dots & \dots & \dots & \dots \\ \rho^{N-1} & \rho^{N-2} & \rho^{N-3} & \dots & 1 \end{bmatrix}, \tag{12.2}$$

като обратната на матрицата в (12.2) е

$$\begin{bmatrix} 1 & -\rho & 0 & \dots & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & \dots & 0 & 0 \\ 0 & -\rho & 1+\rho^2 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1+\rho^2 & -\rho \\ 0 & 0 & 0 & \dots & -\rho & 1 \end{bmatrix}. \tag{12.3}$$

Ковариационната матрица (12.2) има формата на (9.1) с

$$\sigma^2 = \frac{\omega^2}{1-\rho^2}. \tag{12.4}$$

При такова положение модел на инерцията в развитието на един икономически показател y_t в зависимост от определен комплекс фактори за неговата динамика X_t би могъл да се разглежда във формата на (2.2), но като GLS със стохастична компонента от вида на (12.1). Както се обърна внимание в т. 9, намирането на оценка за регресионните параметри в такъв модел, а с това и целият статистически анализ, би се натъкнал на сериозно препятствие – неизвестната величина на авторегресионния параметър ρ. Една възможност за излизане от това положение е трансформиране на линейния модел (1.2) в аналогичен, но за прирастите

$$y_t - y_{t-1} = \sum_{k=1}^{K} \beta_k (x_{kt} - x_{kt-1}) + \varepsilon_t - \varepsilon_{t-1}, \ t = 2, 3, \dots N, \tag{12.5}$$

като за стохастичната компонента се предполага (12.1) с опростяващото условие $\rho = 1$. При такова построение методът OLS е напълно приложим (стохастичната компонента $\eta_t = \varepsilon_t - \varepsilon_{t-1}$ е освободена от хетероскедастичност и автокорелираност) и в сравнение с първоначалния модел се губи само една единица в степените на свобода ($N-1$ случая в производния статистически ред вместо N). Такъв подход е широко използван от Стоун (Stone 1954) при изучаване на потребителското търсене във Великобритания.

Друга възможност се крие в двустепенно приложение на метода на най-малките квадрати по следния начин. Най-напред се прилага OLS за модели от вида на (1.2), като временно се игнорират всички особености на ковариационната матрица като (12.2). От получените резултати, като се използва (3.1), се оценява коефициентът на автокорелация ρ по формулата

$$\hat{\rho} = \frac{\frac{1}{N-1}\sum_{t=2}^{N} \hat{\varepsilon}_t \hat{\varepsilon}_{t-1}}{\frac{1}{N-K}\sum_{t=1}^{N} \hat{\varepsilon}_t^2}. \tag{12.6}$$

Последното се замества в (12.2) и при тези условия се прилага метода GLS. Такъв подход е не по-малко проблематичен, тъй като се разчита на неизместеността на OLS оценките за регресионните параметри при първата степен. Подобно нещо обаче не е валидно за оценките на стохастичната компонента и още по-малко за големината на остатъчната дисперсия, в резултат на което (12.6) представлява изместена оценка за коефициента на автокорелация.

Отделен проблем е, че самите регресори може да се окажат случайни величини, натоварени със собствена стохастика. При такова положение методът на най-малките квадрати би могъл все пак да дава правилния отговор, което трябва обаче да се възприема със съответните му условности и ограничения. За да се побират непосредствено в схемата на OLS от (2.2) и (2.3) линейните модели със стохастични регресори би трябвало да имат следната форма:

$$E(y \mid X) = X\beta,\ E((y - X\beta)(y - X\beta)^T \mid X) = \sigma^2 I, \tag{12.7}$$

което осигурява независимост на ковариационната матрица (във втората част на (12.7)) от величината на регресорите.

Относителността на понятията за екзогенни и ендогенни променливи при иконометричните построения обикновено води до зависимости между стохастичните компоненти и регресорите и невалидност на (12.7), което поставя под съмнение линейността на представяните зависимости, както вече се обърна внимание в т.1. Подобни усложнения може да се демонстрират с един прост модел, често свързван с името на Кейнс (Thiel 1971). Разглежда се годишното потребление y_t в следната пропорционалност спрямо размера на доходите x_t (т. нар. потребителна функция):

$$y_t = \beta x_t + \varepsilon_t, \tag{12.8}$$

като стохастичната компонента ε_t изразява отклоненията от тази тенденция и може да се приеме, че $E(\varepsilon_t) = 0$. Зависимостта (12.8) има макроикономическо съдържание и към нея би могло да се добави, че общият доход x_t за всяка година отива за потребление y_t и спестяване (инвестиции) z_t

$$x_t = y_t + z_t. \tag{12.9}$$

Последното е по-скоро дефиниционно равенство, а не условен модел на релацията между няколко макроикономически величини.

Най-близката цел при подобно макроикономическо моделиране е с използване на серия годишни данни да се определи нормата на потребление – големината на параметъра β в (12.8), като за момент ще се допусне, че стохастичната компонента ε_t е освободена от хетероскедастичност и автокорелация или най-малкото, че ε_t и x_t са некорелирани. След заместване на потреблението от (12.9) в (12.8) се получава

$$z_t + \varepsilon_t = (1-\beta)x_t. \tag{12.10}$$

Последното показва, че ако ε_t и x_t са действително некорелирани, би следвало да има пълна отрицателна корелация между стохастичната компонента и спестяванията при всяко равнище на доходите, което е доста парадоксално и съвсем неприемливо от икономическа гледна точка. Така че е неизбежно да се допусне определена зависимост между доходите и стохастичната компонента в (12.8), което значително отдалечава тази пропорция от условието (12.7) и стандартните линейни статистически модели.

13. ЗА ГРАНИЦИТЕ НА МЕТОДА НА НАЙ-МАЛКИТЕ КВАДРАТИ

Периметърът на приложение на метода на най-малките квадрати е доста широк и изпъстрен както с предметното си многообразие, така и с различни методологически и методически подходи и варианти. Логиката и механизмът на метода при линейните статистически модели почиват на множество хипотетични положения, част от които последователно се подлагат на тест в светлината на използвания статистически материал, а други остават до края като резонни работни хипотези. И в зависимост от резултатите на поредния етап анализът продължава в една или друга посока, като постепенно хипотетичното получава емпирична верификация и свива дела си. Така се формира верига от статистически изводи и решения, натрупва се едно многоетажно построение и логично възниква въпросът доколко устойчива е подобна конструкция и доколко крайните резултати са чувствителни по отношение на евентуални, дори и доста малки отклонения на изходните предположения от действителността. Тази проблематика образува едно сравнително ново направление в статистиката – robust statistics.

Нормалното разпределение на стохастичната компонента например е доста съществено за валидността на изводите, базирани на дисперсионното отношение като (5.1) и в това отношение може да се очаква някакъв ефект от действието на централните гранични теореми, стига използваните статистически редове да имат достатъчно голям обем. Дали обаче при отделните приложения на метода на най-малките квадрати, с използване на конкретни данни, може да се разчита на това? При дисперсионния анализ например и особено при полевите опити, лабораторните изпитания и репрезентативните проучвания е възможно маневриране с един или друг дизайн или дори целенасочено проектиране на самото статистическо наблюдение, така че да се постигне или поне да се доближи моделът до перфектното положение. При регресионния анализ и иконометричното моделиране по макроикономически данни, предимно от административни източници, ситуацията е по-скоро обратната и трудно може да се разчита на някакви универсални статистически закономерности, които да привеждат стохастичната компонента като в (2.2), във формата на нормалното разпределение. В някои по-прости варианти на анализа при линейните статистически модели е показано, че отклонения от хипотезата за нормално разпределение на стохастичната компонента водят само до известна загуба в степените на свобода (Box,

Watson 1962) и, със съответната модификация, критерият и процедурата с дисперсионното отношение все пак остават валидни.

Различни ситуации възникват, ако самите регресори имат случайно поведение или са повлияни от външни случайни фактори, което е доста често срещано при приложенията на регресионния анализ. Причините за това може да се крият в природата на използваните показатели, но и да са породени от закръгления в числата, непрецизност при статистическите наблюдения или измерванията на показателите, групиране или някакво трансформиране и преформатиране на данните. В това отношение доста съществено е дали по такъв начин не се привнася в крайна сметка в модела, не се поражда някаква корелация между стохастичната компонента и регресорите, което, освен всичко друго, драстично би нарушило основното изискване за линейност. Тази проблематика е разгледана нашироко у Маленво (1975).

Независимост на стохастичната компонента от регресорите означава преди всичко, че ковариационната матрица (9.1), макар и с неизвестни величини, не е свързана с обясняващите променливи. Хипотетичното положение за некорелираност на отделните случаи в статистическия ред и условието за хомоскедастичност в (2.3) често е неприемливо в конкретни практически ситуации, каквито са разгледани в т. 11 и 12. Най-доброто решение би било, ако при подготовката на самото статистическо наблюдение детайлите се проектират така, че да може още в плана на проучването да се определи структурата на тази матрица с точност до един общ множител, както е при (11.2), чиято величина може по-нататък да се определи въз основа на събраната информация. При иконометричното моделиране има доста стеснени възможности за някакъв предварителен дизайн и често се прилага двустепенен метод на най-малките квадрати. Като първа степен се оценяват елементите на ковариационната матрица по метода OLS, след което полученото се използва за прилагане на GLS, както бе посочено в т. 12 във връзка с (12.6). При подобно повторно (и многократно) използване на статистическите данни също се губи нещо в степените на свобода, но такива процедури са все пак издържани в духа на законите на големите числа и централните гранични теореми. Благоприятно обстоятелство е, че по метода OLS се получават неизместени оценки за регресионните параметри, дори и когато ситуацията изисква приложение на GLS (вж т.9).

Друг кръг проблеми е свързан с недооценка на факторните признаци и обясняващите променливи в следния смисъл. Изборът на регресорите в (1.1) и (1.2) може да не е съвсем удачен, особено при първите итерации в посока към намирането на най-адекватния модел. Така например вместо по-пълен модел (5.7) може да се разглежда съкратен като (5.9) или вместо действителен (5.9) да се използва пренаситен с излишни регресори във формата на (5.7). При такова положение оценките за регресионните параметри би могло да се окажат изместени. Ако нещо куца в изходните предположения, изместеност може да се появи и при оценките за остатъчната дисперсия, както и при задачите по прогнозирането от т. 8 (Себер 1976).

Известни възможности за приложение на метода на най-малките квадрати съществуват и при нелинейни статистически модели, но в това направление все още има доста бели петна. Най-известният подход е свързан с опита да се линеаризират такива зависимости чрез прилагане на подходящи трансформации на използваните променливи, така че параметрите на модела и стохастичната компонента да "изплуват" в линейна форма. Друга възможност е директна атака за минимизиране на сбора от квадратите на остатъците с използване на съответните математически методи. Съществуват и своеобразни хибриди на тези подходи като метода на Маркуардт (Дрейпер, Смит 1987). Положението обаче не е съвсем задоволително, тъй като подобни процедури често не позволяват разкриването на статистическите свойства на получените резултати и липсва основание за продължаване на

анализа в посока към получаване на доверителни интервали и подлагане на тест на различни варианти на моделите. Трудностите около приложението на метода на най-малките квадрати при нелинейни статистически модели са разгледани у Маленво (1975).

При всички случаи е препоръчително след приложение на метода на най-малките квадрати, в какъвто и да било вариант, внимателно да се изследват оценките за стохастичната компонента (3.1), както това обикновено се прави при динамичните статистически редове, с помощта на методи в духа на критерия на Дърбин-Уотсън (т. 12). Съществуват и редица графични методи за анализ в това направление (Себер 1976), повечето от които са налични в специализирания статистически софтуер.

Литература

Джонсън, Дж. (1980). *Эконометрические методы*. Москва: "Статистика".

Джонсън, Н., Ф. Лион (1981) *Статистика и планирование эксперимента в технике и науке – методы планирования эксперимента*. Москва: "Мир".

Дрейпер, Н., Г. Смит (1987). *Прикладной регресионный анализ*. Москва: "Финансы и статистика.

Кендалл, М., Стьюарт, А. (1966). *Теория распределений*. Москва: "Наука".

Кендалл, М., Стьюарт, А. (1973). *Статистические выводы и связи*. Москва: "Наука".

Кендалл, М., Стьюарт, А. (1976). *Многомерный статистический анализ и временные ряды*. Москва: "Наука".

Кокрен, У. (1976). *Методы выборочного исследования.* Москва: "Стратистика".

Колмогоров, А. (1946) К обоснованию метода наименших квадратов. *Успехи мат. наук*, вып. I., Москва.

Линник, Ю. (1961) *Метод найменьших квадратов и основы обработки наблюдений*. Москва: Физматгиз.

Маленво, Э. (1975). *Статистические методы эконометрии.* Выпуск 1. Москва: "Статистика".

Рао, С. (1968). *Линейные статистические методы и их применения*. Москва: "Наука",

Себер Дж. (1976). *Линейный регресионный анализ*, Москва: "Мир".

Шеффе, Г. (1980). *Дисперсионный анализ.* Москва: "Наука".

Markoff, A. (1900). *Wahrscheinlichkeitsrechnung*, Leibzig: Tebner Verlag.

Aitkin, A. (1935). On least squares and linear combination of observations. *Proc. Roy. Soc. Edin.*, A55, 42-48.

Box, G., G. Watson (1962). Robustness to non-normality of regression tests. *Biometrika*, 49, 93-106.

Moore, E. (1920). On the Reciprocal of the General Algebraic Matrix. (Abstract)" *Bulletin of the American Mathematical Society* 26, 394-5.

Neyman, J., David, F. (1938). Extension of the Markoff theorem on least squares. *Stat. Res. Mem*. 2, 105-116.

Parzen, E. (1961). An approach to time series analysis. Ann. Math. Stat. 32, 951-989.

Penrose, R. (1955). A Generalized Inverse for Matrices. *Proceedings of the Cambridge Philosophical Society* 51, 406-13.

Prais, S., H. Houthakker (1955). *The Analysis of Family Budgets*. Cambridge: The University Press.

Rao, C. (1945). Generalization of Markoff's theorem and tests of linear hypotheses. *Sankhia* 7, 9-15.

Stone, R, (1954). *The Measurement of Consumer's Expenditure and Behavior in the United Kingdom, 1920-1938*. Cambridge Univ. Press.

Theil, H. (1971). *Principles of Econometrics*. New York: John Wiley & Sons.

ЛОГИКА И МЕХАНИЗЪМ НА МЕТОДА НА НАЙ-МАЛКИТЕ КВАДРАТИ

Резюме

Методът на най-малките квадрати и неговото приложение при линейните статистически модели заема централно място в статистическата теория и практика. В студията се разглежда методът на най-малките квадрати в различни направления на статистиката – от регресионния анализ до репрезентативните проучвания, иконометричното моделиране и прогнозиране. Акцентът е поставен върху общата методологична основа на приложенията в иначе отдалечени предметни и проблемни области, в т.ч. на дълбоката връзка между обикновения и обобщения метод на най-малките квадрати и други подходи и методи на статистиката. Набелязват се основните насоки на анализа и най-съществените моменти в съответните методически последователности.

LOGIC AND WORKINGS OF LEAST SQUARES METHOD

Summary

Linear statistical models and the method of least squares play a central role in statistical theory and practice. It is examined in this study the application of the least squares method in linear statistical models in diverge areas of statistics – from regression analysis to sample surveys, econometric modeling and forecasting. The emphasis is on the common methodological groundwork of the applications in otherwise diverge subject and problem areas, including the deep relation between the ordinary and the generalized least squares and other statistical approaches and methods. The main directions of analysis are mapped out and the most important procedural points are underlined.