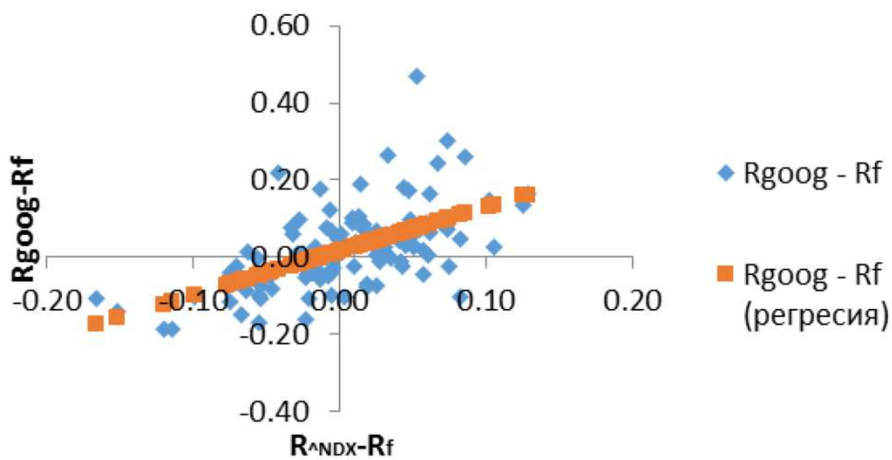
Станимир Кабаиванов

# Иконометрия за финансисти

## Част I




Евдемония продъкшън' 2014

ISBN 978-954-92924-6-6

**СТАНИМИР ИВАНОВ КАБАИВАНОВ**

# ИКОНОМЕТРИЯ
# за финансисти

**ИЗДАНИЕ ПЪРВО**

**ПЛОВДИВ 2014**

**д-р Станимир Иванов Кабаиванов – Автор**

**Рецензенти:**

**доц. д-р Анна Маринова Велева**

**гл. ас. д-р Маргарита Любенова Русева**

**Коректор:**

**гл. ас. Гергана Тодорова Иванова**

**Издател:**

**„Евдемония Продъкшън“ – гр. София**

**Съдържанието е свободно достъпно на [http://www.kabaivanov.net](http://www.kabaivanov.net)**

**ISBN 978-954-92924-6-6**

## Съдържание

# I. Въведение

Развитието на финансовите отношения прави познаването на основните икономически принципи необходимо не само за специалистите, но и за всички онези, които желаят да останат конкурентни и в крак с времето – независимо дали предлагат своя труд като част от нечия организация, или развиват собствени продукти и бизнес. Докосването до общите принципи и закони на икономиката и финансите е интригуващо, но ако не е подкрепено от солиден инструментариум, често може да остави усещането за непълнота на знанието и реална неприложимост. Целта на тази книга е да направи връзката между общите икономически и финансови модели и тяхното конкретно приложение с помощта на методите на математиката и статистиката. В изложението предполагаме, че читателят борави свободно с основни математически зависимости и операции, както и че е запознат с базисните понятия на вероятностните и статистическите изследвания. За разлика от мнозинството аналогичните текстове акцентът е поставен върху практическата приложимост, а математическите доказателства са изнесени в приложения или са представени с подходящи препратки. Този начин на оформяне прави текста достъпен и за онази част от аудиторията, която подхожда с известна доза недоверие и не особено позитивна нагласа към математическите модели и съпътстващите ги изчисления.

## 1. Какво е иконометрия?

Въпреки че използва добре познати и установени математически и статистически методи, иконометрията с право може да се смята за относително млада наука. Първоначалната употреба на термина идва от немското „Oekonometrie” и той е използван през 1910 от Pawel Ciompa за обозначаване на дейностите по събиране, систематизиране и проучване на информация за икономическите дейности [1]. Съвременното значение на термина обаче е фиксирано от Ragnar Frisch. За целите на настоящото изложение е достатъчно да дефинираме иконометрията като *съвместно приложение на математически и статистически методи за анализ на икономически и финансови данни с цел да се постигне по-добро обяснение и разбиране на изследваните процеси*. От дефиницията произтичат три важни за използването на иконометричните познания следствия:

- Прилагането на математически и статистически методи не е произволно, а съобразено с характеристиките на изследваните явления;
- Комбинирането на изследователски подходи от различните области трябва да доведе до общ синергетичен ефект, който да надвишава сумата от единичните „полезности” на отделните използвани аналитични методи;
- Получените резултати следва да позволяват апробирането на финансовите и икономическите модели, с което да потвърдят или отхвърлят тяхната приложимост и в крайна сметка да подпомогнат вземането на важни за практиката решения.

С тези три следствия е съобразено и цялото изложение на иконометрични подходи и методи в тази книга. Тя далеч не включва всички съществуващи модели и инструменти, които могат да бъдат полезни в практиката, но се старае да предложи един фокусиран подход върху основните от тях и тези, до чиято помощ се налага да прибягваме най-често.

## 2. Защо иконометрия за финансисти?

Една значителна част от книгите, учебниците и помагалата по иконометрия се фокусира върху изследването на макроикономически процеси и показатели. Подобен подход налага да се отдели специално внимание не само на макроикономическите модели, но и на проблеми с входните данни, които са специфични за изследването на процеси на макроравнище, но не са валидни в същата степен при анализ на финансовите пазари например. Анализирането на националните икономики е свързано с отчитането на особености, отнасящи се до характеристики на входните данни, като:

- проблеми при измерването и използването на различни по точност оценки и приближения;

- относително кратки времеви редове, чиято периодичност може да бъде от порядъка на една или няколко години;
- сблъскването с редки събития, чиято същност невинаги е добре изяснена и поради малкото налични данни не позволява да се оцени дали става въпрос за наистина прецедент, или за явление с неизяснена/непроучена периодичност.

Независимо дали тези особености са явно или неявно посочени, те дават отражение върху използваните математически и статистически подходи, като често се явяват в ролята на ограничение при избора на един или друг аналитичен инструмент. Благодарение на изобилието от данни, с което се съпътстват процесите на финансовите пазари, част от ограниченията на малките входни съвкупности от данни могат да бъдат преодолени. Това дава шанс да използваме по-разнообразни аналитични инструменти и да постигнем резултати с по-голяма точност. Разбира се, винаги съществува възможността да попаднем на случай с малък брой наблюдения/входни данни, поради което в изложението е обърнато специално внимание на допустимите в тази ситуация средства за анализ.

## 3. Трябва ли ни лист хартия?

Работата с голямо количество данни предполага или много свободно време, или оптималното използване на съществуващите ресурси. И тъй като времето е невъзобновим ресурс, примерите, използвани за илюстриране на представените концепции, са разделени на две основни групи:

- прости и лесни за пресмятане примери, чиято цел е да запознае читателя с конкретното приложение на даден модел, математическа зависимост или концепция. Проследяването на тези пресмятания е тривиално, но важно за бърза проверка дали въпросната концепция е възприета коректно;
- примери, свързани с анализ на по-големи съвкупности от данни, за които са представени най-често само крайните резултати и пътят за тяхното получаване с помощта на Excel.

Изборът на Excel като основно компютърно приложение е обоснован както от широката му употреба, така и от факта, че примерите могат да бъдат адаптирани практически без усилие и към други софтуерни продукти от същата група (напр. LibreOffice Calc[1] или Gnumeric[2]).

[1] http://www.libreoffice.org/features/calc/
[2] https://projects.gnome.org/gnumeric/

## II. Обработка на данни. Общи стъпки при иконометричните изследвания

Данните — в смисъл на неструктурираните факти и параметри, характеризиращи обекта на изследване, са изходната точка за прилагането на всеки един иконометричен метод и аналитично средство. За да можем да се възползваме максимално от наличните данни, е необходимо да ги класифицираме и по този начин да извлечем общите им свойства и особености.

### 1. Класификация на данните

В зависимост от формата на представяне на данните те могат да бъдат символни, числови или комбинирани. За целите на икономическия анализ интерес представляват не само тяхната форма и обем, а преди всичко описателната им способност и съдържащият се в тях допълнителен „принос", с който да се подобри разбирането ни за анализирания обект или процес. В този контекст можем да класифицираме данните според целите, за които ги използваме, а именно:

- статични данни — представящи стойностите на един признак спрямо N (N>1) дадености в една генерална съвкупност, в един времеви контур.

В този случай използването на събраните данни цели да анализира определени вътрешни зависимости или характеристики, вместо да проследи как се променя обектът на изследване с времето. Поради факта, че в този случай разполагаме с данни за интересуващия ни признак в един и същи контур, нямаме основание да говорим за динамика.

- хронологични/времеви редове — представящи стойностите на един признак спрямо една даденост в N (N>1) времеви контури.

Основните характеристики, които определят времевите редове, са видът на времевия контур и методологията за изчисляване на данните. С тези характеристики можем да посочим следните видове времеви редове:

    - според времевия контур те биват моментни редове (например курсът на еврото спрямо щатския долар в 12.00 ч. всеки ден от месец януари) и периодни редове (процент на заетите лица за всеки месец от 2013 година).

- според методологията за изчисляване на данните в периодните редове се открояват две важни характеристики:
  - времева кохерентност – т.е. когато всички времеви контури са периоди с еднаква продължителност и данните са получени с помощта на една и съща методология. Пример за такъв периодичен времеви ред е поредица от данни за равнището на безработица в България, отнасящи се за отделни месеци в рамките на една година.
  - методологична кохерентност – т.е. когато не всички данни в периодичния ред са изчислени по една и съща методология. Например промяна в структурата или методологията на изчисляване на един борсов индекс може да предизвика рязка промяна в размерността на неговите стойности, която обаче не е продиктувана от изменения в икономическата обстановка.

- данни, представящи динамичното развитие в параметрите на множество наблюдавани обекти с течение на времето.

Можем да разглеждаме този тип данни като комбинация от горните два вида. При това освен изследване на вътрешни зависимости и връзки може да се провежда и анализ на тенденциите за промяна в анализирания обект или процес. Трябва да отбележим, че без допълнителна информация за типа на проучваните икономически и финансови явления и процеси не може да се прави предварително подреждане или групиране на данните. Единственото общовалидно групиране и подредба могат да бъдат съобразените с времевия контур, за който се отнасят данните. Този факт още веднъж подчертава, че освен доброто познаване на статистическите и икономическите методи е необходимо да бъде налице информация и за особеностите на обекта на проучване.

Според възможните стойности, които имат изследваните данни, можем да ги разделим на две групи:

- Дискретни данни, при които характеристиките на наблюдаваните процеси предполагат ограничен набор от различни стойности. Не е задължително тези стойности да бъдат само цели числа. Пример за дискретни данни са цените на борсово търгувани ценни книжа, при които има определена стъпка на промяна при обявяване на котировките. Пример за целочислени дискретни данни са броят туристи посетили България през последната година, броят на продадени автомобили за последното тримесечие.
- Непрекъснати данни, при които няма ограничения в множеството от допустими стойности. На практика дори и при непрекъснати данни се налага да работим с определена степен на точност (т.нар. закръгляване след втория, третия и т.н. знак след десетичната точка), което предполага, че неявно се въвеждат ограничения в допустимите стойности. В този случай изборът на степента на точност е решение на

този, който провежда изследването, а не е присъща черта на анализираните икономически явления.

В зависимост от възможностите за сравнение и мащабиране можем да обособим три групи от данни:

- числови данни, представящи стойностите на числов признак с подходяща мерна единица. При тях можем да определим ясна и недвусмислена разлика (още обозначавана като дистанция) между отделните единици в съвкупността. Така например, ако един автомобил струва 50 000 лева, то той е точно два пъти по-скъп от автомобил с цена 25 000 лева.

За целите на следващото изложение е важен фактът, че при този тип всяка една стойност на данните има точен смисъл, което позволява да се съотнася и сравнява с останалите данни.

- ординалните данни представят стойностите на нечислов признак по начин, който позволява да ги подредим. Подреждането ни позволява да сравняваме отделните елементи, но не ни дава точна информация за степента, в която те са различни. Например можем да кажем че автомобилът с цена 50 000 лева е „по-добър" от този с цена 25 000 лв. Но не можем да твърдим, че той е „два пъти по-добър" използвайки само цената като показател. Аналогично не можем да твърдим, че победителя в конкурс за красота е „два пъти по-красив" от заелия второто място и „три пъти по-красив" от третия в класирането.

Ординалните данни, както и числовите, са свързани с фундаменталния принос на Георг Кантор към теорията на множествата. И също както при числовите данни, за целите на разбирането на използваните тук методи е важно да имаме предвид техните основни свойства и факта, че анализирането им изисква специфични методи и подходи.

- номиналните данни представят стойностите на нечислов признак по начин, който не позволява да ги подредим. Те представят само различие между отделните единици от съвкупността, без да дават никаква информация по какъв начин и в каква степен се проявяват тези различия.

Най-честият сблъсък с номинални данни е в случаите, когато имаме кодиране на символна информация или групи посредством цифри. Например кодирането на икономическите специалности в Пловдивския университет с 34, 38, 45 позволява да се добави информация за специалността във факултетния номер.

В зависимост от типа на данните е необходимо да бъдат подбрани и методите за тяхната обработка и анализ. Поради съществените различия в информацията, която носят отделните групи данни, и въпросните методи са специфични и водят до резултати с различна информационна стойност. Използването на модели и средства за изследване, които не са подходящи за типа на данните, може да доведе до получаването на съмнителни и неточни резултати.

## 2. Основни типове финансови данни

Голяма част от практическите задачи, които изискват приложението на иконометрични методи, са свързани с анализ на възвръщаемост и риск. Ако приемем, че рискът може да се дефинира посредством промените и волатилността на очакваната възвръщаемост, е необходимо да отделим внимание на това как се представя и изчислява тя.

Нека разполагаме с набор от входни данни за цената на една ценна книга и единствената доходност, която можем да получим, е свързана с промени в тази цена. Тогава анализът може да бъде направен директно с помощта на цените:

$$P_1, P_2, P_2, \dots, P_n \qquad \text{( II—1)}$$

или посредством изчисляване на тяхната разлика и генерираната по този начин възвръщаемост. Независимо че вторият метод изисква допълнителна трансформация, той е за предпочитане поради следните причини:

- използването на възвръщаемостта, изчислена като процент, ни помага да работим с безразмерни величини и ни освобождава от необходимостта да се грижим за ефекти, свързани с мащаба (абсолютната стойност) на цените;
- участниците на финансовите пазари се интересуват преди всичко от получената доходност, респ. възвръщаемост, вместо от абсолютната стойност на цените;
- анализирането на възвръщаемост вместо на цени позволява по-лесно да се установят и елиминират случаите на „фалшива зависимост”.

Последният довод е много важен за получаването на коректни и точни резултати, тъй като при изследване на икономическите и финансовите процеси се налага да работим под въздействието на няколко фактора (за разлика от други науки, където е възможно провеждането на контролиран експеримент с промяната само на определени параметри). „Фалшива зависимост” имаме в случаите, когато резултатите от корелационния анализ показват силна връзка между два признака (две променливи величини), но тази връзка няма логическа обосновка и съответно няма смисъл. Установяването на такава връзка за два признака X и Y може да означава едно от следните три неща:

- X може да бъде причина за Y;
- Y може да бъде причина за X;

- X и Y може да бъдат причинени от трети признак Z, без да имат пряка връзка помежду си.

В третия от посочените случаи високата степен на корелация може да ни подведе да направим погрешен извод и да приемем наличието на зависимост между X и Y, без да има такава. Ако се върнем на примера с цените, нека да си представим за момент, че имаме акция на фирма X, чиято цена за последните 10 години е нараствала стабилно. Ако разглеждаме този период и изчислим корелацията между цената на акцията и друга нарастваща променлива за същия период (например вашата собствена възраст), ще получим, че двете променливи са силно корелирани. Макар в примера да е очевидно че връзка между тях няма, при по-сложни финансови проблеми подобна ситуация може да доведе до вземането на грешни решения и до правенето на неточни изводи. Използването на възвръщаемост вместо цени позволява да намалим вероятността от изпадане в подобна ситуация (но не я елиминира напълно).

Пресмятането на възвръщаемостта, посредством директно съотнасяне цената на един актив в два съседни периода (или две съседни наблюдения) може да бъде чрез проста възвръщаемост, представена с помощта на темпа на прираст:

$$R = \frac{P_{t+1} - P_t}{P_t}, \text{ където } P_t \text{ е цената за периода } t. \qquad ( \text{II—2})$$

Предимствата на простата възвръщаемост се крият в нейното интуитивно възприятие и лекотата, с която тя може да бъде изчислена. При решаването на по-сложни задачи тази привидна простота може да доведе до усложнения, когато получените възвръщаемости се използват като входни данни на крайните икономически модели и съответно участват при дефинираните комплексни зависимости. Много често в практиката се използва и методът на определяне на компондираната (често наричана и логаритмична) възвръщаемост:

$$Rc = \ln \frac{P_{t+1}}{P_t} \ \textit{(x100\%)} \qquad ( \text{II—3})$$

Кога да предпочетем простата и кога — компондираната възвръщаемост, е въпрос на избор при решаването на конкретен проблем, но когато правим този избор, трябва да имаме предвид математическите и финансовите свойства на двата подхода, посочени в Таблица 1.

Изброените в Таблица 1 особености нямат за цел да дадат дефинитивен и „винаги правилен“ отговор на въпроса кой метод за изчисляване на възвръщаемост трябва да се ползва. И двата метода имат своите предимства и недостатъци и могат да спомогнат за по-бързото решаване на конкретни задачи. Необходимо е да познаваме силните и слабите страни и на двата метода, за да не попадаме в заблуда или да получим некоректни резултати.

Освен характеристиките, произтичащи от избраната математическа форма за представяне на възвръщаемостта, остава винаги актуален въпросът, доколко използваните величини реално отразяват реализираните на практика възвръщаемости. В горните примери използваме навсякъде единствено разликата в цената като източник на доходност. На практика освен капиталова печалба можем да имаме редица други парични потоци, свързани с изследваните обекти (например изплащани дивиденти), които трябва да бъдат отчетени, за да бъде нашият анализ пълен. Поради тази причина във всички примерни ситуации, разгледани по-нататък, ще споменаваме изрично, ако определени аспекти от изследваните процеси са пропуснати с цел опростяване на изложението.

*Таблица 1. Сравнение между проста и компондирана възвръщаемост*

| $R = \frac{P_{t+1}-P_t}{P_t}$ *(x100%)* | $Rc = \ln\frac{P_{t+1}}{P_t}$ *(x100%)* |
|---|---|
| Лесна за изчисляване и обяснение, включително и на неспециалисти във финансите. | По-трудно бихме могли да направим всички изчисления на ръка и се изисква познаване на концепцията за олихвяване и времева стойност на парите. |
| Не може да се направи директно сравнение между алтернативи, за които времевите периоди на изчисляване на възвръщаемостта не съвпадат (защото допълнително ще трябва да се отчете разликата във времето). | Получените възвръщаемости могат да се разглеждат като резултат от процес на непрекъснато олихвяване и по този начин получаваме независимост от разминавания във времевите периоди. |
| Отсъства свойството на адитивност поради това, че отделните възвръщаемости са получени посредством съотнасяне към различна база. $$R_1 = \frac{P_2 - P_1}{P_1}$$ $$R_2 = \frac{P_3 - P_2}{P_2}$$ ………… $$R_{n-1} = \frac{P_n - P_{n-1}}{P_{n-1}}.$$ | Компондираните възвръщаемости са адитивни по отношение на времето, което следва от математическите свойства на логаритмуването: $$R_1 = \ln\frac{P_2}{P_1} = \ln P_2 - \ln P_1$$ $$R_2 = \ln\frac{P_3}{P_2} = \ln P_3 - \ln P_2$$ ………… $$R_{n-1} = \ln\frac{P_n}{P_{n-1}} = \ln P_n - \ln P_{n-1}$$ Сумата от тези възвръщаемости ще бъде: $$\sum_{i=1}^{n-1} R_i = \ln P_{n-1} - \ln P_1 = \ln\frac{P_{n-1}}{P_1}.$$ |
| Простата възвръщаемост е удобна при изчисляване, възвръщаемостта на портфейл от активи, като претеглена средна от възвръщаемостта на $n$ компоненти на портфейла, съответно с дял $w_i$ и възвръщаемост $R_i$: $$R_p = \sum_{i=1}^{n-1} w_i R_i$$ | От свойствата на логаритмичната функция е известно, че: $$\ln(R_1 + R_2 + \cdots + R_{n-1}) \neq \ln R_1 + \ln R_2 +..+ \ln R_{n-1}$$ Следователно не можем да се възползваме от претеглените средни при определяне на възвръщаемостта на портфейл, в случай че възвръщаемостта на отделните активи е изчислена като компондирана такава. |

При анализ на показателите за риск има значително по-голямо разнообразие от показатели, форми на тяхното представяне и характеристики, които следва да се имат предвид. Поради тази причина ще посочим само основните четири свойства, на които трябва да отговаря един добър измерител на риска, както те са дефинирани от Wilmott[3] [2]:

*Таблица 2. Общи свойства на измерителите на риска*

| Свойство | Обяснение |
|---|---|
| Субадитивност | Това свойство е проявление на познатия принцип за диверсификация. То означава, че ако комбинираме две инвестиционни алтернативи, техният риск не може да бъде по-голям (т.е. по-лош) от сумата на техните рискови показатели поотделно. |
| Инвариантност | Това свойство означава, че ако увеличим *n* пъти размера на една инвестиционна алтернатива, то и показателят измерващ риска, следва да нарасне *n* пъти. |
| Положителна хомогенност | Положителната хомогенност означава, че добавянето на пари (или безрискови активи) към една инвестиционна алтернатива (напр. един инвестиционен портфейл) следва да намали риска. |
| Монотонност | Свойството монотонност означава, че ако една инвестиционна алтернатива има по-добро представяне от друга при почти всички възможни сценарии, то тогава рискът на първата алтернатива измерен от показателя, трябва да бъде по-нисък от този на втората алтернатива. |

В зависимост от това, как отделните показатели за риск въплъщават тези основни свойства, можем да анализираме техните характеристики и ограничения. Много често при демонстрациите на определени модели или концепции се използва само средното квадратично отклонение като измерител на риска. Там, където е възможно, в изложението са дадени препратки и информация и към други възможни инструменти и е дискутирана тяхната ефективност, предимствата и недостатъците им.

## 3. Общи стъпки при иконометричните изследвания

Подходите към реализирането на финансови и икономически анализи са толкова многобройни, колкото и различните проблеми, с които се сблъсква съвременната наука (а често могат да бъдат и по-многобройни). В този контекст посочените тук стъпки нямат за цел да ограничат читателите, нито да им наложат стил на работа. Спазването на описаните

[3] Идеята за кохерентните измерители на риска е развита от Artzner, Delbaen, Eber и Heath [21].

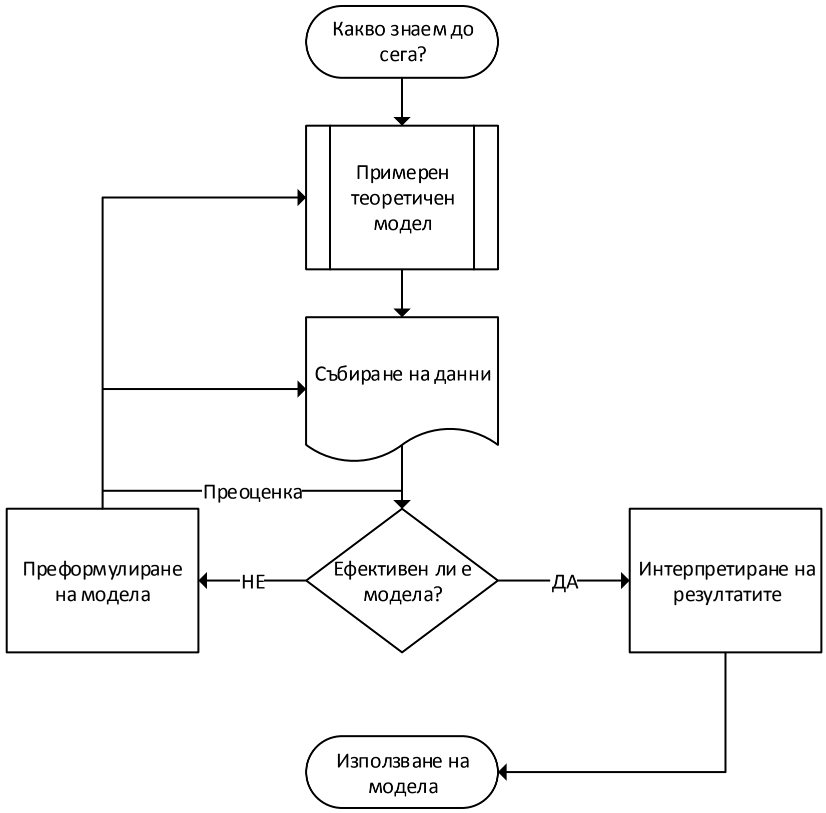
правила и последователност обаче могат да спестят много главоболия и да допринесат за по-лесното разбиране на модели и концепции, които изглеждат сложни и необятни на пръв поглед.

Първата стъпка, както е показана на Фигура 1, е свързана с точното идентифициране на наличното до момента познание за изследвания проблем. Въпреки че изглежда естествено именно това да е началният етап от всяко едно изследване, е важно да откроим два момента:

- Идентифицирането на наличното познание е свързано с точното дефиниране и на изследвания проблем. Една съществена част от трудностите по-нататък произтичат от неточно или непълно определяне на основните ни задачи и цели.
- В рамките на наличното познание се включват и други изследвания по същата или близка тематика. Колкото и привлекателно да звучи самата дума „изследване“, в много малка част от случаите става дума за анализ на напълно непознат обект. Проучването на това, какво другите са направили досега за решаването на близки по тип задачи, дава отговор на два важни въпроса:
    - какви методи можем да използваме и за кои аналитични инструменти е доказано, че не вършат работа за подобен тип проблеми – това може да спести не само време, но и да намали вероятността от грешка още при избора на аналитични средства;
    - в какви граници се движат получените от други икономисти резултати и ако нашите собствени се различават значително от тях, на какво се дължи това.

Втората стъпка е свързана с определянето на примерен теоретичен модел. Както отбелязва Brooks [3], много е вероятно този модел да не е перфектен и да не обхваща всяка една характеристика на анализирания обект. Добрият първоначален модел позволява с помощта на няколко последователни промени да се постигне достатъчно точно и адекватно теоретично представяне на изследваните процеси и зависимости.

Коректното регистриране на данни е третата стъпка и е от изключително значение за получаването на качествени крайни резултати. В зависимост от това, дали става въпрос за първични данни (например събирани посредством обработка на попълнени анкети), или за вторични данни (които вече са събрани и евентуално обработени по определен начин от друг), подходите използвани на този етап, са различни. И в двата варианта е необходимо да се подходи отговорно и честно към осигуряване на необходимата информация, без да се стремим към такива времеви отрязъци или източници, които са „удобни“ за доказването на определена теза.




*Фигура 1. Процес и стъпки при иконометричното изследване*

Оценката за ефективност на избрания модел е четвъртата стъпка в изследването и може да бъде декомпозирана на две взаимносвързани дейности:

- избор на критерий за оценка – критерият за оценка зависи от основните характеристики на подбрания модел (например неговата размерност, линейност / нелинейност) и анализираните процеси;
- оценка за статистическата коректност на избрания модел – целта на тази оценка е да отговори дали моделът съответства на заложените предположения и дали описва точно характеристиките на наличните данни;
- оценка за обясняващата сила и съвместимостта на модела със съществуващите икономически и финансови теории.

Окончателната приложимост на изградения модел и резултатите от него е крайният ефект, към който се стреми цялото иконометрично изследване. Особеностите на този етап са свързани с интегрирането на получените резултати в цялостната финансова/инвестиционна стратегия и стремежа да генерираме не просто преработени данни, а максимален принос към познанията ни за анализираните икономически и финансови процеси.

## 4. Общи означения и използвана нотация

В изложението използваме концепции и понятия, които не могат да бъдат изразени пълно без подходящи формули и математически зависимости. Винаги когато е възможно ще използваме обозначенията и нотацията описани в Таблица 3.

*Таблица 3. Общи означения и нотация*

| Показател | Обяснение |
|---|---|
| $x_i, y_i$ | Изразява $i$-тата стойност на променливите $x$ и $y$. Тази нотация ще се използва и в случаите, когато говорим абстрактно за стойностите на променливата, без да имаме предвид анализираните данни. |
| $\bar{x}, \bar{y}$ | Изразява средната стойност (средна аритметична, ако не е изрично посочено друго) на $x$ и $y$. |
| $\hat{x}$ (или $\hat{y}$) | Тази нотация се използва, за да отбележим, че става дума за оценка, получена от теоретичен модел за променливата $x$ (или $y$). Например $\hat{y}_i$ означава $i$-тата изчислена стойност за $y$. Примерна употреба на тази нотация има във Фигура 3 и Фигура 4. |

## 5. Задачи и въпроси за размисъл

1. Отговаря ли средното квадратично отклонение на всички изисквания за кохерентен измерител на риска?
2. Как бихте определили възвръщаемостта на инвестиционен портфейл, ако възвръщаемостта на отделните компоненти в него е зададена като компондирана (с помощта на логаритъм)?
3. Как постепенното преформулиране на икономическия модел спомага за повишаване на точността му?

# III. Линейни регресионни модели. Същност, валидност и допускания. Методи за оценка на параметрите и проверка за валидност на линейните регресионни модели

Историята на регресионния анализ е относително кратка, на фона на другите математически методи, които използваме във финансите. Споменатият факт в никакъв случай не влияе на значението, което този инструментариум има за развитието и практическата приложимост на икономическите изследвания. Можем да кажем, че регресионният анализ е не просто в основата на познанията, които са ни необходими за решаването на широк кръг от финансови проблеми, а че той е и най-широко използваният метод в практиката.

Регресионният анализ изучава и оценява връзката между две или повече променливи величини. Характерното за него е, че той се концентрира както върху изследване на наличието на причинно-следствена зависимост между две и повече променливи величини, така и върху търсенето на функционална връзка между тях.

## 1. Регресия и корелация

Инструментите на регресионният анализ не са единствените аналитични средства, с които можем да изследваме наличието на зависимост между различни променливи величини. Много често регресионният анализ се представя съвместно с един друг показател – на корелация, като при това рядко се посочват разликите между тях.

Корелацията е мярка за зависимост между две случайни величини, която не предполага и не изисква наличието на причинно-следствена връзка между тях. Т.е. корелацията асоциира измененията в разглежданите променливи, без да дава допълнителна информация дали тези изменения имат връзка помежду си. Нека разгледаме една примерна ситуация, при която летните температури са изключително високи. Това ще доведе — от една страна до повишена консумация на газирани напитка, а от друга — до увеличаване на броя на слънчевите удари. Ако съпоставим само крайните величини, ще видим, че има положителна зависимост между консумираните газирани напитки и слънчевите удари, но това не означава в никакъв случай, че газираните напитки са причина за този тип здравословни неприятности. Корелацията се представя с коефициент $\rho$, наричан коефициент на корелация, който за две случайни величини има вида:

$$\rho = \frac{cov(x,y)}{\sqrt{var(x).var(y)}}. \qquad ( \text{III—1} )$$

А когато разглеждаме само $n$ на брой данни за променливите $(y_i, x_i)$, $\rho$ се определя като:

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}. \qquad \text{( III—2 )}$$

Регресионният анализ, както споменахме, се концентрира върху причинно-следствените връзки и се стреми да направи оценка на конкретната сила на тези връзки. В контекста на горния пример регресионният анализ ще постави въпроса за връзката между средните температури през летните дни и консумацията на напитки, респ. броя на слънчевите удари.

## 2. Линейни регресионни модели. Допускания и характеристики

Линейните зависимости намират широко приложение във финансите и икономиката, благодарение на това, че са лесни за употреба и дават достатъчно точни за практически цели резултати. За да бъде оправдано прилагането на регресионните модели, трябва да имаме обосновано предположение за връзка и причинно-следствена зависимост между анализираните целева променлива и фактори (обясняващи променливи).

За да демонстрираме приложението на линейните регресионни модели, ще използваме една основна зависимост във финансите, а именно връзката между риск и възвръщаемост. Тази зависимост е фундаментална за много теоретични модели, като ние ще използваме модела за оценка на капиталови активи (МОКА/CAPM). CAPM е добре познат от изучавания курс по финанси, използва се в практиката, а освен това представя връзката между риск и възвръщаемост с линейна зависимост, което го прави много подходящ за демонстриране на възможностите на регресионния анализ. Основната зависимост, която ни дава моделът за оценка на капиталови активи е:

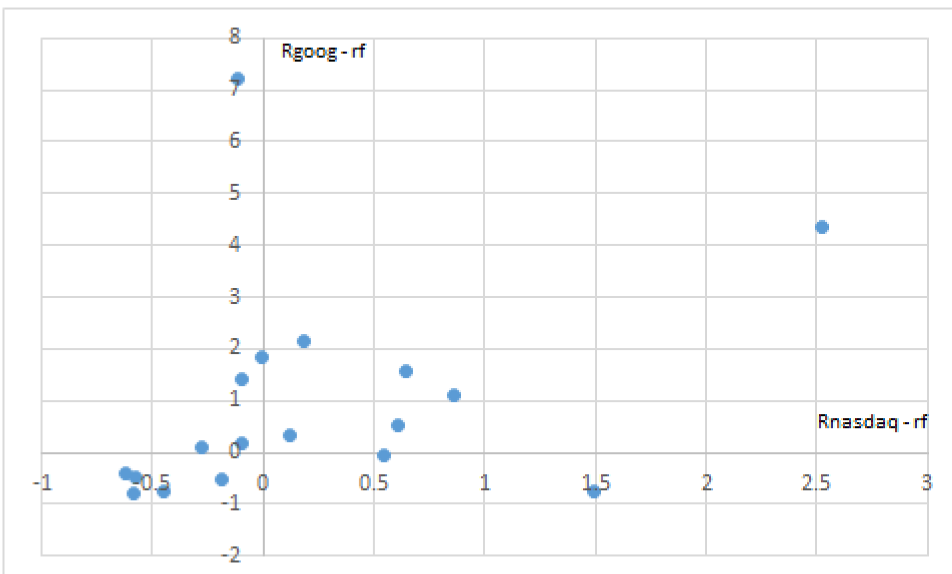
$$R = r_f + \beta(R_m - r_f) \text{ или } R - r_f = \beta\left(R_m - r_f\right), \qquad \text{( III—3 )}$$

където R е очакваната възвръщаемост на един актив, $r_f$ е безрисковата възвръщаемост, $R_m$ е възвръщаемостта на пазарния портфейл, а $\beta$ е коефициент, измерващ как промените във възвръщаемостта на пазара се отразяват на възвръщаемостта на анализирания актив (респ. ценна книга). Тъй като е трудно да проследим и оценим портфейл, който включва всички търгувани на пазара ценни книжа и активи (каквато е дефиницията на пазарния портфейл), обикновено $R_m$ се определя с помощта на пазарен индекс. Този метод не е перфектен, но ако индексът е съставен добре и включва в себе си достатъчен брой представителни ценни книжа/активи, то грешката генерирана от използването на „заместител“ вместо теоретично изисквания пазарен портфейл, ще бъде достатъчно малка.

Нека предположим, че имаме само необработени данни и не знаем нищо за линейните регресионни модели. Първата стъпка, с която можем да проверим за връзка между възвръщаемост на пазарния портфейл и на дадена ценна книга, е да визуализираме наличната информация на т.нар. точкови диаграми (още срещани като X-У или скатър диаграми). Фигура 2 представя една такава зависимост, основана на реални данни за Google и индекса Nasdaq[4], и показва, че наистина можем да търсим зависимост между двете величини.

Ако връзката между двете променливи е наистина линейна (нека си припомним, че в нашата въображаема ситуация не знаем все още нищо за съществуването на CAPM), то можем да я представим във вида:

$y = \alpha + \beta x$, където $\alpha$ и $\beta$ са коефициенти (т.е. константи), y е целевата променлива (в случая — свръх възвръщаемостта на ценната книга), а $x$ е факторната променлива (свръх възвръщаемостта на пазарния индекс).




*Фигура 2. Диаграма на рисковата премия (R—rf) на акция и пазарен индекс*

Така описаната линейна зависимост предполага, че след като веднъж определим стойностите на коефициентите $\alpha$ и $\beta$, ще разполагаме с модел, който ефективно и със сигурност ще дефинира каква е свръх възвръщаемостта на ценната книга при определена

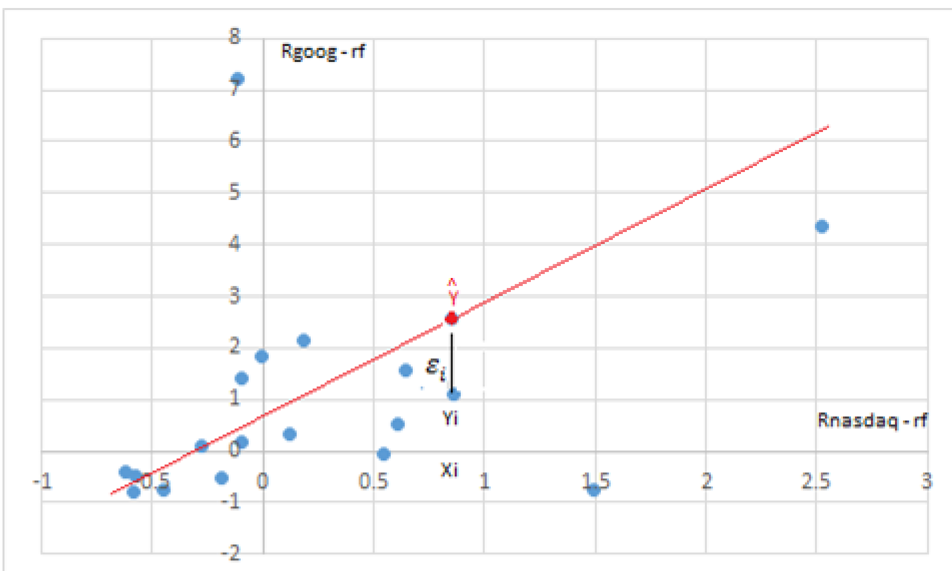
[4] Рисковата премия, или свръх-възвръщаемостта е изчислена на база 20 годишни държавни ценни книжа на САЩ, които са използвани като еталон за безрисков актив. На фигурата са представени само част от наличните данни с цел по-добрата им интерпретация.

възвръщаемост на пазарния индекс и известна $r_f$. В един свят на идеална линейна зависимост правата линия, определена с това уравнение, би трябвало да описва перфектно всички налични данни и да минава през всички точки, показани на Фигура 2.

Дори и ограниченото количество стойности видни от графиката показва, че подобна ситуация не е реалистична. Това означава, че правата, дефинирана от линейната зависимост, не може да минава през всички точки и наличието на определено ниво на точност (респ. на грешка) е неизбежно. Следователно пълният вид на зависимостта между двете изследвани променливи трябва да бъде:

$$y_i = \alpha + \beta x_i + \varepsilon_i, \qquad (\text{III—4})$$

където $\varepsilon_i$ представя именно нивото на грешка — отместването на конкретната стойност на целевата променлива спрямо правата линия, дефинирана от коефициентите $\alpha$ и $\beta$. Графично стойността на $\varepsilon_i$ е показана на Фигура 3 и именно тази разлика (наричана понякога и остатъчна стойност) предопределя методите, използвани в регресионния анализ, за да намерим стойностите на коефициентите $\alpha$ и $\beta$.




*Фигура 3. Остатъчна стойност, грешка и линейна регресия*

Въвеждането на компонент за грешка в линейната зависимост води след себе до:

- Необходимост да признаем, че определената от линейната зависимост права линия винаги ще съдържа в себе си елемент на приближение и неточност.

- Необходимост да познаваме причините, които налагат използването на приближение. Без да бъдем изчерпателни, ще посочим, че съществуват следните две обективни причини, поради които линейните регресионни модели не могат да игнорират наличието на грешка:

    - изследваните променливи и процеси могат да бъдат подвластни на случайни събития и външни фактори, които не се поддават на устойчиво прогнозиране.

Например финансовите пазари могат да бъдат повлияни от терористична атака или ураган с по-голяма от очакваната сила. Подобни събития не са периодични (а част от тях можем да приемем и за уникални поне в обозрим времеви период) и не се поддават на моделиране, следователно не могат директно да бъдат включени в нашата линейна зависимост.

    - Възможно е стойностите на целевата променлива да се определят и от други фактори, които не са включени в модела. В примера, който разглеждаме имаме само един фактор (рисковата премия на индекса), но в реалността възвръщаемостта на отделните акции може да зависи и от събития, които са специфични за компанията и не са отчетени в нашия модел.

В случаите, когато допълнителните фактори подлежат на измерване и оценка, можем да увеличим точността на модела посредством добавянето им към линейната зависимост. Когато обаче част от допълнителни фактори не са предварително изяснени или не могат да бъдат измерени, то тяхното интегриране в линейния модел не е толкова лесно.

    - Възможно е да има отклонения и грешки при измерването на стойностите на целевата променлива и факторите в регресионния модел.

Отклоненията могат да бъдат резултат от чисто технически грешки (в нашия пример тази възможност е минимизирана) или да се появят вследствие избраните методи за предварителна обработка на данните. Както вече коментирахме, изчисляването на възвръщаемост само в резултат от промените в цените пренебрегва допълнителната доходност от получаване на дивидент. И ако в примера с Google този ефект не е от съществено значение, в друга ситуация може да допринесе за увеличаване на общата стойност на грешката.

Преди да демонстрираме методите за оценка на коефициентите $\alpha$ и $\beta$, е необходимо да посочим допусканията, които трябва да бъдат в сила, за да можем да използваме адекватно линейния регресионен анализ.

*Таблица 4. Допускания и предпоставки на линейните регресионни модели*

| Допускане | Значение |
|---|---|
| Средната стойност на грешката трябва да е 0: $\mathrm{E}(\varepsilon) = 0.$ | Значението на това допускане е свързано с определянето на точните стойности на константите (параметрите) на регресионната зависимост. В случай че средната/очакваната стойност на грешките не е нула, то и получените параметри ще бъдат „изместени“ спрямо техните истински (верни) стойности. |
| Вариацията на грешката има крайна стойност, която е константна: $var(\varepsilon) = const < \infty.$ | Значението на това допускане е, че няма връзка между вариацията на грешката и стойностите на факторната променлива. Ако наистина липсва такава връзка, говорим за хомоскедастицитет на грешката, който за съжаление невинаги е изразен при финансовите и икономическите данни[5]. |
| Грешките (остатъчните стойности) при отделните наблюдения са независими една от друга: $cov(\varepsilon_i, \varepsilon_j) = 0$ за $\forall\ i \neq j.$ | Подобна независимост невинаги е налице при анализа на финансови данни. Ако данните съдържат информация за много на брой фирми през няколко периода, то винаги съществува възможност да сме изложени на въздействието на:<br>- общи икономически фактори, които ще окажат едновременно въздействие върху всички фирми в определен период;<br>- дълготрайни и специфични за фирмата процеси, които ще генерират корелация между данните за фирмата през отделни периоди;<br>- наличието на бизнес цикли за отделните индустрии, които са източник на зависимости за различни фирми през различни периоди от време. |
| Не трябва да има зависимост между грешката и някой от факторите: $cov(\varepsilon_i, x_j) = 0.$ | Това изискване е свързано с необходимостта ясно да бъдат дефинирани ендогенните и екзогенните променливи в един икономически модел. В случай че е налице зависимост между грешката и фактор, то тогава оценката на параметрите на регресионното уравнение е възможно да доведе до получаване на неточни („изместени“) стойности. |
| Грешката трябва да следва нормално разпределение: $\varepsilon \sim N(0, \sigma_\varepsilon^2).$ | Допускането за следвано нормално разпределение е важно, защото ни казва по какъв начин реално наблюдаваните стойности се отличават от получените с помощта на модела. С допускането, че тези разлики са нормално разпределени, ние реално казваме, че е малко вероятно те да получат екстремни стойности (т.е. малко вероятно е да имаме изключително голяма грешка). Освен това, дефинирайки $\sigma$ на разпределението на грешката, с помощта на допускането за нормално разпределение можем да оценим интервалите, в които с фиксирана вероятност ще се движи тази грешка. |

[5] Пример за хетероскедастицитет е връзката между образователна степен и нива на заплащане. Вариациите в заплатите на най-ниско образованите служители са по-малки, отколкото при тези с висока степен на образование и тясна специализация.

С помощта на изброените допускания и методите за оценка на параметрите на регресионната зависимост можем да дефинираме и някои по-важни характеристики на получените резултати.

*Таблица 5. Свойства на линейните регресионни модели*

| Характеристика | Значение |
|---|---|
| Точност и ефективност на получените резултати | Точността и ефективността на получените резултати означават, че например за линейната зависимост, с която анализираме CAPM, получените резултати за $\alpha$ и $\beta$ ще съответстват на техните истински стойности. |
| Консистентност на получените резултати | Консистентността на получените резултати, означава, че стойностите на параметрите на регресионната зависимост ще се доближават все повече до тяхното истинско значение с увеличаване на обема на анализираните данни. |
| Липса на изместване при оценка на параметрите на регресионната зависимост | Спрямо разглеждания пример това изискване означава, че $\mathrm{E}(\alpha) = \alpha$ и $\mathrm{E}(\beta) = \beta$. Иначе казано, няма систематично подценяване или надценяване на параметрите на регресионната зависимост спрямо техните истински/верни стойности. |

Анализирането на грешките е от важно значение за качеството на постигнатите резултати и ще бъде в центъра на нашето внимание при почти всеки разглеждан метод или модел. Трябва да имаме предвид обаче, че когато говорим да остатъчни стойности или грешка, ние имаме предвид значенията им в рамките на регресионния анализ. Качествата и допусканията на линейните регресионни модели не могат да отменят нашата собствена отговорност, когато избираме променливите на модела и задаваме връзката помежду им. Така например, който и метод за оценка на параметрите на линейна регресия да използваме, няма да получим добри резултати, ако истинската връзка между целевата и обясняващите променливи има силно изразен нелинеен характер.

## 3. Методи за оценка на параметрите на линейни регресионни модели

Съществуват различни методи, с чиято помощ можем да намерим параметрите на една линейна регресионна зависимост. Те са с различна степен на сложност и позволяват да отслабим в различна степен общите допускания на регресионните модели, което ги прави подходящи за решаването на различни практически задачи. В рамките на това изложение ще разгледаме тези от тях, които се използват най-широко, докато за останалите любознателният читател може да потърси информация в специализирани текстове по статистика.

## 4. Метод на най-малките квадрати

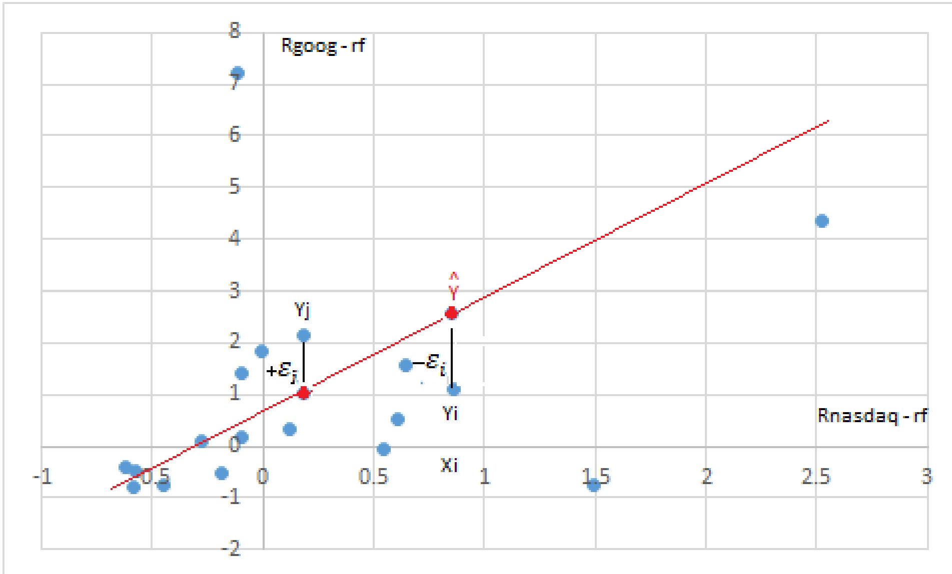
Методът на най-малките квадрати е най-простият за прилагане аналитичен инструмент за оценка на параметрите на една линейна регресионна зависимост. Както показва името му, неговата цел е да достигнем до такива резултати, които минимизират сумата от квадрата на разликите между реално наблюдаваните стойности и тези, получени с помощта на регресионното уравнение. Както е показано на Фигура 2, тази разлика е равна точно на остатъчната стойност или грешката за всяко едно измерване – $\varepsilon$. Следователно същността на метода на най-малките квадрати е да минимизира $\sum \varepsilon^2$.

Изборът да минимизираме квадрата на грешката не е случаен. Ако изберем да минимизираме простата сума от грешките (респ. остатъчните стойности), рискуваме да не получим точни резултати поради факта, че някои от разликите ще бъдат положителни, други — отрицателни, и те взаимно ще се компенсират. По този начин можем да посочим повече от една права линия, която ще минимизира тази проста сума – достатъчно е само тя да минава през точката, определена от средните стойности на наблюдаваните фактори и целева променливи (в нашия пример това е точката определена от $[\bar{x}, \bar{y}]$). Когато използваме квадрата на грешката обаче, няма да имаме проблем с това, дали реално наблюдаваните стойности са над или под тези, получени с помощта на модела (т.е. дали знакът на остатъчната стойност е положителен или отрицателен), като по този начин всяка разлика допринася за по-голяма обща сума на грешките без значение в каква посока е самото отклонение.

Ако се върнем към примерът с Google и единичен фактор то ще получим:

$$y_i = \hat{\alpha} + \hat{\beta} x_i + \varepsilon_i = \hat{y_i} + \varepsilon_i \ . \qquad \text{( III—5 )}$$

Трябва да отбележим, че получените крайни резултати за параметрите на регресионната зависимост ще бъдат също оценки – т.е. ще получим $\hat{\alpha}$ и $\hat{\beta}$. За момента не можем да кажем дали това са „истинските“ стойности на параметрите, поради което е по-коректно да използваме нотацията, посочваща, че става дума само за оценки на истинските стойности на $\alpha$ и $\beta$.




*Фигура 4. Метод на най-малките квадрати*

А функцията която трябва да минимизираме ще има вида:

$$E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i)^2 \qquad ( \text{III—6} )$$

Тъй като търсим оптимум на E, намираме първата производна спрямо $\alpha$ и $\beta$. Тези производни трябва да бъдат равни на нула, което ни дава:

$$\frac{\partial E}{\partial \hat{\alpha}} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\alpha} - \hat{\beta} x_i\right) = 0 \,, \qquad ( \text{III—7} )$$

$$\frac{\partial E}{\partial \hat{\beta}} = -2 \sum_{i=1}^{n} x_i \left(y_i - \hat{\alpha} - \hat{\beta} x_i\right) \quad = 0 \,. \qquad ( \text{III—8} )$$

Като имаме предвид, че при анализ на реална съвкупност от данни индексът $i$ в горните уравнения ще бъде от 1 до $n$, където $n$ е броят на данните за x и y, ще получим:

От израза $\frac{\partial \mathrm{E}}{\partial \alpha} = 0$ ще получим:

$$\sum_{i=1}^{n}\left(y_i - \hat{\alpha} - \hat{\beta}x_i\right) \quad = 0$$

$$\sum_{i=1}^{n} y_i - n\hat{\alpha} - \hat{\beta}\sum_{i=1}^{n} x_i = 0$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \rightarrow \sum_{i=1}^{n} y_i = n\bar{y}$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \rightarrow \sum_{i=1}^{n} x_i = n\bar{x}$$

$$n\bar{y} - n\hat{\alpha} - \hat{\beta}n\bar{x} = 0$$

$$\bar{y} - \hat{\alpha} - \hat{\beta}\bar{x} = 0$$

От израза $\frac{\partial \mathrm{E}}{\partial \beta} = 0$ ще получим:

$$\sum_{i=1}^{n} x_i\left(y_i - \hat{\alpha} - \hat{\beta}x_i\right) = 0$$

$\bar{y} - \hat{\beta}\bar{x} = \hat{\alpha}$ (от резултата от колона 1)

$$\sum_{i=1}^{n} x_i\left(y_i - \bar{y} + \hat{\beta}\bar{x} - \hat{\beta}x_i\right) = 0$$

$$\sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i + \hat{\beta}\bar{x}\sum_{i=1}^{n} x_i - \hat{\beta}\sum_{i=1}^{n} {x_i}^2 = 0$$

$$\sum_{i=1}^{n} x_i y_i - n\bar{y}\bar{x} + n\hat{\beta}\bar{x}^2 - \hat{\beta}\sum_{i=1}^{n} {x_i}^2 = 0$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{y}\bar{x}}{\sum_{i=1}^{n} {x_i}^2 - n\bar{x}^2}$$

По аналогичен начин могат да бъдат намерени параметрите на една линейна регресионна зависимост и в случаите на повече от един фактор. С тази разлика, че тогава ще трябва да решим една система с брой уравнения, колкото е броят на факторите в регресионния модел плюс едно за константата.

Ще разгледаме приложението на описания метод на най-малките квадрати, като използваме само тази част от данните за възвръщаемостта на Google и Nasdaq, които са показани на Фигура 2 и зададени в Таблица 6.

От данните в таблицата можем да изчислим необходимите за пресмятане на $\hat{\alpha}$ и $\hat{\beta}$ стойности:

$\bar{x} = 0.4956, \bar{y} = 0.5814$, n=17

$$\sum x_i \hat{y_i} = 0.1722 * 0.1722 + 0.3446 * 0.3446 + \cdots + (-0.7715 * -0.5884) = 23.21237$$

$$\sum {x_i}^2 = 0.1722^2 + 0.3446^2 + \cdots + (-0.5884)^2 = \ 22.0343$$

*Таблица 6. Примерни входни данни за линеен регресионен модел с един фактор*

| Google ($R_{GOOG}-r_f$) — Y | Nasdaq Composite 100 ($R_{NDX}-r_f$) — X |
|---|---|
| 0.1722 | 0.1722 |
| 0.3446 | 0.3446 |
| −0.7362 | −0.7362 |
| 2.1598 | 2.1598 |
| 1.5623 | 1.5623 |
| 1.8268 | 1.8268 |
| −0.0415 | −0.0415 |
| −0.4129 | −0.6197 |
| −0.5334 | −0.1914 |
| 0.5223 | 0.6113 |
| 1.1201 | 0.8611 |
| −0.7523 | 1.4945 |
| 1.4245 | −0.0946 |
| 0.1169 | −0.2803 |
| 4.3542 | 2.5224 |
| −0.4776 | −0.5779 |
| −0.7715 | −0.5884 |

След като заместим в изразите за $\alpha$ и $\beta$, получаваме:

$$\hat{\beta} = \frac{23.21237 - 17\text{x}0.5814\text{x}0.4956}{22.0343 - 17\text{x}(0.4956)^2} = 1.025488$$

$$\hat{\alpha} = 0.5814 - 1.025488\text{x}0.4956 \quad = 0.073168$$

Или получената линейна зависимост ще има вида: $y = 0.073168 + 1.025488x$, откъдето, като си припомним, че всъщност x и y обозначават разликата между възвръщаемостта на Nasdaq Composite 100 и Google получаваме:

$$R_{GOOG} - r_f = 0.073168 + 1.025488(R_{NDX} - r_f).$$

## 5. Модификации на метода на най-малките квадрати

Разгледаният до момента подход за оценка на параметрите на една линейна регресионна зависимост се основава на минимизиране на сумата на грешките (остатъчните стойности) и предлага ефективен и лесен за употреба математически апарат. Освен предимства той има някои недостатъци, свързани с изискванията, на които наблюдаваните процеси трябва да се подчиняват.

При анализа на допусканията, съпътстващи употребата на линейни регресионни модели, споменахме, че е необходимо стойностите на грешката да са независими (некорелирани) помежду си, както и вариацията на грешката да остава константна (т.нар. хомоскедастицитет). След като обективно наблюдаваме отклонения от тези изисквания в данните за реалните икономически и финансови процеси, остава отворен въпросът дали е оправдано да използваме демонстрирания метод на най-малките квадрати. За щастие методът на най-малките квадрати може да се модифицира, така че да запазим неговата валидност и в случаите, когато имаме хетероскедастицитет (т.е. вариацията на грешката се променя) или определена степен на корелация между остатъчните стойности (разликите между реално наблюдаваните стойности на целевата променлива и тези, получени с помощта на регресионния модел).

### 5.1. Метод на най-малките квадрати с теглови коефициенти

Използването на тази модификация на метода на най-малките квадрати позволява да се анализират процеси, при които няма взаимна корелация между отделните грешки (респ. остатъчни стойности), но е налице хетероскедастицитет. За да прибегнем до услугите на така модифицирания метод, трябва предварително да е известна формата, под която вариацията на грешката се променя.

При стандартния метод на най-малките квадрати стремежът е да се минимизира изразът:

$$\mathrm{E} = \sum (y_i - \widehat{y}_i)^2 = \sum (y_i - \hat{\alpha} - \hat{\beta} x_i)^2 \,, \qquad \text{( III—9 )}$$

докато при наличие на хетероскедастицитет можем да достигнем до ефективни оценки за параметрите $\hat{\alpha}$ и $\hat{\beta}$, ако намерим минимума на:

$$\mathrm{E}_{WLS} = \sum \omega_i (y_i - \widehat{y}_i)^2 = \sum \omega_i (y_i - \hat{\alpha} - \hat{\beta} x_i)^2, \qquad \text{( III—10 )}$$

където тегловите коефициенти са $\omega_i = \frac{1}{var(\varepsilon_i)}$ — обратно пропорционални на вариацията на грешката.

Процесът по установяване на стойностите на параметрите е същият, с тази разлика, че при израза на първата производна на $\mathrm{E}_{WLS}$ ще участват и тегловите коефициенти на грешката.

### 5.2. Метод на най-малките квадрати с последователни итерации

Тази модификация на метода на най-малките квадрати се прилага в случаите, при които е налице хетероскедастицитет и/или корелация между остатъчните стойности за анализираните данни, но без да е предварително известно как се променя вариацията на грешката и как се проявява корелацията.

Основната идея на тази модификация е да прилага последователно стандартния метод на най-малките квадрати или този с теглови коефициенти, докато се получи конвергенция на изчислената грешка (иначе казано — докато последователните итерации не водят до намаляване на грешката). Тъй като модификацията се използва в случаи, когато нямаме предварително информация за измененията в грешките и корелацията помежду им, стремежът е да минимизираме израза:

$$E_{IRLS} = \sum \omega_i(\beta)(y_i - \hat{y_i})^2 = \sum \omega_i(\beta)(y_i - \hat{\alpha} - \hat{\beta} x_i)^2 \,. \qquad \text{( III—11 )}$$

Изразът $\omega_i(\beta)$ показва, че тегловите коефициенти не са фиксирана величина (както е в предната модификация на метода на най-малките квадрати), а всяка следваща итерация използва изчислената структура на тегловите коефициенти до момента и се стреми да я подобри по начин, който да намали $E_{IRLS}$.

## 6. Метод на максималното правдоподобие

Този метод е приложим в случаите, когато е предварително известно какво разпределение следват остатъчните стойности (респ. грешката). В основните, допускания свързани с линейните  регресионни зависимости, изтъкнахме, че предположението за нормално разпределение на остатъчните стойности е важно, но не винаги е изпълнено на практика.

В случай че се наблюдават отклонения от това изискване за нормалност (например налице са големи отклонения и се наблюдават екстремни стойности), но можем  да намерим друго теоретично разпределение, което да отговаря на наблюдаваните остатъчни стойности, то методът на максимално правдоподобие позволява да направим реалистична оценка на параметрите на регресионната зависимост. Казано по друг начин — методът основан на максималното правдоподобие, може лесно да бъде адаптиран там, където основните допускания за метода на най-малките квадрати не са в сила, както и да бъде използван при по-тесен набор от предварителни допускания. В допълнение методи, свързани с принципите на максимална правдоподобност, могат да се използват за оценка на параметри, които са част и от нелинейни зависимости.

От статистическа гледна точка, когато анализираме една съвкупност от данни $\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_k}$, ние можем да приемем, че те са представители на една по-голяма съвкупност $\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_{k,\dots}}, \boldsymbol{x_n}$, и да дефинираме функцията на вероятността да наблюдаваме точно определена стойност. Ако разгледаме проблема от тази гледна точка, можем да представим изграждането на модел като търсене на такава функция, определяща вероятността за получаване на определена стойност, която да съответства максимално добре на разполагаемите данни и характеристиките на реалния процес, който те представят. Тогава анализът се свежда до намирането на такава функция на плътността на разпределението на $x$, която да представя най-добре извадката със стойности $\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_k}$. Тогава и решението на проблема ще бъде оценката на параметрите на разпределението (които участват и във функцията на плътността), или ако става въпрос за единствен параметър $w$, ще търсим $p(x_i) = f(x|w)$. Функцията $f(x|w)$ определя не едно, а множество от разпределения (т.е. фамилия от разпределения), а задачата на метода на максималното правдоподобие е да определим тази стойност на $w$, която отговаря най-точно на анализираните данни (т.е. едно от всички разпределения от съответната фамилия).

Ако $f(x|w)$ е функция на плътността на въпросната фамилия вероятностни разпределения и данните $\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_k}$ са независими една от друга, то тогава можем да запишем, че:

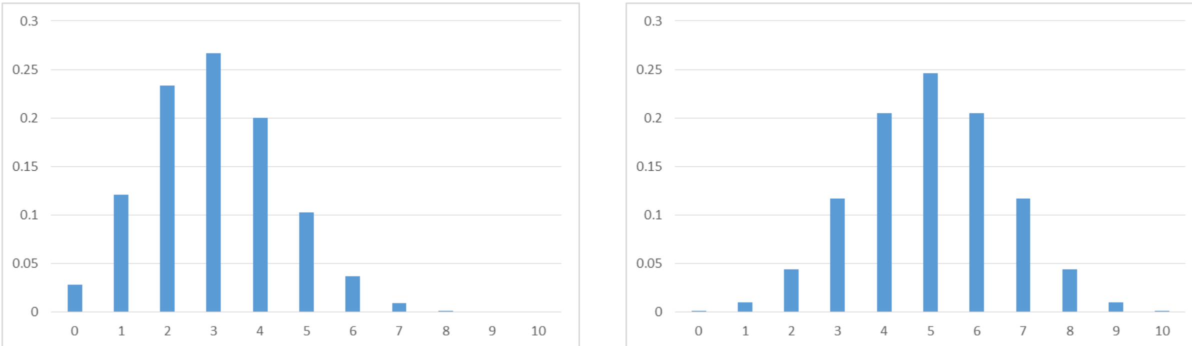
$$f(x = (x_1, x_2, \dots, x_k)|w) = f(x_1|w)f(x_2|w)\dots f(x_k|w) \qquad \text{( III—12 )}$$

Трябва да отбележим, че с $w$ бележим вектор, съдържащ параметрите на разпределението, чиято плътностна функция е $f(x|w)$. Т.е. методът може да се използва за оценка на повече от един параметър, като размерността на вектора $w$ зависи от вида на разпределението.

Ще разгледаме ситуация, подобна на представената от Myung [4], при която посоката на движение на цената на една акция е случайна величина и може да се покачва с вероятност $w$=0.3 и да спада с вероятност 1-0.3 = 0.7. Тогава, ако нас ни интересува вероятността да имаме $x$ на брой повишения в цената за последните 10 дни, функцията $f(x|w)$ ще има следният вид:

$$f(x|n = 10, w = 0.3) = \frac{10!}{\mathrm{x}!\,(10-\mathrm{x})!}0.3^{\mathrm{x}}(1-0.3)^{10-\mathrm{x}},$$

където параметрите на функцията са както броят повишения на цената (които, естествено могат да бъдат максимално 10), така и броят на дните, през които наблюдаваме движенията в цените. Ако променяме вероятността, с която цената може да се повиши, това ще доведе и до различна функция, респективно и до различно разпределение.




*Фигура 5. Вероятност за n повишения при w=0.3 (ляво) и w=0.5 (дясно)*

Както показва Фигура 5, различията в стойностите на параметрите могат да променят значително и формата на цялостното разпределение. При реалните изследвания използването на метода на максималната правдоподобност е свързано с решаване на задача, „обратна“ на токущо демонстрираната – да намерим какъв е този параметър $w$ (в случая вероятност от повишаване на цената), ако вече имаме набор от предварително известни данни (например наблюдения как се е движела цената на акцията през последните 3 месеца).

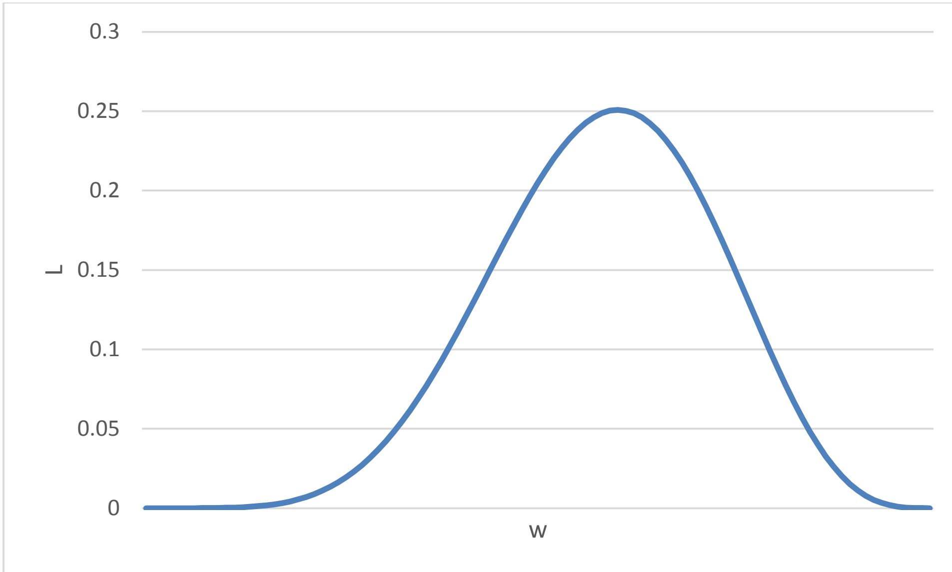
За да намерим оптималната стойност за параметъра $w$ дефинираме функцията за подобие като: $L(w|x) = f(x|w)$, което показва точно „обратният път“ по който се движим – как $w$ зависи от наблюденията, с които разполагаме. Ако вземем предвид равенство ( III—12 ) ще получим, че в общия случай функцията на подобие ще има вида:

$$L\big(w\big|x = (x_1, x_2, \dots x_k)\big) = f(x = (x_1, x_2, \dots x_k)|w) = \\ = f(x_1|w)f(x_2|w) \dots f(x_k|w) \qquad \text{( III—13 )}$$

Нека за цената на акцията имаме само едно наблюдение, което съответства на шест повишения. Тогава функцията на подобие L ще има вида (в съответствие с ( III—13 ) ):

$$L(w|n = 10, x = 6) = f(n = 10|x = 6, w) = \frac{10!}{6!\,(10-6)!} w^6 (1-w)^{10-6}\,.$$

Тъй като от двата параметъра (брой на повишенията и вероятност) променлива е само вероятността, то и графиката на функцията L е крива линия, както е показано на Фигура 6.




*Фигура 6. Графика на функцията за подобие L(w|n=10, x=6)*

Методът на максимално правдоподобие може да бъде използван и за повече от един параметър, при което вместо от крива линия L ще бъде представена от (хипер)равнина (с размерност броя на променливите параметри).

Стъпките за определяне на параметрите следват същата логика, както при метода на най-малките квадрати. Но докато там искахме да намерим минимума на сумата от остатъчните стойности (респ. показателя за грешка — сумата на квадрата на остатъчните стойности), в този случай търсим максималната стойност на показателя за подобност L:

$\frac{\partial L(w|x)}{\partial w} = 0$ и $\frac{\partial^2 L(w|x)}{\partial w^2} < 0$, като второто условие е проверка за намерен максимум.

Тъй като при много от използваните теоретични разпределения се налага да работим със степенуване, при метода на максималното правдоподобие често се използва един „трик", за да се улесни анализът – вместо функцията $L(w|x)$ се търси максимум за логаритъм от L – т.е. $\ln\big(L(w|x)\big)$. По този начин се избягва работата със степенни показатели и се улеснява значително решаването на горепосоченото уравнение, което вече има вида:

$$\frac{\partial \ln L(w|x)}{\partial w} = 0 \text{ (а проверката за максимум става } \frac{\partial^2 \ln L(w|x)}{\partial w^2} < 0\text{).} \qquad \text{( III—14 )}$$

Ако приложим техниката към примерната функция използвана за генерирането на Фигура 6, то ще получим:

$$\ln L(w|n = 10, x = 6) = \ln\left(\frac{10!}{6!\,(10-6)!} w^6 (1-w)^{10-6}\right) = \ln\frac{10!}{6!\,4!} + 6\ln w + 4\ln(1-w)$$

Оставяме любознателния читател да се убеди, че намирането на първата производна и приравняването ѝ на 0 е по-удобно в случаите, когато се използва логаритъм вместо първоначалния вид на функцията на подобност. За да намерим стойността на $w$, която ще максимизира стойността на L в нашия пример ще трябва да решим:

$$\frac{\partial ln(L)}{\partial w} = \frac{6}{w} - \frac{4}{1-w} = 0 = \frac{6-6w-4w}{w(1-w)} = \frac{6-10w}{w(1-w)} => w = 0.6\,.$$

Нека сега да предположим, че разполагаме вече с две наблюдения, при което имаме две стойности – 6 и 1. Тогава функцията за подобие, която ще трябва да максимизираме, ще има вида:

$$\ln L(w|n=10, x=(6,1)) = \ln(\frac{10!}{6!\,(10-6)!}w^6(1-w)^{10-6}\frac{10!}{1!\,(10-1)!}w^1(1-w)^{10-1}).$$

Или след използване на свойствата на логаритмичната функция при произведение:

$$\ln L(w|n=10, x=(6,1)) \\ = \ln\frac{10!}{6!\,4!} + 6\ln w + 4\ln(1-w) + \ln\frac{10!}{1!\,(10-1)!} + \ln w + 9\ln(1-w).$$

След като намерим производната на тази функция и я приравним на нула, ще получим:

$$\frac{\partial ln(L)}{\partial w} = \frac{6}{w} - \frac{4}{(1-w)} + \frac{1}{w} - \frac{9}{(1-w)} = 0 = \frac{6-6w-4w+1-w-9w}{w(1-w)} => w \approx 0.35$$

Трябва да отбележим, че не във всички случаи е възможно да намерим лесен за преобразуване израз на функцията за подобност. Тогава за намирането на оптималната стойност на търсените параметри се използват числени методи, които дават приблизителни до достатъчно точни резултати.

Използването на метода на максимално правдоподобие е съпътствано често и от един съществен проблем – необходимостта да покажем, че намерените стойности дефинират глобален, а не локален максимум. Този проблем е особено актуален в случаите, когато работим с по-сложни функции и се налага използването на числени методи за решаване на уравнението $\frac{\partial \ln L(w|x)}{\partial w} = 0$. Тогава решаването на задачата е свързано с прилагането на алгоритми (включително и евристични такива), които са извън обхвата на този текст.

## 7. Точност и грешки при линейните регресионни модели

При оценката на параметрите на линейната регресионна зависимост можем да използваме различни подходи в зависимост от характеристиките на анализираните процеси. Независимо от инструментариума, с който разполагаме, ще бъдем подвластни на един важен ефект – получените резултати ще зависят от данните, с които работим. Казано по друг начин — ако приложим същите аналитични инструменти към друга съвкупност от данни, ще получим и други стойности за параметрите $\alpha$ и $\beta$. Този неприятен факт изисква не просто да намерим конкретни стойности за отделните параметри, а да проверим до каква степен те са адекватни и заслужават доверие.

За да преценим доколко добри са получените резултати, е необходимо първо да анализираме остатъчните стойности или т.нар. грешка при съпоставянето на първоначалните данни и тези, получени от прилагане на регресионната зависимост. Като израз за това, как се движат тези остатъчни стойности, ще използваме средното квадратично отклонение, което е:

$$\varepsilon_i = (y - y_i)$$
$$var(\varepsilon) = \sigma_\varepsilon^2 = \sum (\varepsilon_i - E(\varepsilon))^2 \quad \text{( III—15 )}$$

Тази зависимост ни е позната и е свързана с изчисляването на средно квадратично отклонение на една случайна величина, но в контекста на линейните регресионни модели може да се опрости допълнително. Първото от допусканията, които дефинирахме в началото на главата, гласеше:

*Средната стойност на грешката трябва да е 0:*

$$\mathrm{E}(\varepsilon) = 0.$$

Откъдето следва, че израза за средно квадратичното отклонение ще бъде:

$$var(\varepsilon) = \sigma_\varepsilon^2 = \sum (\varepsilon_i)^2 \, , \quad \text{( III—16 )}$$

или на базата на разполагаемите данни ще имаме:

$var(\varepsilon) = \sigma_\varepsilon^2 = \frac{\sum(\varepsilon_i)^2}{n}$, или неизместената[6] оценка ще бъде $var(\varepsilon) = \sigma_\varepsilon^2 = \frac{\sum(\varepsilon_i)^2}{n-2}$, откъдето:

$$\sigma_\varepsilon = \sqrt{\frac{\sum(\varepsilon_i)^2}{n-2}} \, . \quad \text{( III—17 )}$$

[6] Ще припомним, че една оценка се счита за неизместена, ако очакваната стойност на разликата между оценките и истинската стойност е нула – т.е. $E(\hat{x} - x) = 0$.

Този израз се нарича стандартна грешка на регресията и се използва като един от основните показатели за точност и адекватност на регресионната зависимост. Колкото по-малка е стойността на този показател, толкова по-точно регресионната зависимост ще описва наблюдаваните данни (графично това означава, че точките на Фигура 4 ще бъдат по-близо до правата линия).

След като разполагаме с метод за изчисляване на стандартната грешка на регресията, можем да получим и оценка за точността (респ. на границите, в които ще се колебаят стойностите) на параметрите $\alpha$ и $\beta$:

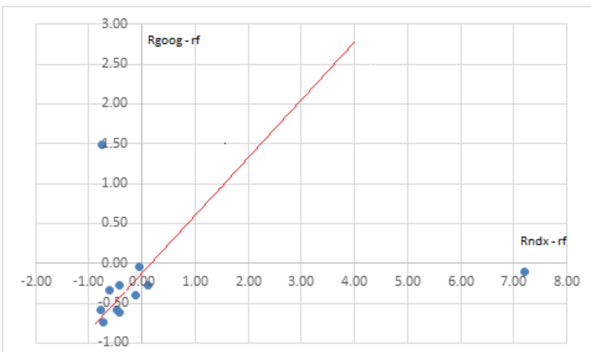
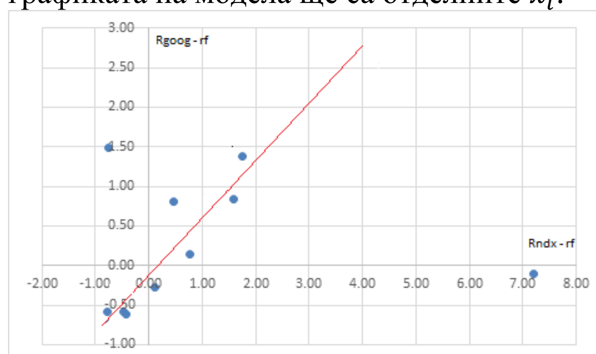
$$\sigma_\alpha = \sigma_\varepsilon \sqrt{\frac{\sum x_i{}^2}{n\sum(x_i - \bar{x})^2}}, \qquad ( \text{III—18} )$$

$$\sigma_\beta = \sigma_\varepsilon \sqrt{\frac{1}{\sum(x_i - \bar{x})^2}} = \sqrt{\frac{1}{\sum x_i{}^2 - n\bar{x}^2}} \;. \qquad ( \text{III—19} )$$

При практическото използване на регресионния анализ е важно, че колкото по-малки са тези стандартни отклонения, толкова по-точни (като средно) ще бъдат оценките на двата параметъра. Разбира се, до каква степен може да намалим така изчислените грешки, зависи от количеството и качеството на разполагаемата информация. Hill, Judge, Griffiths [5] посочват и систематизират следните важни принципи, които следват от изразите за точността на $\alpha$ и $\beta$:

*Таблица 7. Особености при анализа на стандартната грешка на линейната регресия*

| **Зависимост** | **Обяснение** |
|---|---|
| С колкото по-голяма съвкупност от данни работим, толкова по-малки ще са грешките и колебанията в стойностите на параметрите на регресионната зависимост. | Формално това се вижда от факта, че броят на данните $n$ участва в знаменателя на изразите за $\sigma_\alpha$ и $\sigma_\beta$.<br>Логически е ясно, че ако всеки един елемент от данните, с които разполагаме, носи в себе си информация за анализираните процеси, колкото повече такива елементи имаме, толкова повече информация ще се съдържа в цялата съвкупност. |

<table>
<tr><td>Колебанията в стойностите на параметрите на регресията зависят от стандартната грешка ($\sigma_{\varepsilon}$).</td><td>Колкото по-голяма е стойността на стандартната грешка, толкова по-отдалечени (като цяло) от линията на регресията ще бъдат реално наблюдаваните стойности. По този начин се намалява и общата точност на модела, а следователно и точността на получените оценки за параметрите на регресионната зависимост.</td></tr>
<tr><td>Връзката между $\sum(x_i - \bar{x})^2$ и колебанията в оценките на параметрите е обратнопропорционална.</td><td>Причината освен във формалния запис можем да намерим и в смисъла на този израз. Колкото по-малка е разликата между всяка една от наблюдаваните стойности $x_i$ и $\bar{x}$, толкова по-близки и събрани около графиката на модела ще са отделните $x_i$.</td></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td colspan="2">Фигура 7. Връзка между (x−xi) и стандартната грешка</td></tr>
<tr><td></td><td>Графично това е представено в лявата част на Фигура 7. Групирането на отделните наблюдения прави по-трудно да преценим дали получените параметри на регресионната зависимост са валидни и дали показаната линия е точна. Дясната част на графиката показва случая, при който имаме „разпръснати точки“, откъдето можем да оценим по-ефективно и дали правата линия „съответства“ на изследваните процеси.</td></tr>
<tr><td>Изразът $\sum x_i{}^2$ не участва при определяне на стандартното отклонение за коефициента $\beta$.</td><td>Причината за това е, че този израз е свързан с мащаба (иначе казано — отстоянието на наблюдаваните величини от $y$ (целевата променлива) на анализираните данни, но не и с начина, по който измененията на фактора влияят върху целевата променлива.</td></tr>
</table>

Необходимо е да познаваме характеристиките и методите за изчисляване на стандартна грешка на линейните регресионни зависимости поради следните причини:

- защото това е важен показател доколко данните, с които разполагаме, могат да бъдат използвани при търсене на общовалидна зависимост, с която да разкрием същинските характеристики на анализираните финансови процеси;
- защото ни позволяват да преценим дали получените оценки за параметрите на регресията съответстват да стойностите, които финансовата теория посочва.

Последното е особено съществено в случаите, когато трябва да проверим дали получените резултати наистина се движат в рамките, които са предписани от теоретичните финансови и икономически модели.

## 8. Линейни регресионни модели с повече от един фактор

В много случаи се налага да работим с повече от един фактор, който оказва влияние върху целевата променлива. Това се дължи както на сложността на изследваните финансови проблеми, така и на факта, че често искаме да обхванем различни аспекти от поведението на анализирания процес. В случая с анализа на възвръщаемостта на акциите на Google можем да направим проверка не само как тя зависи от движенията на индекса Nasdaq, но и как се влияе от инфлацията, броя на обявени фалити през годината, равнището на безработица и т.н. Колко точно фактора да анализираме, е въпрос, който е свързан както с обясняващата сила на създадения модел (за което ще говорим отново с подробности в следващата част), така и с нормалната икономическа логика.

При повече от един фактор линейната регресионна зависимост придобива следния вид:

$$y_i = \alpha + \beta_1 x_{i1} + \ \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \ . \qquad \text{( III—20 )}$$

Значението на отделните коефициенти бета в горната зависимост е близо до познатото от курсовете по микро- и макроикономика състояние „при равни други условия“. Иначе казано — коефициентът $\beta_2$ представя влиянието на фактора $x_2$, при условие че останалите променливи не се променят или тяхното влияние е елиминирано. Например как ще се промени възвръщаемостта на акциите на Google, ако всички други характеристики на икономиката се запазят, а само темпът на инфлация се увеличи с 2%.

Ако си представим, че константата в горната зависимост е всъщност коефициент $\alpha = \beta_0$, то можем да запишем последното уравнение като:

$$y_i = \beta_0 . 1 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \ . \qquad \text{( III—21 )}$$

Добавянето на единица след $\beta_0$ не е случайно, понеже ни позволява да въведем една измислена обясняваща променлива $x_0$, чиято стойност е винаги 1 (т.е. тя не променя съществуващата зависимост, но ни дава много по-удобен и хомогенен запис на регресионното уравнение). Тогава, ако разполагаме с *n* наблюдения на целевата променлива, ще имаме *n x k* стойности за факторите и *n* на брой остатъчни стойности. Това ни позволява да използваме матричен запис на анализирания модел и с негова помощ да получим съкратен израз за намиране на стойностите на параметрите на линейната регресионна зависимост:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n \times 1} = \begin{bmatrix} 1 & x_{11} & x_{12} & & x_{1k} \\ 1 & x_{21} & x_{22} & \vdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & & x_{nk} \end{bmatrix}_{n \times k} \begin{bmatrix} \beta_0 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}_{k \times 1} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n \times 1} \quad \text{( III—22 )}$$

Колоната от единици в горния запис идва точно от ново въведената обясняваща променлива $x_0$, чиято стойност по дефиниция е винаги 1. Следователно матричният запис на горната зависимост ще има вида:

$$Y = X\beta + \varepsilon\,, \quad \text{( III—23 )}$$

където Y, X, $\beta$ и $\varepsilon$ са матриците (респ. вектор стълбовете) от зависимост ( III—23 ).

Матричният запис ни позволява да използваме метода на най-малките квадрати, при това за *k* на брой фактори.

$$\varepsilon = Y - X\hat{\beta} \quad \text{( III—24 )}$$

$$[\varepsilon_1 \quad \varepsilon_2 \quad \dots \quad \varepsilon_n]_{1 \times n} \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{bmatrix}_{n \times 1} = [\varepsilon_1\varepsilon_1 + \varepsilon_2\varepsilon_2 + \dots + \varepsilon_n\varepsilon_n]\,. \quad \text{( III—25 )}$$

Като имаме предвид, равенствата ( III—24 ) и ( III—25 ) то можем да запишем, че сумата от квадратите на остатъчните стойности в линейната регресия е равна на:

$$\begin{gathered} \varepsilon'\varepsilon = \left(Y - X\hat{\beta}\right)'\left(Y - X\hat{\beta}\right), \\ \varepsilon'\varepsilon = Y'Y - \hat{\beta}'X'Y - Y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta}\,. \end{gathered} \quad \text{( III—26 )}$$

Резултатите от зависимост ( III—26 ) се получават, като отчетем свойствата на транспонирането на сума и произведение на матрици:

$$(A + B)' = A' + B' \ \text{ и } \ (\mathrm{AB})' = \mathrm{B}'\mathrm{A}' \ \text{ и } \ (\mathrm{A}')' = \mathrm{A}$$

Отново с отчитането на тези свойства, може да опростим допълнително ( III—26 ):

$$Y'X\hat{\beta} = (Y'X\hat{\beta})' = \hat{\beta}'X'Y$$
$$\varepsilon'\varepsilon = Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta} \qquad \text{( III—27 )}$$

За да намерим минимума на сумата на квадратите на остатъчните стойности (т.е. $\varepsilon'\varepsilon$) трябва както в Част III.3 да приравним на нула първата производна на ( III—26 ):

$$\frac{\partial\hat{\beta}'X'X\hat{\beta}}{\partial\hat{\beta}} = \frac{\partial\hat{\beta}'A\hat{\beta}}{\partial\hat{\beta}} = 2A\hat{\beta}$$ [7]

$$\frac{\partial e'e}{\partial\hat{\beta}} = -2X'Y + 2X'X\hat{\beta} = 0$$
$$X'Y = (X'X)\hat{\beta} \qquad \text{( III—28 )}$$
$$(X'X)^{-1}(X'X)\hat{\beta} = (X'X)^{-1}X'Y$$
$$\hat{\beta} = (X'X)^{-1}X'Y$$

Матричният запис изглежда по-труден за възприемане, но позволява да работим с една и съща формула при различен брой фактори. Освен това той позволява да автоматизираме изчисленията, поради което се използва широко от различни софтуерни приложения.

Възможно ли е да подберем такива променливи, при които няма смислена икономическа зависимост, и въпреки това да търсим оценка на параметрите на регресията? Технологично нищо не ни пречи да го направим, но остава въпросът за каква цел ще ползваме получените резултати и дали откритата „зависимост“ не е просто случайност. Методът на най-малките квадрати, неговите модификации или методът на максималното правдоподобие не ни дават окончателен отговор дали оценените параметри описват логически издържана или валидна зависимост. За да разширим нашето познание за изследваните процеси отвъд технологичните процедури за намиране на коефициентите на регресионната зависимост, е необходимо да познаваме методите за проверка на хипотези и оценка на обясняващата сила на съставените модели.

Ще демонстрираме предимствата на матричния запис, като разширим линейния регресионен модел за възвръщаемостта на Google с използването на три обясняващи променливи (а не само на една). За целта ще използваме три-факторния модел на Fama-

[7] Подробности за получения резултат има в част 2.3. за диференциране на матрици на The Matrix Cookbook на Kaare Brandt Petersen и Michael Syskind Pedersen [24].

French [6], който включва освен стойността на пазарен индекс разликата между възвръщаемостта на фирмите с малка и голяма пазарна капитализация, както и разликата във възвръщаемостта на акциите на установени компании спрямо тези на фирми с потенциал за растеж в бъдеще. Крайната зависимост има вида:

$$\left(R_{GOOG} - r_f\right) = \beta_0.1 + \beta_1\left(R_{NDX} - r_f\right) + \beta_2\left(SMB - r_f\right) + \beta_3\left(HML - r_f\right) + \varepsilon_i\,, \quad (\text{III—29})$$

където SMB е разликата във възвръщаемостта на фирмите с малка и голяма пазарна капитализация, а HML е разликата във възвръщаемостта на акциите на установени компании спрямо тези на фирми с потенциал за растеж в бъдеще. Данните за последните два показателя за реалните финансови пазари се публикуват на сайта на Кенет Френч[8].

Като използваме данните от Таблица 8 можем да определим вида и стойностите на необходимите ни матрици, а с помощта на познатите от линейната алгебра действия по транспониране, умножение и инвертиране на матрици да достигнем до крайния резутат от израз ( III—31 ).

$$Y = \begin{bmatrix} 0.1722 \\ 0.3446 \\ -0.7362 \\ 2.1598 \\ 1.5623 \\ 1.8268 \\ -0.0415 \\ -0.4129 \\ \vdots \\ -0.4776 \\ -0.7715 \end{bmatrix} = \begin{bmatrix} 1 & 0.1722 & 0.01 & -0.27 \\ 1 & 0.3446 & 0.2 & -0.14 \\ 1 & -0.7362 & 0.43 & 0.31 \\ 1 & 2.1598 & 0.23 & 0.18 \\ 1 & 1.5623 & 0.31 & -0.04 \\ 1 & 1.8268 & 0.25 & -0.44 \\ 1 & -0.0415 & 0.15 & -0.63 \\ 1 & -0.6197 & 0.56 & 0.02 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & -0.5779 & 0.45 & -0.29 \\ 1 & -0.5884 & 0.04 & 0.05 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \vdots \\ \varepsilon_{16} \\ \varepsilon_{17} \end{bmatrix} \quad (\text{III—30})$$

След като приложим операциите от зависимост ( III—28 ) върху данните от ( III—30 ) ще получим вектор-стълб за коефициентите $\hat{\beta}$:

$$\hat{\beta} = \begin{bmatrix} 0.18451953 \\ 1.07106275 \\ -0.50291555 \\ -0.12464224 \end{bmatrix} \quad (\text{III—31})$$

Откъдето получаваме, че линейната регресионна зависимост ще има вида:

$$(R_{GOOG} - r_f) = 0.18452 + 1.071063(R_{NDX} - r_f) - 0.503(SMB - r_f) + -0.1246(HML - r_f)$$

[8] http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

*Таблица 8. Линейна регресионна зависимост основана на три-факторния модел на Фама-Френч*

| **Google ($R_{GOOG}$-$r_f$) - Y** | **NDX 100 ($R_{NDX}$-$r_f$) – $X_1$** | **SMB (R-rf) – $X_2$** | **HLM (R-rf) – $X_3$** |
|---|---|---|---|
| 0.1722 | 0.1722 | 0.01 | −0.27 |
| 0.3446 | 0.3446 | 0.2 | −0.14 |
| −0.7362 | −0.7362 | 0.43 | 0.31 |
| 2.1598 | 2.1598 | 0.23 | 0.18 |
| 1.5623 | 1.5623 | 0.31 | −0.04 |
| 1.8268 | 1.8268 | 0.25 | −0.44 |
| −0.0415 | −0.0415 | 0.15 | −0.63 |
| −0.4129 | −0.6197 | 0.56 | 0.02 |
| −0.5334 | −0.1914 | 0.11 | −0.33 |
| 0.5223 | 0.6113 | 0.26 | 0.52 |
| 1.1201 | 0.8611 | 0.23 | 0.05 |
| −0.7523 | 1.4945 | 0.81 | −0.39 |
| 1.4245 | −0.0946 | 0.05 | −0.56 |
| 0.1169 | −0.2803 | 0.23 | −0.39 |
| 4.3542 | 2.5224 | 0.85 | −0.20 |
| −0.4776 | −0.5779 | 0.45 | −0.29 |
| −0.7715 | −0.5884 | 0.04 | 0.05 |

## 9. Задачи и въпроси за размисъл

1. Защо при метода на най-малките квадрати се използва разликата между наблюдаваните и получените с помощта на модела стойности, повдигната на втора степен? Защо не на трета степен?
2. Какво смятате, че ще се случи, ако някои от предпоставките, на които стъпва методът на най-малките квадрати, не са изпълнени за изследваните процеси?
3. Опитайте се да получите сами резултатите от Фигура 5 с Excel, като използвате функцията BINOMDIST. Потърсете в интернет какви са параметрите на тази вградена функция.
4. Вярно ли е, че ако една стойност на $w$ максимизира $\ln L(w|x)$, то тя ще максимизира и $L(w|x)$?
5. Опитайте да възпроизведете графиката на Фигура 6, като използвате Excel и вградената функция FACT. За какво служи тази функция и какви са нейните параметри?

## IV. Проверка на хипотези и използване на резултатите от линейните регресионни модели. Значимост и адекватност на получените резултати.

В заключението на предната част посочихме, че един от важните въпроси, на който трябва да отговорим, е дали получените от регресионния анализ стойности на отделните параметри отговарят на това, което финансовите и икономическите модели предполагат. Когато анализираме дадена зависимост, е най-добре ако разполагаме с истинските стойности за тези параметри, но обикновено те не са известни и това налага да работим с оценките, които сме получили на базата на разполагаемите данни.

Нека разгледаме отново получените резултати от предишната част. Като използвахме метода на най-малките квадрати и стойностите от Таблица 6, ние достигнахме до извода, че:

$$R_{google} - r_f = 0.073168 + 1.025488\left(R_{NDX} - r_f\right).$$

Или иначе казано − в тази линейна регресионна зависимост $\beta = 1.025488$ и $\alpha = 0.073168$. Тези стойности обаче бяха получени на базата на 17-те двойки стойности от таблицата. Ако вземем други 17 стойности от историческите данни за възвръщаемостта на Nasdaq Composite 100 и Google, можем да получим близки (но не същите) стойности. Но ако компанията се е променила значително или икономическите условия са други, е възможно да получим и много различни резултати.

Като финансисти нас ни интересува не само да познаваме начина, по който можем да оценим параметрите на регресията, а как получените оценки се съотнасят към стойностите, описващи истинската зависимост между възвръщаемостта на Google и Nasdaq. Казано по друг начин, как можем да проверим дали истинската стойност на коефициента $\beta$ е 1.02 или пък е 0.5? И изобщо как можем да сме сигурни, че линейната връзка, която моделът за оценка на капиталови активи предполага, е адекватна? Възможно ли е освен възвръщаемостта на пазара (в нашият пример представена с индекса Nasdaq Composite 100) да има и други фактори, които определят възвръщаемостта на акциите на Google? И ако има такива фактори, каква е тяхната тежест и значимост?

Отговори на всички тези въпроси ще се опитаме да дадем в рамките на тази част, като използваме статистическите техники за проверка на хипотези и показателите за статистическо съответствие.

## 1. Статистическа проверка на хипотези

След като проверяваме една хипотеза, винаги имаме поне две възможност – тя да бъде правилна или неправилна. При статистическата проверка на хипотези тези две възможности съответстват на:

- Нулева хипотеза, което съвпада (съответства на проверяваната хипотеза). Обикновено се отбелязва с $H_0$или $H_n$. В контекста на първия от въпросите които зададохме, нулевата хипотеза ще бъде „истинската стойност на коефициента $\beta$ е 1.02“.
- Алтернативна хипотеза, която съответства на останалите възможни варианти (т.е. това, което не съвпада с нулевата хипотеза). В контекста на първия от въпросите, които зададохме, алтернативната хипотеза ще бъде „истинската стойност на коефициента $\beta$ е различна от 1.02“. В различните литературни източници обозначаването на алтернативната хипотеза може да бъде $H_1$или $H_{\text{a}}$.

Алтернативната хипотеза може да бъде зададена по различни начини, които определят и смисъла, който влагаме в нея. В горния пример „истинската стойност на коефициента $\beta$ е различна от 1.02“ означава, че тя може да бъде както по-малка ($\beta < 1.02$), така и по-голяма ($\beta > 1.02$). Такъв тип хипотези се наричат двустранни, тъй като са възможни стойности на параметъра в два отделни интервала. При изследването на някои икономически процеси е възможно да имаме наложени ограничения (или от използвания теоретичен модел, или като резултат от други изследвания), които да изменят алтернативната хипотеза. Например, ако е известно, че стойността на коефициента $\beta$ не може да бъде по-голяма от 1.02, то тогава ще имаме едностранната $H_1$: $\beta < 1.02$.

По какъв начин можем да се ориентираме кога да използваме едностранна и кога − двустранна алтернативна хипотеза? Отговорът на този въпрос не е универсален, а зависи от конкретната проблемна ситуация. Най-често изборът между едностранна и двустранна хипотеза зависи от това дали:

- са налице определени теоретични изисквания към стойността на изследвания параметър (например, ако става въпрос за бета коефициент на ценна книга, може да очакваме, че неговата стойност ще бъде неотрицателна[9]);
- са налице специфични за проблемната ситуация фактори, които лимитират стойността на изследвания параметър. Ако проучваме как ще се отрази на доходността на акциите на Google навлизането на силно конкурентен пазар, то можем да дефинираме предварително, че очакваме увеличение на риска (а следователно и бета ще се увеличава);

[9] В рамките на модела за оценка на капиталови активи се допуска да има и отрицателни бета коефициенти (например за златото като търгуем актив).

- са налице някакви особени изисквания към резултатите от нашият анализ, които да налагат използването само на едностранна или двустранна алтернативна хипотеза.

След като дефинираме хипотезите, се нуждаем от метод, с чиято помощ да вземем решение за приемането или отхвърлянето им. Съществуват два принципни подхода, които можем да използваме, за да решим този проблем – посредством проверката за значимост и с използването на доверителни интервали. Проверката на хипотези в тези случаи е свързана със сравняването (което, разбира се, трябва да стане по статистически оправдан начин) на получените с помощта на наличните данни оценъчни стойности и тези, залегнали в нулевата/алтернативната хипотеза. Колкото по-голяма е разликата между заложената в хипотезата и получената на практика оценка, толкова по-вероятно е хипотезата да не бъде вярна и да трябва да я отхвърлим. Например, ако оценката е $\beta = 1.025488$, то много по-вероятно е да отхвърлим нулева хипотеза, че $\beta = 7$, отколкото нулева хипотеза, че $\beta = 1.02$. За да преценим обаче колко точно е „по-вероятно“, е необходимо да дефинираме начина по който определяме разликата, и какво означава това сравнение да бъде статистически оправдано.

## 2. Проверка на хипотези в контекста на линейните регресионни модели

Можем да улесним проверката на хипотези в контекста на линейните регресионни зависимости, ако се позовем на последното допускане от Таблица 4, а именно че грешката трябва да следва нормално разпределение:

$$\varepsilon \sim N(0, \sigma_\varepsilon^2). \qquad \text{( IV—1 )}$$

Допускането за нормално разпределение е важно, понеже от него произтича, че изчислените стойности за параметрите на регресионната зависимост ще бъдат нормално разпределени „около“ тяхната истинска стойност: $\alpha \sim N(\alpha_{\text{истинско}}, {\sigma^2}_\alpha)$ и $\beta \sim N(\beta_{\text{истинско}}, {\sigma^2}_\beta)$. Ако искаме да работим със стандартно нормално разпределение (при което ще припомним, че очакваната/средната стойност е 0, а стандартното отклонение е 1), то можем да преобразуваме горните изрази като:

$$\frac{\alpha - \alpha_{\text{истинско}}}{\sigma_\alpha} \sim N(0,1) \text{ и } \frac{\beta - \beta_{\text{истинско}}}{\sigma_\beta} \sim N(0,1). \qquad \text{( IV—2 )}$$

За съжаление, ние не знаем истинските стойности на параметрите $\alpha$ и $\beta$, а разполагаме само с получените оценки и техните грешки, които сме изчислили по описаните в предната част формули:

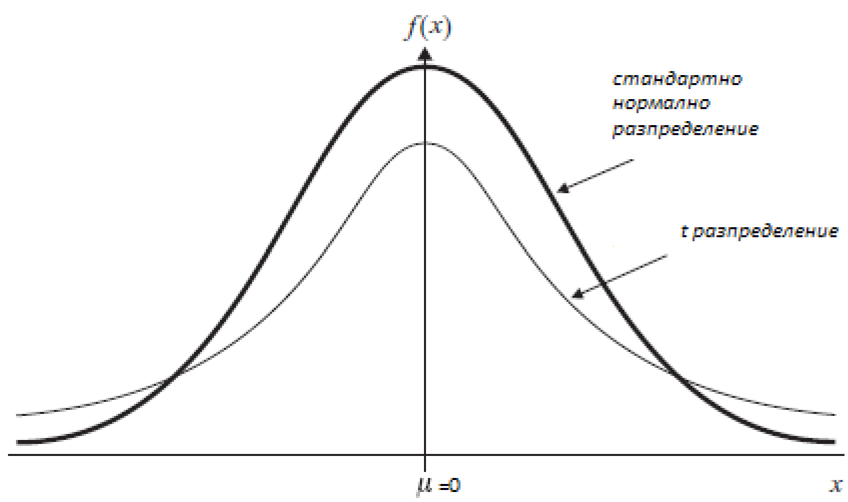
$$\sigma_\alpha = \sigma_\varepsilon \sqrt{\frac{\sum x_i^2}{N \sum (x_i - \bar{x})^2}} \quad \text{( IV—3 )}$$

$$\sigma_\beta = \sigma_\varepsilon \sqrt{\frac{1}{\sum (x_i - \bar{x})^2}} \quad \text{( IV—4 )}$$

Но ако използваме изчислената вариация с помощта на нашите формули за грешка, това означава, че въвеждаме допълнителен източник на неопределеност и тогава изразите за разликата между оценките на параметрите и тяхната истинска стойност, съотнесени към стандартната грешка, ще следват t разпределение с N-2 степени на свобода:

$$\frac{\alpha - \alpha_{\text{истинско}}}{\sigma_\varepsilon \sqrt{\frac{\sum x_i^2}{N \sum (x_i - \bar{x})^2}}} \sim t_{N-2} \ \text{ и } \ \frac{\beta - \beta_{\text{истинско}}}{\sigma_\varepsilon \sqrt{\frac{1}{\sum (x_i - \bar{x})^2}}} \sim t_{N-2}. \quad \text{( IV—5 )}$$

Както показва Фигура 8, стандартното нормално и t разпределение си приличат по форма. При увеличаване на степените на свобода на t разпределението ще наблюдаваме все по-голямо съответствие, като в случай че имаме безкрайно много степени на свобода, двете разпределения ще съвпадат.




*Фигура 8. Стандартно нормално разпределение и t разпределение*

След като познаваме връзката между стандартното нормално разпределение и t разпределението, можем да представим и стъпките, с чиято помощ да проверим нашите хипотези. За да избегнем разминавания с терминология, които не са рядкост, предварително ще посочим основните понятия, техните синоними и значение.

### TDIST(X, СТЕПЕНИ НА СВОБОДА, ЕДНО- /ДВУСТРАННО ОТНОШЕНИЕ)

Резултатът от тази функция е остатъчната вероятност на t разпределението. Например, ако получената стойност е 0.05, това ще съответства на 95% (0.95=1-0.05) равнище на значимост при отхвърлянето на нулевата хипотеза. Когато имаме едностранно отношение последният параметър е 1, а при двустранна зависимост е 2. В последните версии на Excel тази функция се поддържа, но се препоръчва използването на TDIST.RT и TDIST.2T.

### TINV(ВЕРОЯТНОСТ, СТЕПЕНИ НА СВОБОДА)

Резултатът от тази функция е стойността на t разпределението при зададените вероятност и степени на свобода. Тази функция намира такава стойност x, при която TDIST(x, степени на свобода, 2) да е равно на посочената вероятност.

*Таблица 9. Основни понятия при проверката на хипотези*

| **Понятие** | **Значение** |
|---|---|
| Ниво на съгласие<br>α–ниво (алфа-ниво) | Този показател се бележи с α и означава вероятността да отхвърлим нулевата хипотеза, при условие че тя е вярна. |
| Ниво на доверие<br>Гаранционна вероятност<br>1-α | Този показател се дефинира като 1-α и означава вероятността, че нулевата хипотеза е наистина вярна, при условие че с проверката сме я потвърдили като вярна (т.е. да потвърдим правилно нулевата хипотеза). |
| β–ниво | Този показател се бележи с β и означава вероятността да НЕ отхвърлим нулевата хипотеза, при условие че тя не е вярна. |
| Мощност на теста<br>Мощност на проверката<br>1-β | Този показател се дефинира като 1-β и е вероятността да отхвърлим нулевата хипотеза, при положение че тя не е вярна (т.е. да отхвърлим правилно нулевата хипотеза). |

Тъй като символите α и β вече използвахме при дефиниране на линейната регресионна зависимост, когато говорим за проверка на хипотези, ще използваме пълното име на понятията, а не тяхното обозначение.

## 2.1. Алгоритъм за проверка за значимост

1. Изчисляваме с помощта на метода на най-малките квадрати (или негова модификация) стойностите за параметрите на регресионната зависимост и стандартната грешка.

2. Определяме стойността на t разпределението, като използваме тези стойности на параметрите, заложени в нулевата хипотеза:

$$t = \frac{\beta - \beta_{\text{нулева хипотеза}}}{\sigma_\varepsilon \sqrt{\frac{1}{\sum (x_i - \bar{x})^2}}}. \qquad ( IV—6 )$$

Например, ако сме получили оценка за коефициента $\beta$ от $1.025488$, а нулевата хипотеза е, че истинската стойност е 1.02, ще имаме: $\beta = 1.025488$ и $\beta_{\text{нулева хипотеза}} = 1.02$.

3. В зависимост от степените на свобода (които ще определим от броя на данните, с чиято помощ е получена оценката за $\beta$) определяме теоретичната стойност на t разпределението. Както посочихме, степените на свобода са равни на $n - 2$, където $n$ е броят на наблюденията, с чиято помощ сме изчислили $\beta$.

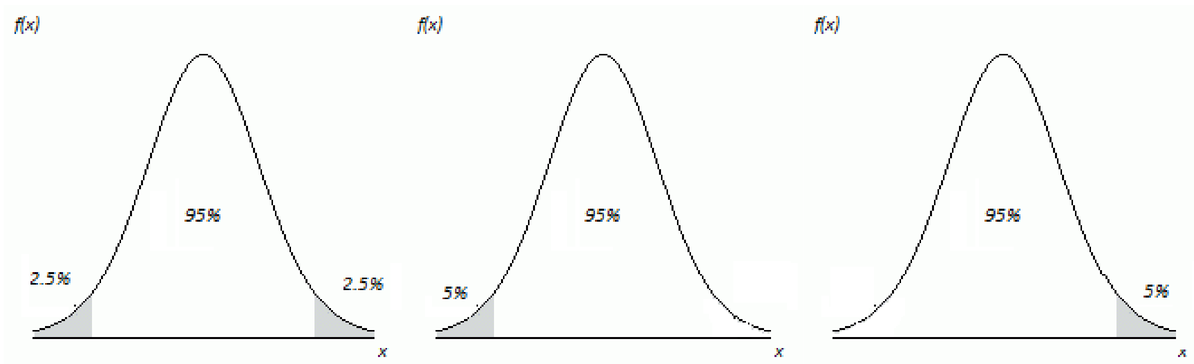
В примера с приложението на метода на най-малките квадрати използвахме общо 17 двойки стойности за възвръщаемостта на Google и Nasdaq, откъдето следва, че търсим теоретичната стойност на t разпределение с 15 (= 17 − 2) степени на свобода.

Точната теоретична стойност можем да получим директно от таблица или с помощта на компютър. Excel предлага две основни функции за работа с t разпределение – TDIST и TINV.

4. Следващата стъпка при проверката на хипотези е свързана с избиране на ниво на значимост. Изборът на конкретно ниво на значимост е важен не само за постигането на по-висока увереност в качествата на получените резултати, но и за да може да имаме сравнимост между отделните изследвания. Поради тази причина в иконометричните изследвания се използват „стандартизирани“ стойности за ниво на съгласие – най-често 5% (α = 0.05) или 1% (α = 0.01).

Графично нивото на значимост може да се представи с разделянето на фигурата, определена от статистическото разпределение, на две части – област на отхвърляне и не-отхвърляне на анализираната хипотеза. На Фигура 9 са представени как изглеждат регионите за приемане и отхвърляне на хипотези при ниво на съгласие α = 0.05 (5%) и:

- двустранна алтернативна хипотеза, например $\beta < 1.02$ и $\beta > 1.02$. Този случай е изобразен на лявата графика.
- еднострaнна алтернативна хипотеза, например $\beta < 1.02$. Този случай е изобразен на средната графика.
- еднострaнна алтернативна хипотеза, например $\beta > 1.02$. Този случай е изобразен на крайната дясна графика.




*Фигура 9. Региони на отхвърляне и не отхвърляне на хипотези*

Трите изброени случая са свързани с разглеждания пример, при който, както споменахме, нулевата хипотеза беше, че $\beta = 1.02$. Ако изберем различно ниво на съгласие, принципът който е показан на Фигура 9, се запазва, като се променят единствено лицата на регионите на отхвърляне и не-отхвърляне. Например при ниво на съгласие от 1% ще имаме съответно регион на не отхвърляне от 0.99 и на отхвърляне 0.01 или 0.005 при двустранна алтернативна хипотеза.

При анализа на финансови данни много често разполагаме с голямо количество стойности и дълги времеви редове. Изобилието от данни обикновено означава, че значението на случайните отклонения спрямо реалното развитие на изследвания процес намалява. Едновременно с това използването на повече данни при оценката на параметрите на линейната регресионна зависимост ще доведе до намаляване и на стандартната грешка (спрямо случаите, при които разполагаме с малко на брой наблюдения), защото повече данни обикновено означава и повече полезна информация, закодирана в тях. Съответно с колкото повече данни разполагаме (и ако нашият модел и тип зависимост са правилни), оценъчните стойности на параметрите на регресията ще се приближават до техните истински стойности. Ако използваме едно и също (и то голямо) ниво на значимост, можем да достигнем до ситуация, при която изобилието от данни само по себе си да спомогне за отхвърлянето на нулевата хипотеза. Причината е, че с намаляването на стандартната грешка оценката за стойностите на t, получени от стъпка 2 на алгоритъма, се увеличават (стандартната грешка е в знаменателя). Поради тази причина при използването на голям обем от данни е препоръчително също да се направи проверка и с нива на значимост, по-малки от „стандартните“ 5% [7].

5. Последната стъпка от алгоритъма е свързана с проверка дали получената стойност за t разпределението попада в региона на отхвърляне, в който случай можем да отхвърлим нулевата хипотеза.

Забележете, че концентрираме вниманието си преди всичко върху възможността да отхвърлим дадена хипотеза. Ако хипотезата не може да се отхвърли, това не означава автоматично, че тя е вярна и трябва автоматично да се приеме за вярна! Нека си представим, че сме разгледали поотделно две различни двойки хипотези:

$$H_0 : \beta = 1.02 \text{ и } H_1 : \beta < 1.02;$$

$$H_0 : \beta = 1.01 \text{ и } H_1 : \beta < 1.01.$$

Какво ще се случи, ако наличните данни не са достатъчни да отхвърлим нулевата хипотеза и в двата случая? Очевидно, ако не отхвърлянето е еквивалентно на приемане, ще трябва да обявим, че $\beta = 1.02$ и едновременно с това $\beta = 1.01$. Но истинската стойност на коефициента $\beta$ е една-единствена и следователно не може да е и 1.02, и 1.01!

## 2.2. Алгоритъм за проверка с ниво на съгласие

Алтернативен подход към проверката на хипотези може да се реализира, като поставим въпроса „обратно“, а именно: „с какво ниво на съгласие можем да твърдим, че проверяваната хипотеза е правилна?“. Например, ако при 95%-на гаранционна вероятност имаме хипотезата, че $\beta$ се намира в интервала [1.05; 1.3], то това означава, че очакваме при провеждането на множество опити (или респ. проучването на множество аналогични случаи) истинската стойност на коефициента в 95% от случаите да попада във въпросния интервал.

Съществува пряка зависимост между метода за проверка на хипотези с помощта на степени на значимост и доверителни интервали – изборът на 5% интервал на значимост практически означава и избор на 95% (1−0.05) доверителен интервал. Стъпките при проверка на хипотези с помощта на доверителни интервали са близки до тези, изложени в предната част.

1. Изчисляваме с помощта на метода на най-малките квадрати (или негова модификация) стойностите за параметрите на регресионната зависимост и стандартната грешка.
2. Определяме ниво на съгласие (ако гаранционната вероятност е 95% или 99%, това съответства на ниво на съгласие съответно 5% и 1%).
3. В зависимост от степените на свобода (които ще определим от броя на данните, с чиято помощ е получена оценката за параметрите на регресията) определяме теоретичната стойност на t разпределението. Както посочихме, степените на свобода са равни на $n - 2$, където $n$ е броят на наблюденията, с чиято помощ сме изчислили въпросните параметри.

4. Определяме региона (аналогичен на посочените области на Фигура 9), в който нулевата хипотеза не може да бъде отхвърлена по следния начин:

$$-t_{\text{теорет.}} < \frac{\beta - \beta_{\text{нулева хипотеза}}}{\sigma_\varepsilon \sqrt{\frac{1}{\sum(x_i - \bar{x})^2}}} < +t_{\text{теорет.}}, \qquad (\text{IV—7})$$

където $t_{\text{теорет.}}$ е получената теоретична стойност за t разпределението от стъпка 3 на алгоритъма.

5. Последната стъпка от алгоритъма е свързана с проверка дали стойността на регресионния параметър, заложена в нулевата хипотеза (т.е. $\beta_{H_0} = \beta_{\text{нулева хипотеза}}$), попада/съответства на интервала, определен в стъпка 4 на алгоритъма.

Проверката можем да извършим директно чрез заместване във формулата от стъпка 4 на алгоритъма:

$$\beta - t_{\text{теорет.}}(\sigma_\varepsilon \sqrt{\frac{1}{\sum(x_i - \bar{x})^2}}) < \beta_{H_0} < \beta + t_{\text{теорет.}}(\sigma_\varepsilon \sqrt{\frac{1}{\sum(x_i - \bar{x})^2}}). \qquad (\text{IV—8})$$

Как да изберем кой от двата алгоритъма за проверка на хипотези да използваме? И в двата случая би трябвало да достигнем до един и същи извод, така че основният аргумент в полза на един от подходите е преди всичко доколко е удобно да се използва в конкретната ситуация.

### 2.3. Кой взема крайното решение?

Изборът на метод, по който да извършим статистическата проверка на нужните ни хипотези, зависи от конкретните изследователски цели. Тъй като и двата метода дават едни и същи резултати, правилното им прилагане би трябвало да доведе и до едни и същи изводи относно значимостта и отхвърлянето/не отхвърлянето на тестваните хипотези.

Невинаги обаче резултатите от статистическата проверка на хипотези автоматично водят до вземането на финансови и икономически решения. Можем да групираме причините за това по следният начин:

- Причини, непосредствено свързани с характеристиките на резултатите

По дефиниция статистическата проверка на хипотези ни дава определен вероятностен резултат. Казано по друг начин, посредством посочените по-горе t тестове ние достигаме до определено ниво на вероятност, с което можем да твърдим, че данните показват зависимости, които съответстват на тестваните хипотези (а не са просто случаен резултат). Наличието на вероятност означава, че при прехвърлянето на статистическите резултати към финансови решения ще трябва да отчетем съществуването на несигурност, както и абсолютната стойност на получената вероятност.

Едновременно с това трябва да имаме предвид, че невъзможността да отхвърлим една хипотеза не означава автоматично, че тя е правилна и трябва да бъде приета. Изпадането в подобна заблуда може да доведе до вземането на погрешни решения и тълкуването по некоректен начин на получените резултати.

- Причини, свързани с допускането на грешки при статистическата проверка на хипотези

Както е известно от курса по статистика [8] [9], при статистическата проверка на хипотези можем да допуснем две принципни грешки:

- грешка от първи род, при която отхвърляме нулевата хипотеза, въпреки че тя е вярна;
- грешка от втори род, при която не отхвърляме нулевата хипотеза, въпреки че тя е грешна.

Двете възможности при които допускаме грешки, са показани в Таблица 10.

*Таблица 10. Възможни резултати и грешки при статистическата проверка на хипотези*

<table>
<tr><td colspan="2" rowspan="2"></td><td colspan="2">В действителност</td></tr>
<tr><td>Вярна е нулевата хипотеза</td><td>Грешна е нулевата хипотеза</td></tr>
<tr><td rowspan="2">Резултати от проверката</td><td>Отхвърляме нулевата хипотеза</td><td>Грешка от първи род<br>$P(\text{гр. от първи род}) = \alpha$</td><td>ОК<br>$P(OK) = 1 - \beta$</td></tr>
<tr><td>Не можем да отхвърлим нулевата хипотеза</td><td>ОК<br>$P(OK) = 1 - \alpha$</td><td>Грешка от втори род<br>$P(\text{гр. от втори род}) = \beta$</td></tr>
</table>

Можем да намалим вероятността за грешка от първи род, като намаляваме нивото на съгласие α (респ. като увеличаваме гаранционната вероятност), но това ще доведе автоматично до увеличаване на вероятността да допуснем грешка от втори род. Това означава, че не можем да постигнем по-висока точност в рамките на същата съвкупност от данни (в нашия пример изследваните седемнадесет двойки стойности за възвръщаемост на Google−Nasdaq) само и единствено посредством промяна на нивото на съгласие α / гаранционната вероятност 1−α.

Ако желаем да увеличим точността на нашите изследвания, ще трябва да се погрижим да анализираме по-голям набор от данни и/или да подобрим финансовия модел, на чиято база извършваме нашия анализ.

- Причини, свързани с тълкуването на резултатите

Статистическата значимост е важна и увеличава нашето знание за изследваните финансови процеси. Но сама по себе си тя невинаги води до промяна в пазарното поведение на икономическите агенти и начина, по който те вземат решения. В контекста на нашият пример е напълно възможно да получим, че хипотезата за стойности на коефициента $\beta$ в интервала [1.02;1.3] е статистически значима, но това да не доведе до промяна в нашето пазарно поведение.

Можем да обобщим, че методите за статистическа проверка на хипотези са един много полезен инструмент за анализ и проверка на нашите предположения и резултати, но крайните решения трябва да бъдат вземани в един по-широк контекст, който отчита всички аспекти на изследвания проблем, както и поставените за решаване финансови цели и задачи.

## 2.4. Съвместна проверка на повече от една хипотези

Като използваме t разпределението и t тест, можем да направим проверка на различни прости хипотези (например като разгледаните примери за стойностите на параметрите на регресията). В случаите, когато се налага да работим с повече от една обясняваща променлива и искаме да проверим хипотези, отнасящи се до едновременното въздействие на няколко от тях, t разпределението не е достатъчно, за да осъществим намеренията си.

Защо е необходимо да разглеждаме специален случай при едновременната проверка на няколко допускания, дефинирани като една хипотеза? Не е ли по-просто да направим t тест за всяко едно от условията поотделно и да анализираме всички резултати едновременно? За съжаление, едновременната проверка на отделните условия с помощта на отделни t тестове не е достатъчна и ще се опитаме да го илюстрираме с пример.

Нека в нашия модел да въведем още една променлива (освен възвръщаемостта на Nasdaq) – например пазарния дял на телефоните с Android (Google е основният поддръжник и двигател в развитието на тази операционна система). Тогава нашата регресионна зависимост ще изглежда като:

$$R = r_f + \beta_1\left(R_m - r_f\right) + \beta_2 \mathrm{A}_{\text{дял}} + \varepsilon\,. \qquad (\text{IV—9})$$

В този случай параметърът $\beta_2$ указва влиянието, което има пазарният дял на мобилните телефони с Android върху възвръщаемостта на акциите на Google. Ако искаме да тестваме хипотезата, че пазарният дял на Android телефоните няма значение за възвръщаемостта на акциите, можем да опитаме с t тест и нулева хипотеза от вида $\beta_2 = 0$. Ако искаме да тестваме хипотезата, че движението на индекса Nasdaq не оказва влияние на възвръщаемостта на акциите на Google, то тогава можем да опитаме с t тест и нулева хипотеза от вида $\beta_1 = 0$.

Причините да не можем да използваме резултатите от горните два t теста, за да проверим дали „едновременно и индекса и пазарният дял на Android телефоните не оказват влияние върху възвръщаемостта на акциите на Google" са:

- Възможно е t тестът да не може да отхвърли нулевата хипотеза (в нашия случай − че факторът няма значение), ако данните от извадката са много малко или съдържат много съществени случайни колебания. Наслагвайки двата t теста обаче, ние не можем да кажем със сигурност „колко [много] случайни" трябва да бъдат колебанията, така че да не можем да отхвърлим влиянието и на двата фактора;
- Като „съберем" двата t теста, не можем да определим директно какъв ще бъде доверителният интервал (респ. степента на сигурност) на така получената съвкупност;
- Ако имаме корелация между факторите в регресионния модел, това ще доведе до увеличаване на стандартната грешка.

Трябва да отбележим, че включването на повече от една променлива при дефинирането на тестваната хипотеза не означава автоматично, че не може да използваме t тест. За да прибегнем до други аналитични средства, трябва не просто да имаме хипотеза с повече от една променлива, а наличието едновременно на няколко условия. Така например „стойността на коефициента $\beta_1$ е равна на стойността на коефициента $\beta_2$" съдържа в себе си едно условие и позволява използването на t тест, докато „стойността на $\beta_1$ е нула и едновременно с това $\beta_2$ е нула" дефинира две условия и налага прилагането на други методи за проверка на хипотези.

Основният от тези методи е наречен F−тест по името на известният статистик Роналд Фишер и работи посредством дефинирането на две регресионни зависимости – неограничена и ограничена. При първата от тях стойностите на параметрите се определят от данните в извадката, докато при втората регресионна зависимост са изкуствено наложени ограничения върху стойностите на отделните коефициенти $\beta_1$ и $\beta_2$. Въз основа на получените остатъчни стойности при двете регресионни зависимости се дефинира стойността на F−теста:

$$F\text{тест} = \frac{RSS_{\text{ограничена}} - RSS_{\text{неограничена}}}{RSS_{\text{неограничена}}} \cdot \frac{\text{брой наблюдения} - \text{брой фактори}}{\text{брой ограничения}}, \qquad (\text{IV—10})$$

където $RSS_{\text{ограничена}}$ е сумата от квадратите от остатъчните стойности за регресионната зависимост с наложени ограничения, а $RSS_{\text{неограничена}}$ е сумата от квадратите на остатъчните стойности за регресионната зависимост без наложени ограничения.

Идеята на получения резултат е да сравним каква ще бъде сумата от всички отклонения при съществуването и липсата на ограничения върху параметрите на регресионната зависимост. Ако разликата в тази сума е малка при двата случая (т.е. $RSS_{\text{ограничена}}$ и $RSS_{\text{неограничена}}$ са близки като стойност), то това ще означава, че наличните данни подкрепят налагането на тези ограничения. Казано по друг начин – ограниченията не противоречат на това, което данните съдържат в себе си като информация за анализираните процеси. В случай че лимитирането на стойностите на част от регресионните параметри доведе до значителна разлика $RSS_{\text{ограничена}} - RSS_{\text{неограничена}}$, това означава, че въпросните ограничения водят до по-голяма грешка и следователно не се подкрепят от използваните данни.

Възможен е случай, при който сумата от квадратите на остатъчните стойности е идентична за двете регресионни зависимости и тогава $RSS_{\text{ограничена}} - RSS_{\text{неограничена}}$ ще бъде равна на нула. Такъв резултат бихме имали тогава, когато въпросните ограничения са изначално заложени в изследваните данни. За читателя предоставяме въпроса защо и дали е реалистично да има ситуация, при която $RSS_{\text{ограничена}} < RSS_{\text{неограничена}}$. Останалите елементи от формулата за F‒теста служат за „калибриране“ на получения резултат спрямо броя на елементите в анализираната съвкупност и броя на наложените ограничения.

След като веднъж получим резултат от формулата за $F$тест, той трябва да бъде сравнен с теоретичната стойност на F разпределението, което зависи от два параметъра – брой на наложените ограничения и *(брой на единиците в анализираната съвкупност – брой фактори)*. Окончателната проверка на заложените хипотези (като не забравяме, че получените от формулата стойности за F‒теста представят нулевата хипотеза) може да направим по вече описаните алгоритми посредством доверителни интервали или проверка за значимост.

Ще проследим етапите от прилагането на F‒теста, като използваме примера за Google, възвръщаемостта на индекса Nasdaq и пазарния дял на мобилните устройства с операционна система Android. Регресионната зависимост има вида $R = r_f + \beta_1(R_m - r_f) + \beta_2 \text{А}_{\text{дял}} + \varepsilon$, а ние желаем да проверим хипотезата, че сумата $\beta_1 + \beta_2 = 1$.

За да използваме F‒тест, е необходимо първо да определим две неща:

- броя на ограниченията в ограничената регресионна зависимост

Броят на ограниченията е важен елемент от цялостната проверка на хипотезата, защото той участва пряко при оценката на стойността на показателя $F$тест и непряко ‒ при определяне на сумата на квадратите на остатъчните стойности на регресията с наложени ограничения (т.е. $RSS_{\text{ограничена}}$). Колко са ограниченията, можем да определим, като анализираме логически, как те се отразяват върху стойностите, които могат да получат параметрите на регресионната зависимост. Едновременно с това бърз метод да проверим дали полученият от нас брой е верен е да го сравним с броя на равенствата, описващи ограниченията – броят на дефинираните равенства трябва да бъде равен на броя на наложените ограничения.

В конкретния пример имаме налице само едно ограничение (забележете, че независимо от факта, че в $\beta_1 + \beta_2 = 1$ участват два параметъра, ограничението е само едно – за тяхната сума)[10].

- точния вид на ограничената регресионна зависимост

От посочените ограничения и първоначалния вид на регресионната зависимост ще получим следната система:

$$\begin{gathered} \beta_1 + \beta_2 = 1 \\ R = r_f + \beta_1 (R_m - r_f) + \beta_2 \text{А}_{\text{дял}} + \varepsilon \, , \end{gathered} \qquad \text{( IV—11 )}$$

откъдето, като заместим $\beta_1 = 1 - \beta_2$, ще получим:

$$R = r_f + (R_m - r_f) + \beta_2 \left( \text{А}_{\text{дял}} - (R_m - r_f) \right) + \varepsilon \qquad \text{( IV—12 )}$$

или след преобразуване:

$$R - (R_m - r_f) = r_f + \beta_2 \left( \text{А}_{\text{дял}} - (R_m - r_f) \right) + \varepsilon \, . \qquad \text{( IV—13 )}$$

При практическото изчисляване на параметрите на регресионната зависимост може да се използва директно получената ограничена регресионна зависимост. Тъй като в повечето случаи изчисленията се правят с помощта на специализиран софтуер, за улеснение могат да се въведат и нови променливи, като $V = R - (R_m - r_f)$ и $\text{Т} = \left( \text{А}_{\text{дял}} - (R_m - r_f) \right)$. Идеята на тези „изкуствени променливи" е, че не е необходимо софтуерът да поддържа сложни формули за анализ на ограничената регресия, а можем предварително да изчислим V и T[11] и да предоставим на софтуера за анализ една съвсем „обикновена" регресионна зависимост:

$$V = r_f + \beta_2 T + \varepsilon \, .$$

---

[10] Ако ограниченията бяха зададени във вида $\beta_1 = 1, \beta_2 = 0.5$, то щяхме да имаме две равенства и съответно две ограничения.

[11] Можем предварително да ги изчислим, понеже отделните стойности за $R, R_m$ и $\text{А}_{\text{дял}}$ са ни известни.

Допълнително улеснение за нас е фактът, че в контекста на F–теста за нас от първостепенно значение е само сумата от квадратите на остатъчните стойности на ограничената регресия. Следователно не е необходимо да търсим и анализираме икономически и финансов смисъл зад променливите V и T – те служат единствено и само да улеснят нашите изчисления и да намалят вероятността да допуснем техническа грешка.

### 2.5. Обясняваща сила на регресионната зависимост

Дефинирането на хипотези, включващи повече от едно ограничение между параметрите на регресионната зависимост, има един интересен частен случай – когато нулевата хипотеза е, че всички $\beta_i (i > 0)$ са нулеви. Казано иначе – ако не можем да отхвърлим нулевата хипотеза, сме изправени пред случай, при който нито един от факторите не ни помага да обясним измененията в целевата променлива. Тогава изследваната от нас регресионна зависимост няма да допринесе съществено за разбирането на проучваните финансови и икономически процеси и е необходимо да потърсим и включим в анализа си други обясняващи променливи.

Когато имаме достатъчно основание, за да отхвърлим нулевата хипотеза, е добре да разполагаме и със средство, с което да оценим степента, в която подбраните фактори обясняват промените в стойностите на целевата променлива. От дефиницията на отделните елементи на регресионната зависимост можем да кажем, че колкото по-малка е сумата от остатъчните стойности (респ. грешката), толкова по-голяма част от измененията на анализираните финансови показатели могат да бъдат обяснени с помощта на участващите в регресията променливи. И тъй като са важни отклоненията както в положителна, така и в отрицателна посока, е логично да се използва сумата от квадратите на остатъчните стойности (RSS).

Поради факта, че финансовите показатели имат различна размерност, не можем да използваме директно абсолютната стойност на RSS, тъй като това би довело до трудности при сравняването на модели с различна размерност. Няма как да определим дали сума от квадратите на остатъчните стойности от например 152 показва висока или ниска точност на регресионния модел, без да вземем предвид всички негови характеристики. За да се постигне сравнимост и да се отчетат всички характеристики на регресионната зависимост, се използват различни методи за „мащабиране“ на RSS, които да доведат до интуитивни и лесни за използване показатели за обяснителната сила на получените резултати.

Един от тези показатели е $R^2$, който се дефинира като квадрата на коефициента на корелация между стойностите на обясняващата променлива от анализираната извадка и стойностите, изчислени посредством регресионната зависимост. Тази дефиниция е удобна поради факта, че ни дава директно и начин, по който може да се изчисли $R^2$, но не ни казва какъв е смисълът на показателя и защо неговата употреба е оправдана. За да отговорим на тези въпроси, ще се опитаме да достигнем до концепцията за $R^2$ посредством сумата на квадратите на остатъчните стойности на регресията. Определянето на $R^2$ с помощта на RSS дава представа за интервала, в който се движат получените стойности, и какъв е техният смисъл с контекста на икономическите и финансовите модели.

Движенията на обясняващата променлива могат да бъдат представени на базата на отклоненията от средната им стойност – например за възвръщаемостта на Google, това ще бъде $R_{GOOG} - \overline{R_{GOOG}}$. Когато сумираме тези отклонения за всички единици от анализираната съвкупност и ги повдигнем на квадрат (за да не се компенсират взаимно стойностите над или под средната), ще получим, че:

$$\begin{aligned}\sum(R_{GOOG} - \overline{R_{GOOG}})^2 &= \\ &= \sum(R_{GOOG\,(\text{регресия})} - \overline{R_{GOOG}})^2 + \sum(R_{GOOG} - R_{GOOG\,(\text{регресия})})^2 .\end{aligned} \qquad (\text{ IV—14 })$$

Горната зависимост показва, че отклоненията спрямо средната стойност могат да бъдат разделени на две части – такива, които вече са отчетени от регресионната зависимост ($\sum(R_{GOOG\,(\text{регресия})} - \overline{R_{GOOG}})^2$) и такива, които не са „уловени“ от нея ($\sum(R_{GOOG} - R_{GOOG\,(\text{регресия})})^2$). Колкото по-голям е относителният дял на отчетените отклонения, толкова по-добра е точността и обясняващата сила на анализираната регресионна зависимост. Следователно $R^2$ може да се дефинира и като относителен дял на „уловените“ колебания:

$$R^2 = \frac{\sum(R_{GOOG\,(\text{регресия})} - \overline{R_{GOOG}})^2}{\sum(R_{GOOG} - \overline{R_{GOOG}})^2} . \qquad (\text{ IV—15 })$$

Ако в горното отношение отчетем, че:

$$\begin{aligned}\sum(R_{GOOG\,(\text{регресия})} - \overline{R_{GOOG}})^2 &= \\ &= \sum(R_{GOOG} - \overline{R_{GOOG}})^2 - \sum(R_{GOOG} - R_{GOOG\,(\text{регресия})})^2 ,\end{aligned} \qquad (\text{ IV—16 })$$

то тогава можем да обясним и защо стойностите на $R^2$ са винаги в интервала [0;1] – тъй като:

$$\begin{aligned}R^2 &= \frac{\sum(R_{GOOG} - \overline{R_{GOOG}})^2 - \sum(R_{GOOG} - R_{GOOG\,(\text{регресия})})^2}{\sum(R_{GOOG} - \overline{R_{GOOG}})^2} = \\ &= 1 - \frac{\sum(R_{GOOG} - R_{GOOG\,(\text{регресия})})^2}{\sum(R_{GOOG} - \overline{R_{GOOG}})^2}\end{aligned} \qquad (\text{ IV—17 })$$

В случаите, когато получените с помощта на регресионната зависимост стойности съвпадат напълно, то тогава за всяка една стойност от анализираната съвкупност ще имаме $R_{GOOG} = R_{GOOG\,(\text{регресия})}$ и получената регресионна зависимост ще обяснява перфектно промените в целевата променлива. Тогава и показателят $R^2$ ще бъде равен на единица:

$$R^2 = 1 - \frac{\sum\left(R_{GOOG} - R_{GOOG\,(\text{регресия})}\right)^2}{\sum(R_{GOOG} - \overline{R_{GOOG}})^2} = 1 - \frac{0}{\sum(R_{GOOG} - \overline{R_{GOOG}})^2} = 1\,. \qquad (\text{IV—18})$$

В случаите, когато регресионната зависимост не обяснява нищо от промените в целевата променлива, то сумата $\sum\left(R_{GOOG} - R_{GOOG\,(\text{регресия})}\right)^2 = \sum(R_{GOOG} - \overline{R_{GOOG}})^2$ и тогава показателят $R^2$ ще бъде равен на нула:

$$R^2 = 1 - \frac{\sum(R_{GOOG} - R_{GOOG\,(\text{регресия})})^2}{\sum(R_{GOOG} - \overline{R_{GOOG}})^2} = 1 - 1 = 0 \qquad (\text{IV—19})$$

С помощта на показаните формули можем да изчислим $R^2$ и да се възползваме от този лесен за възприемане показател до каква степен използваната регресионна зависимост обяснява промените в стойностите на обясняващата променлива.

Освен предимства $R^2$ има и редица ограничения и особености, които могат да доведат до получаването на недотам коректни резултати и да засилят стремежа при изследване на реални финансови проблеми фокусът да се измести от получаването на издържани и практически полезни резултати към самоцелен стремеж за повишаване на $R^2$.

Таблица 11 показва основните особеност и недостатъци, които могат да доведат до непълни или неточни изводи при използването на $R^2$. Особено в случаите, при които няма солидна теоретична основа, с чиято помощ да бъдат подбрани факторите в регресионното уравнение, стремежът да се увеличи стойността на $R^2$ може да доведе до ненужно разширяване на моделите и добавянето на прекалено много обясняващи променливи.

За да се избегне излишното утежняване на модела, се използва модифициран показател за $R^2$, който елиминира втория ефект от Таблица 11.

$$R^2_{\text{мод}} = 1 - \left[\frac{n-1}{n-k-1}(1-R^2)\right]. \qquad (\text{IV—20})$$

*Таблица 11. Особености и недостатъци на $R^2$*

| Особености/недостатъци | Причини |
|---|---|
| Определянето на $R^2$ посредством средната величина на целевата променлива го прави негоден инструмент за сравняване на модели, при които имаме промяна на целевата променлива. | Причината за тази особеност е, че $R^2$ не отразява пряко абсолютната стойност на сумата от квадратите на остатъчните стойности (респ. грешката), а относителния дял на обяснените колебания на целевата променлива спрямо средната ѝ величина. |
| Стойностите на $R^2$ не намаляват (а могат само да останат същите или да нарастват) при добавянето на нови фактори към модела. | Добавянето на нова обясняваща променлива може да доведе до два ефекта:<br>- оценката за нейния коефициент $\beta$ да бъде 0, при което тази нова променлива няма да окаже влияние върху стойността на $R^2$;<br>- оценката за нейния коефициент $\beta$ е различен от нула, при което общата обясняваща сила на модела ще се увеличи (макар и с пренебрежимо малко в най-лошият случай).<br>Тази особеност може да доведе до трудности при определянето на това кои фактори да бъдат включени в регресионната зависимост и дали (и колко) нови да бъдат добавени. |
| При много близки стойности на $R^2$ за различни регресионни зависимости не може да се даде ясен отговор коя от тях обяснява по-добре измененията в целевата променлива. | |

Във формула ( IV—20 ) $n$ е броят на единиците в анализираната съвкупност, а $k$ е броят на факторите в модела. Следователно $R^2_{\text{мод}}$ може да бъде използвано, за да се провери дали включването на допълнителни променливи към модела допринася за повишаване на неговата обясняваща сила, като при това повишението е такова, че не се дължи на случайността (т.е. подобрението е по-силно, отколкото ако би било под въздействието само на случайни фактори).

Много често вниманието се концентрира върху тази силна страна, като се пренебрегва фактът, че именно благодарение на нея стойностите на модифицирания $R^2$ имат различно значение и трябва да бъдат тълкувани различно от тези на обикновения $R^2$. В Таблица 12 са посочени три основни правила, които трябва да отчитаме, когато прилагаме или се позоваваме на $R^2_{\text{мод}}$.

*Таблица 12. Важни моменти при сравнението между $R^2$ и $R^2_{\text{мод}}$*

| | |
|---|---|
| $R^2$ е статистически показател за обясняващата сила на модела. Стойност на показателя от 1 показва, че имаме пълно съвпадение между регресионна права и наличните данни. | $R^2_{\text{мод}}$ е статистически показател, който може да се използва за оценка колко ефективно различни набори от фактори обясняват измененията в целевата променлива. |
| В случаите, когато $R^2$ се използва за оценка на линейна регресия, стойностите на показателя се движат в интервала [0,1]. | Стойностите на $R^2_{\text{мод}}$ са по-малки или равни от стойността на $R^2$. Трябва да отбележим, че $R^2_{\text{мод}}$ може да има и отрицателни стойности. |
| Използването само на $R^2$ може да стимулира необоснованото разширяване на модела, без това да носи реална полза. | Използването на $R^2_{\text{мод}}$ има смисъл да се прилага в случаите, когато анализираме извадка, а не цялата съвкупност. В последния случай резултатът от използването на $R^2$ и $R^2_{\text{мод}}$ ще бъде еднакъв. |

Съществува и разширено тълкуване на $R^2$, което използва концепцията за обясняващата сила на модела, но прилагана и към нелинейни модели. В този контекст са разработени и генерализирани и други показатели за оценка. Интересуващият се читател може да намери подробна информация и сравнение на техните качества в изследването на Schunn, Wallach [10].

## 3. Задачи и въпроси за размисъл

1. Кои са „стандартно“ използваните нива на значимост при проверка на хипотези и в кои случаи е добре да използваме и други нива освен тях? По-високите нива на съгласие дават ли по-голяма сигурност при проверката и ако използваме α=100%, това означава ли, че сме абсолютно сигурни?
2. Какво е особеното при проверката на хипотези с помощта на доверителни интервали?
3. Опитайте се да получите графика на t разпределението с помощта на Excel и задаването на 20 различни комбинации от степени на свобода и вероятност.
4. Винаги ли е по-добре да използваме $R^2_{\text{мод}}$ вместо стандартния $R^2$ като показател за оценка на обясняващата сила на регресионния модел?

# V. Валидност на линейните регресионни модели и метода на най-малките квадрати.

Линейните регресионни модели заедно с метода на най-малките квадрати са предпочитан инструментариум за анализ на редица проблеми не само в областта на финансите, но и на социалните науки изобщо. При това, вперили поглед в резултатите, нерядко пропускаме да отбележим два съществени момента:

- В много от приложенията на линейните регресионни модели въпросът за техните недостатъци и ограничения или отсъства, или не е разгледан задълбочено.
- Неявно се правят допускания, че са изпълнени всички условия, които посочихме в Таблица 4. Ако допусканията не са налице, то това означава, че използваме неправилен инструмент и получените резултати ще бъдат неточни.

И докато в контекста на една научна статия или учебник най-неприятното, което може да се случи, е никой да не ги прочете, то в практическите финансови анализи резултатите ще бъдат съвсем реални загуби или печалби. В тази глава ще разгледаме начините за проверка дали изискванията на линейните регресионни модели са изпълнени и как можем да се защитим от изпадането в заблуда.

Проверката дали са изпълнени условията, посочени в Таблица 4, е свързана с използването на две вероятностни разпределения: $\chi^2$ (хи-/чи- квадрат) и F разпределение (разпределение на Фишер-Снедекор). $\chi^2$ разпределението е свързано с работите на немския математик Хелмерт и английския му колега Пиърсън и представлява разпределението на сумата от квадратите на $k$ обясняващи променливи, всяка от които следва стандартно нормално разпределение[12]. Броят на променливите $k$ е основна характеристика на $\chi^2$ и също така се обозначава като степен на свобода на разпределението. Плътността на вероятностното разпределение на $\chi^2$ е дефинирана, както следва:

$$f(x,k) = \begin{cases} \dfrac{x^{\frac{k}{2}-1} e^{\frac{-x}{2}}}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} & \text{за } x \geq 0 \\ 0 & \text{за } \mathrm{x} < 0 \end{cases}, \qquad \text{( V—1 )}$$

където $\Gamma(\frac{k}{2})$ е Гама функция, която се дефинира като

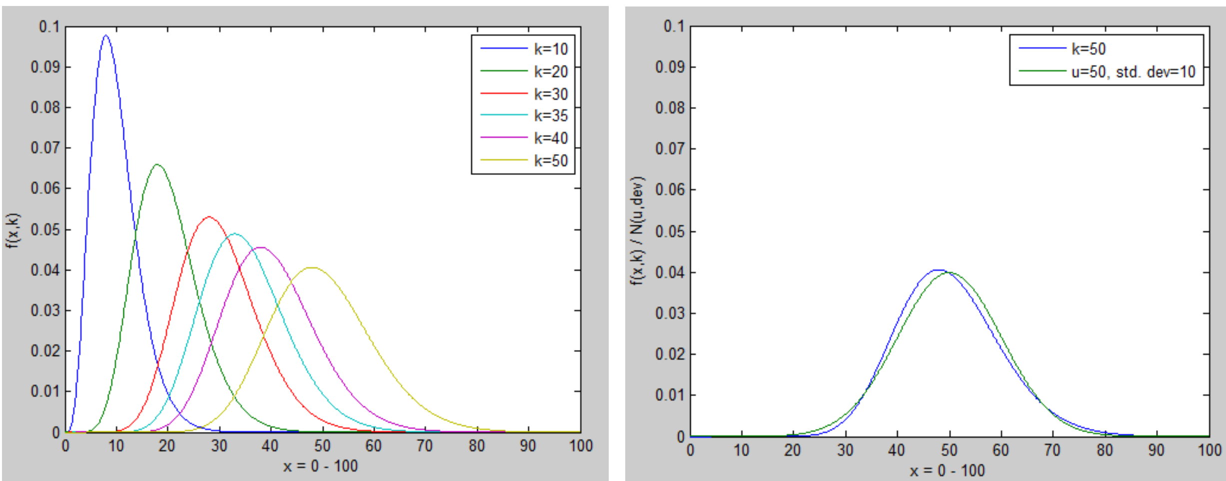
$$\Gamma(\alpha) = \int_0^\infty x^{a-1} e^{-x}\, dx\,, a > 0 \qquad \text{( V—2 )}$$

[12] Ще припомним, че стандартно се нарича такова нормално разпределение, за което средната стойност е 0, а стандартното отклонение е 1.

и има следните три основни свойства:

$$
\begin{aligned}
&\Gamma(0) = 1 \\
&\Gamma(a) = (a-1)\Gamma(a-1), \text{за } \forall\, a \\
&\Gamma(a+1) = a! \text{ за } a \text{ цяло число}
\end{aligned}
\qquad (\text{V—3})
$$

С нарастване на степените на свобода $k$, $\chi^2$ разпределението се доближава до нормално разпределение $N(\mu = k, \sigma = \sqrt{2k})$, като за практически цели се приема, че при *k > 50* полученото разпределение е достатъчно близо до стандартното нормално разпределение (т.е. ако една случайна величина X има $\chi^2(k)$ разпределение, то с нарастването на $k$ разпределението на $\frac{X-k}{\sqrt{2k}}$ ще клони към стандартно нормално N(0,1) ).




*Фигура 10. Графично представяне на $\chi^2$ разпределението*

Както се вижда от Фигура 10, при степени на свобода, равни или по-големи от 50, $\chi^2$ разпределението наистина е близо до нормално разпределение N(50,10). В рамките на този текст ние няма да представяме в дълбочина характеристиките на $\chi^2$ разпределението, но е необходимо да познаваме поне основните му свойства и как изглежда то.

F разпределението се свързва с имената на английския статистик Роналд Фишер и американския математик и негов колега Джордж Снедекор. Поради факта, че то заема важно място като инструмент за статистическа проверка на хипотези, е необходимо да познаваме неговия вид и основни характеристики. За разлика от $\chi^2$ разпределение при F разпределението се налага да работим с два параметъра, обикновено обозначавани като $d_1$и $d_2$, и функция на вероятностна плътност ( V—4 ):

$$f(x, d_1, d_2) = \begin{cases} \dfrac{\sqrt{\dfrac{(d_1 x)^{d_1} d_2^{d_1}}{(d_1 x + d_2)^{d_1 + d_2}}}}{xB\left(\dfrac{d_1}{2}, \dfrac{d_2}{2}\right)} & \text{за } x > 0 \\ 0 & \text{за } x \leq 0 \end{cases} \qquad (\text{V—4})$$

където $B(\frac{d_1}{2}, \frac{d_2}{2})$ е Бета функция, която може да се представи като:

$$\mathrm{B}(p, q) = \frac{(p-1)!(q-1)!}{(p+q-1)!} = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}, \qquad (\text{V—5})$$

а Γ(α) е вече използваната при $\chi^2$ разпределението гама-функция. Освен използването на гама-функция съществува още една важна характеристика, която предопределя факта, че обикновено $\chi^2$ и F разпределенията вървят ръка за ръка и се използват за статистическа проверка на хипотези. Тази особеност е, че една случайна променлива, следваща F

## Как мога да получа графики, подобни на фигура 10?

Фигура 10 може да получим дори и с помощта на Excel. По-лесно е, ако разполагаме със специализиран софтуер като Matlab или Maple. Командите (кодът) на Matlab с чиято помощ е получена фигурата, е:

```
figure
int=0:0.1:20;
pdf1=chi2pdf(int, 1);  % chi2pdf връща стойността на вероятностната плътност за интервала int и k=1
pdf2=chi2pdf(int, 2);
pdf3=chi2pdf(int, 3);
pdf6=chi2pdf(int, 6);
pdf9=chi2pdf(int, 9);
pdf50=chi2pdf(int, 50);
plot(int, pdf1, int, pdf2, int, pdf3, int, pdf6, int, pdf9, int, pdf50 );
hleg = legend('k=1','k=2','k=3','k=6','k=9','k=50');
xlabel('x = 0 - 20');
ylabel('f(x,k)');
int2=0:0.1:100;
pdf50=chi2pdf(int2, 50);
norm50=normpdf(int2, 50, 10);
plot(int2, pdf50, int2, norm50);
hleg = legend('k=50','u=50, std. dev=10');
xlabel('x = 0 - 200');
ylabel('f(x,k) / N(u,dev)');
```

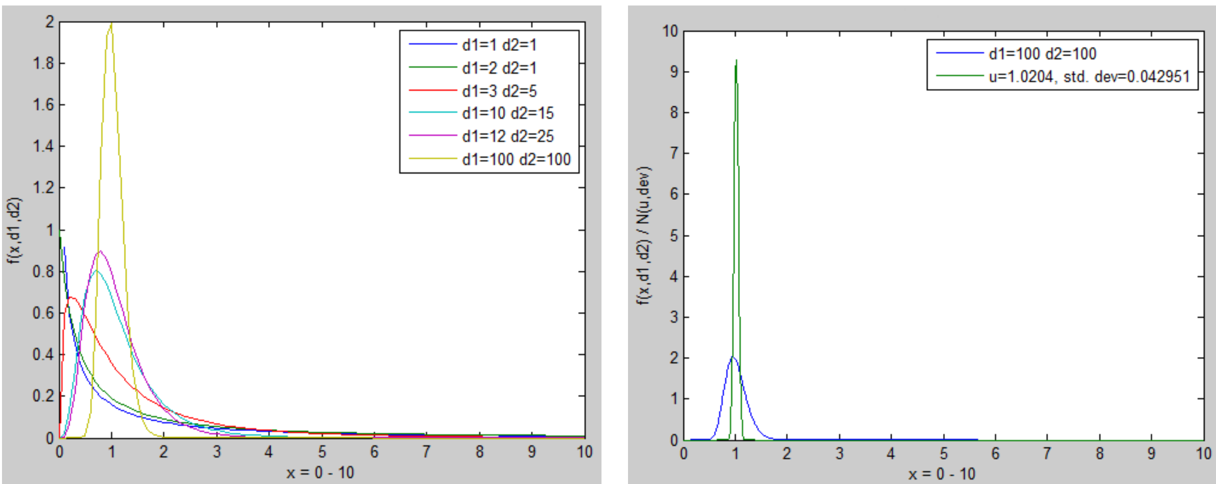
## CHISQ.DIST(x, ст. на свобода, кумулативна ф-я)

Тази функция на Excel може да се използва, за да изчислим стойността на кумулативната функция на χ2 за желана стойност *x* и степени на свобода *k*.

разпределение, може да бъде представена като отношение на две други обясняващи променливи с $\chi^2$ разпределения и степени на свобода съответно $d_1$и $d_2$:

$$X = \frac{\frac{R_1}{d_1}}{\frac{R_2}{d_2}} \text{ и } X \sim F(d_1, d_2). \qquad (\text{V—6})$$

Фигура 11 дава представа как изглежда F разпределението при различни комбинации от степени на свобода $d_1$ и $d_2$. Трябва да отбележим, че за разлика от $\chi^2$ разпределението при големи стойности на $d_1$ и $d_2$ разликите между F разпределението и нормалното се запазват.




*Фигура 11. Графично представяне на F разпределението*

Внимателният читател вероятно се е запитал защо в дясната част на фигури 10 и 11 поставяме нормално разпределение с точно тези средна стойност и стандартно отклонение. Отговорът се крие в

Таблица 13, която представя не само функциите на вероятностна плътност, но и средните и вариацията на двете разпределения.

Освен за представените методи за проверка на хипотези, $\chi^2$ и F разпределенията намират и други приложения във финансите. Така например F разпределението се използва при анализа на вариацията (т. нар. ANOVA методи), който пак може да бъде сведен до статистически тест, но с различни допускания и различни практически цели.

## Как мога да получа графики, подобни на фигура 11?

Фигура 11 е получена с помощта на Matlab. Командите (кодът) на Matlab с чиято помощ са изчертани графиките, е:

```
Figure
int=0:0.1:10;
pdf11=fpdf(int, 1,1);
pdf21=fpdf(int, 2,1);
pdf35=fpdf(int, 3,5);
pdf1015=fpdf(int, 10,15);
pdf1225=fpdf(int, 12,25);
pdf100100=fpdf(int, 100, 100);
plot(int, pdf11, int, pdf21, int, pdf35, int, pdf1015, int, pdf1225, int, pdf100100 );
hleg = legend('d1=1 d2=1','d1=2 d2=1','d1=3 d2=5','d1=10 d2=15','d1=12 d2=25','d1=100 d2=100');
xlabel('x = 0 - 10');
ylabel('f(x,d1,d2)');
int2=0:0.01:10;
pdf100100=fpdf(int2, 100, 100);
norm100=normpdf(int2, 1.0204, 0.042951);
plot(int2, pdf100100, int2, norm100);
hleg = legend('d1=100 d2=100','u=1.0204, std. dev=0.042951');
xlabel('x = 0 - 10');
ylabel('f(x,d1,d2) / N(u,dev)');
```

## F.DIST(x, ст. на свобода, ст. на свобода, кумулативна ф-я)

Тази функция на Excel може да се използва за да изчислим стойността на кумулативната функция на F разпределението за желана стойност x и степени на свобода съответно $d_1$ и $d_2$.

*Таблица 13. Основни характеристики на $\chi^2$ и F разпределенията*

| $\chi^2$ разпределение | F разпределение |
|---|---|
| $f(x,k)=\begin{cases}\dfrac{x^{\frac{k}{2}-1}e^{\frac{-x}{2}}}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} & \text{за } x\geq 0\\ 0 & \text{за x} < 0\end{cases}$ | $f(x,d_1,d_2)=\begin{cases}\dfrac{\sqrt{\dfrac{(d_1x)^{d_1}d_2^{d_1}}{(d_1x+d_2)^{d_1+d_2}}}}{xB(\frac{d_1}{2},\frac{d_2}{2})} & \text{за } x>0\\ 0 & \text{за } x\leq 0\end{cases}$ |
| средна стойност $=k$ | средна стойност $=\dfrac{d_2}{d_2-2}, d_2>2$ |
| вариация $=2k$ | вариация $=\dfrac{2d_2^2(d_1+d_2-2)}{d_1(d_2-2)(d_2-4)}, d_2>4$ |
| мода $=\max(k-2,0)$ | мода $=\dfrac{(d_1-2)}{d_1}\dfrac{d_2}{(d_2+2)}$ |

## 1. Проверка за нулева средна стойност на грешката

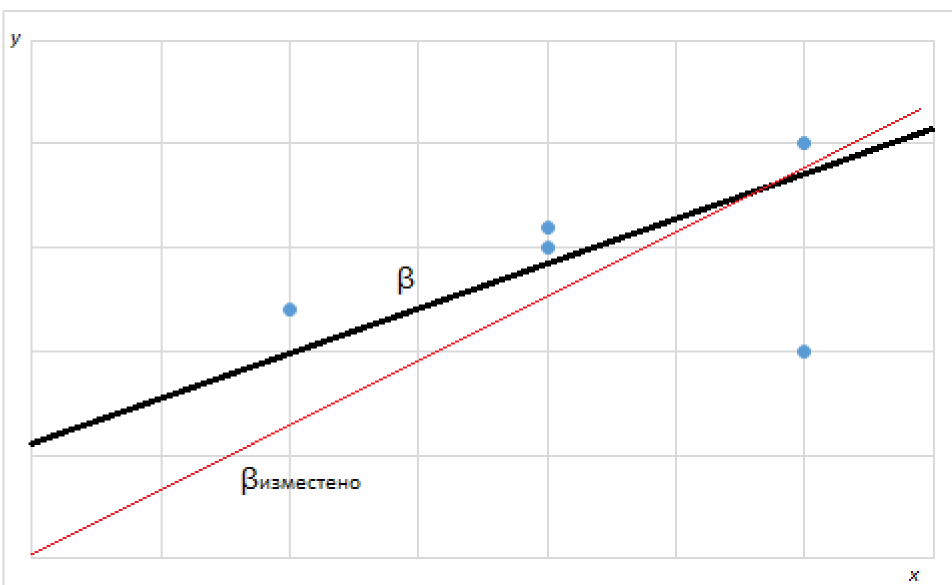
За да преценим значението на това допускане, нека припомним отново използвания общ вид на линейния регресионен модел:

$$y_i = \alpha + \beta_1 x_{i1} + \ \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i, \qquad i = \overline{1,n}\,, \tag{V—7}$$

или в случай на една обясняваща променлива:

$$y_i = \alpha + \beta x_i + \varepsilon_i\,. \tag{V—8}$$

Ако условието за нулева средна стойност на грешката не е изпълнено, то тогава, както дискутирахме в Таблица 4, получените оценки на параметрите $\alpha$ и $\beta$ ще бъдат неточни или казано по друг начин – „изместени“. Доколко опасно е това изместване и дали може да сме сигурни, че в получените резултати няма да са заложени изкривени и неотговарящи на нашите цели стойности? В случаите, когато изграждания финансов модел предполага стойност на $\alpha$, която да е различна от нула, то ефектът от ненулева средна стойност на грешката ($\mathrm{E}(\varepsilon) \neq 0$) ще доведе до друга стойност $\alpha'$ на параметъра $\alpha$. В случай че просто прилагаме метода на най-малките квадрати, ние ще получим стойността $\alpha'$ и ако във финансовия модел няма допълнителни ограничения, този резултат ще съответства идеално на изискванията – ще получим константа, но няма как да знаем (освен ако финансовият модел не ни подскаже) дали тази константа е изместена, или не.




*Фигура 12. Ефект от ненулева средна стойност на грешката и изискване за α=0*

По-различна е ситуацията в случаите, когато икономическата логика изисква стойността на константата в модела да бъде нула, или казано по друг начин − правата да минава през началото на нашата координатна система. В този случай ефектът от ненулева средна стойност на грешката, комбиниран с изискването $\alpha' = \alpha = 0$, ще доведе до получаване на изместена оценка за наклона на правата – т.е. коефициента $\beta$.

На Фигура 12 е показан ефектът от подобна ситуация, при което червената и по-тънка права линия показва изместената оценка за $\beta$. Некоректните резултати не само водят до неточности при използването на модела, но могат и да бъдат придружени от изненадващи оценки за обясняващата сила на регресионното уравнение. Така например поради изместената оценка за $\beta$ може да се окаже, че $R^2$ е отрицателна величина! Това става възможно, тъй като по дефиниция $R^2$ е свързан с това, каква част от измененията в стойностите на целевата променлива успява да обясни моделът, а резултатите в контекста на „изместените“ оценки може да се окаже, че обясняват дори по-големи изменения от реално наблюдаваните[13].

При прилагането на линейните регресионни модели за решаване на икономически и финансови проблемни ситуации изискването за нулева стойност на константата $\alpha$ се среща много често. Причините за това могат да бъдат два типа:

- необходимост от нулева стойност поради теоретичните изисквания на използвания икономически или финансов модел

Съществуват редица модели, при които изначално е необходимо константата $\alpha$ в линейната зависимост да бъде нула или просто ненулевите стойности нямат логично обяснение. Така например, ако изследваме производствената функция на Коб-Дъглас и логаритмуваме, ще получим линейна зависимост за коефициентите на еластичност на труда и капитала:

$$Y = L^a K^b \text{ и } \ln Y = alnL + blnK \,. \qquad ( V—9 )$$

В този модел с две променливи няма константа, което съответства на стойност на константата нула. Разбира се, възможно е да се въведе изкуствено ненулева константа в модела но това ще означава, че ние дефинираме състояние, при което с никакъв капитал и труд можем да постигнем определено ненулево ниво на производство.

- необходимост, възникваща вследствие като резултат от прилагането на трансформации върху първоначалните данни

[13] В част IV ние коментирахме, че математически можем да обясним защо стойностите на $R^2$ се намират в интервала [0,1]. Обяснението тук не влиза в противоречие с това твърдение, тъй като ние коментираме ситуация, при която условията за валидност на линейните регресионни модели са нарушени.

Няма да разглеждаме всички възможни преобразувания, които могат да доведат до линеен модел с нулева константа. Но ще обърнем внимание на едно от тях, което се използва много често и води точно до анализираната ситуация. В много изследвания акцентът се поставя не върху абсолютните стойности на целевата променлива, а на това, как тя се променя във времето. Когато тези промени се разглеждат в абсолютна стойност и ние имаме линеен модел, ще получим:

$$y_t - y_{t-1} = \alpha + \beta x_t - (\alpha + \beta x_{t-1}) = \beta(x_t - x_{t-1}) \,, \qquad \text{( V—10 )}$$

което отново води до модел с нулева константа и необходимост да проверим изрично дали $E(\varepsilon) = 0$.

## 2. Проверка за константна и крайна стойност на вариацията на грешката

Допускането, че вариацията на грешките е константна (и крайна) величина е известно още като хомоскедастицитет (а неговата противоположност − като хетероскедастицитет). Проверката за наличието му е важна поради следните причини:

*Първата причина е, за да сме сигурни, че използваме правилния метод за оценка на параметрите/коефициентите на регресията.*

От техническа гледна точка няма причина да не използваме определен метод (например най-малките квадрати) дори и когато вариацията на грешките не е константа. Но в този случай получените резултати няма да бъдат максимално точни. Под „максимално точни“ тук имаме предвид, че те няма да минимизират колебанията в оценките на регресионните параметри. В коментара към точността на линейната регресия отбелязахме, че границите в които се изменя параметърът β (ако разглеждаме модел с една обясняваща променлива), са:

$$\sigma_\beta = \sigma_\varepsilon \sqrt{\frac{1}{\sum(x_i - \bar{x})^2}} \,. \qquad \text{( V—11 )}$$

Но това е валидно само в случай че е изпълнено условието за хомоскедастицитет. В противен случай, както отбелязват Hill, Griffiths, Guay [11], изразът, който определя границите на измененията на β, има вида:

$$\sigma_\beta{}' = \sqrt{\sum_{i=1}^{N} \frac{(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} \sigma_i^2},$$

където $\sigma_i^2 = var(\varepsilon_i)$ е израз на промените във вариацията на грешката.

Ако предварително е известно, че вариацията на грешките се променя, то тогава може да се използват специални методи за оценка на параметрите на регресията, които да елиминират ефекта на хетероскедастицитет и да подобрят точността на получените стойности. Има две основни форми на зависимост на вариацията на грешките, с които се сблъскваме и които трябва да бъдат коригирани:

- Когато е налице зависимост между вариацията на грешките и определена променлива величина

Важно е да имаме предвид, че в подобна ситуация променливата, свързана с вариацията на грешките, не е задължително да бъде един от факторите в регресията. Тя може да бъде и външна за нашия модел, но това не я прави по-малко важна. В подобни случаи използването на метода на най-малките квадрати ще ни позволи да получим неизместена оценка на параметрите на регресията, но стойността на стандартната грешка ще бъде неправилна. При това отклоненията в стандартната грешка ще зависят от начина, по който вариацията на грешките се изменя, или казано иначе – от формата на хетероскедастицитет.

Ако предварително е известно каква е връзката между вариацията на грешките и дадена променлива, то тогава можем да използваме модификация на метода на най-малките квадрати, за да избегнем проблемите, свързани със стандартната грешка. Линейната регресионна зависимост, когато имаме хомоскедастицитет, е:

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i\,,\ cov(\varepsilon) = const\,. \qquad (\text{V—12})$$

Ако предположим, че измененията на грешките зависят от променлива $v$ (която може да е, но може и да не е част от факторите $x_1, x_2, \dots, x_k$), така че $cov(\varepsilon) = const.v^2$, то можем да преобразуваме регресионната зависимост, като разделим двете страни на уравнението на $v$:

$$\frac{y_i}{v_i} = \frac{\alpha}{v_i} + \beta_1 \frac{x_{i1}}{v_i} + \beta_2 \frac{x_{i2}}{v_i} + \cdots + \beta_k \frac{x_{ik}}{v_i} + \frac{\varepsilon_i}{v_i}\,,\ cov\left(\frac{\varepsilon}{v_i}\right) = \frac{const.{v_i}^2}{{v_i}^2} = const\,. \qquad (\text{V—13})$$

Получената нова регресионна зависимост е „претеглена“ спрямо стойностите на променливата $v$, откъдето идва и името на този подход – метод на претеглените най-малки квадрати. За разлика от първоначалната форма на уравнението тук няма константа (тъй като първият елемент също е претеглен спрямо $v - \frac{\alpha}{v_i}$), което означава, че и тук са в сила аргументите, дискутирани в предната част. Предимството на претеглените най-малки квадрати е, че по този начин се елиминира ефектът върху стандартната грешка, свързан с хетероскедастицитета. Основният проблем при използването на този подход е, че предварително трябва да е известна зависимостта между вариацията на грешките и променливата $v$. Ако тази връзка не е известна или е направена на базата на опростяващи допускания, то използването на претеглените най-малки квадрати ще елиминира един източник на неточност за сметка на друг (т.е. опростеното допускане за $cov(\varepsilon) = f(v)$ ).

- Когато вариацията на грешките се променя с времето

В този случай се налага използването на група методи, известни в статистиката като ARCH (AutoRegressive Conditional Heteroskedasticity). Без да навлизаме в подробности, основна характеристика на тези методи е предположението, че е налице зависимост между грешките през отделните времеви периоди.

*Втората важна причина е, за да се уверим, че получените оценки на параметрите на регресията могат да бъдат тествани с вече разгледаните методи за проверка на хипотези.*

Ако допускането за хомоскедастицитет е нарушено, както видяхме, това няма да се изрази в изместване на оценките за параметрите на регресионната зависимост. Но то ще повлияе на границите, в които те се колебаят. Поради тази причина, ако използваме метода на най-малките квадрати в условията на хетероскедастицитет, можем да бъдем подведени на последващия етап от анализа – при използването например на t-тестове за проверка на валидността и значимостта на получените оценки.

За щастие най-сложната част при проверката за наличие на хомоскедастицитет е правилното изговаряне на термина. Съществуват много методи за проверка дали вариацията на грешките остава постоянна, като те в крайна сметка се свеждат до статистическа проверка на хипотези – разликата е в конкретния тест за проверка и как точно е формулирана хипотезата. Ще представим един от възможните подходи – теста на White [12]. За да направим проверката за хомоскедастицитет, използваме вече познатата ни регресионна зависимост:

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \,, \qquad (\text{V—14})$$

както и факта, че вариацията на грешките е $var(\varepsilon) = E[(\varepsilon_i - E(\varepsilon_i))^2]$. Ако обаче е изпълнено първото условие за адекватност на линейния регресионен модел, то би трябвало средната стойност на грешките да бъде равна на нула – т.е. $E(\varepsilon_i) = 0$, откъдето ще получим, че $var(\varepsilon) = E[{\varepsilon_i}^2]$. Тази зависимост White използва, за да представи тествана хипотеза с помощта на второ регресионно уравнение (за прегледност е посочено как изглежда то в случай на 2 обясняващи променливи):

$${\varepsilon_i}^2 = \gamma_1 + \gamma_2 x_{i1} + \gamma_3 x_{i2} + \gamma_4 {x_{i1}}^2 + \gamma_5 {x_{i2}}^2 + \gamma_6 x_{i1} x_{i2} + v_i \,. \qquad (\text{V—15})$$

Ако вариацията на грешката е константа, би следвало всички параметри на втората регресионна зависимост с изключение $\gamma_1$ да бъдат нулеви. Причината да няма изискване за нулева стойност на константата е свързана с това, че хомоскедастицитет означава константна вариация, но не и нулева такава. Т.е. фактът че не е нужно $\gamma_1$ да е нула, съответства на това, че все пак може да има колебания в стойностите на грешката (ще припомним, че първото условие, което разгледахме, се отнася до нулева средна стойност на грешките, а не до нулева стойност за всяка грешка).

С въведената от White втора регресионна зависимост проверката за хомоскедастицитет се свежда до проверка на хипотезата, че $\gamma_2 = \gamma_3 = \cdots = \gamma_l = 0$. Приемането или отхвърлянето на тази хипотеза може да стане посредством използването на стандартен F-тест, който вече дискутирахме. За целта обаче е необходимо да въведем трета регресионна зависимост (т.нар. ограничена регресия), в която да фигурира само константата $\gamma_1$.

### Примерен тест за хомоскедастицитет

Приемете, че сте изградили линеен регресионен модел за това, как зависи размерът на заплатата от образованието и броя на говорените чужди езици. Оценката на параметрите на регресията е базирана на проучване, включващо 90 участници. Проверете дали вариацията на грешките е константа, ако е известно, че показателят $R^2$ за регресионната зависимост свързваща грешките и факторите в модела, е 0.25

### CHIINV(вероятност, степени на свобода)

$TR^2 = 90*0.25 = 22.5$
$\chi^2(5)$ за ниво на съгласие $\alpha$=0.05 (5%) е CHIINV(0.05, 5) = 11.07049.
Тъй като $TR^2 > \chi^2(5)$ (22.5 > 11.07049), то тогава можем да отхвърлим нулевата хипотеза за наличие на хомоскедастицитет.

Съществува и по-кратък начин за проверка с помощта на т.нар. множители на Лагранж и показателя за значимост на втората регресионна зависимост:

$$nR^2 \sim \chi^2(l), \qquad \text{( V—16 )}$$

където $n$ е броят на единиците в анализираната съвкупност, $R^2$ е показателят за обясняваща сила на втората регресионна зависимост, а $l$ е броят на събираемите в регресионната зависимост, без да броим константата (т.е. $l$ е броят на всички събираеми в уравнението минус единица). По този начин окончателната проверка ще изглежда, както следва:

Ако $nR^2 < \chi^2(l)$ при определено ниво на съгласие, то тогава не можем да отхвърлим нулевата хипотеза, че $\gamma_2 = \gamma_3 = \cdots = \gamma_l = 0$.

Ако $nR^2 > \chi^2(l)$ при определено ниво на съгласие, то тогава можем да отхвърлим нулевата хипотеза, че $\gamma_2 = \gamma_3 = \cdots = \gamma_l = 0$, и следователно не е изпълнено условието за хомоскедастицитет.

## 3. Проверка за независимост на грешките (остатъчните стойности)

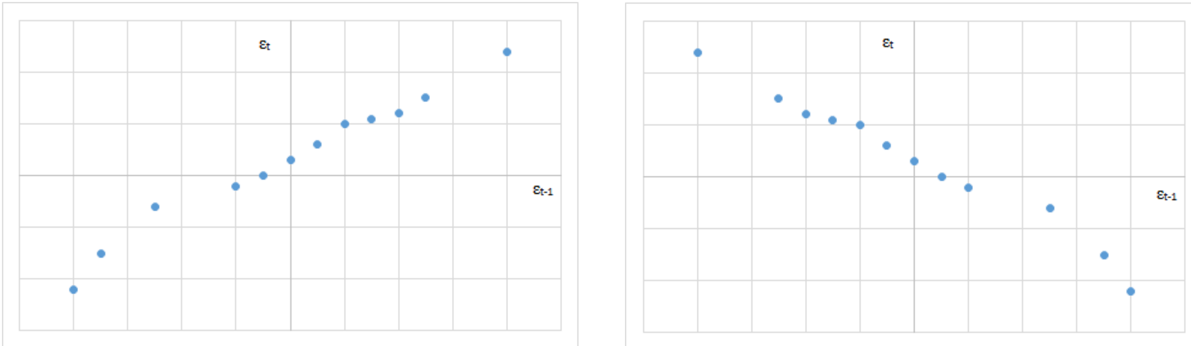
Независимост на грешките на практика означава допускането, че различните остатъчни стойности не са корелирани – т.е. $cov(\varepsilon_i, \varepsilon_j) = 0$. Финансовото тълкуване на липсата на зависимост е различно в зависимост от приложението на регресионните модели. Един пример е, когато се анализират данни от времеви ред и ние предполагаме, че остатъчните стойности за всеки период са независими от тези стойности от предходните периоди.

Анализ на времеви редове и динамичните модели са предмет на специално внимание по-нататък в изложението, като засега е важно да споменем, че независимостта може да бъде отнасяна както спрямо отделни времеви периоди, така и в рамките на статични модели за отделните наблюдения от анализираната съвкупност и извадка. Когато допускането за независимост на грешките е необосновано, казваме, че е налице автокорелация или серийна корелация и това налага определени ограничения при използването на линейните регресионни модели и тълкуването на техните резултати. Крайните ефекти от липсата на независимост между остатъчните стойности на регресията са близки до тези, разглеждани в предния параграф – когато имаме хетероскедастицитет. Автокорелацията води до получаването на некоректни стойности за стандартната грешка, откъдето се увеличава вероятността за грешки от I и II род при използване на регресионния модел и статистическата проверка на хипотези. Наличието на ненулева ковариация между отделните остатъчни стойности също така компрометира изчисленията за обясняващата сила на изградения модел и може да доведе до прекалено оптимистичното приемане на получените резултати. Всичко това налага от − една страна − да направим проверка дали допускането е налице при всяко едно конкретно изследване, а от друга − да предприемем действия, за да коригираме негативните ефекти на автокорелацията, когато тя обективно е налице.

Съществуват различни методи за проверка за независимост на грешките в линейните регресионни модели. Ние ще посочим само два от тях поради широката им приложимост и интуитивната дефиниция:

- графичен метод за проверка за автокорелация

Този подход е най-малко обременяващ откъм необходимостта от изчисления, но от друга страна е и най-податлив на субективна преценка. Същността на метода е да представим графично отделните остатъчни стойности и да проверим дали е налице определен шаблон или зависимост между тях. Фигура 13 представя един от най-простите възможни случаи на автокорелация, при който анализираме един времеви ред и остатъчната стойност във всеки един период зависи от грешката в предния.




*Фигура 13. Пример за положителна и отрицателна автокорелация*

Дясната графика представя положителна зависимост между остатъчните стойности в два последователни периода, което означава, че с нарастване на грешката през период t-1, нараства и грешката през период t. Лявата графика представя точно огледалната зависимост – т.е. отрицателна автокорелация. В случаите, когато имаме ясно изразена зависимост този метод ни дава бърз и удобен отговор на въпроса дали допускането за $cov(\varepsilon_i, \varepsilon_j) = 0$ е нарушено. За съжаление, графичният метод невинаги е сполучливо решение поради следните причини:

- когато анализираме много данни, точковата диаграма става трудна за четене и анализ;
- ако трябва да изразим едновременна проверка за зависимост между няколко грешки, то графиката няма да бъде двумерна (а хиперравнина) или ще трябва да работим с много последователни графики;
- освен отговора за наличието/липсата на зависимост понякога се налага да изследваме и силата и посоката на тази зависимост. А това е нещо, което графичният подход не може да предостави като информация.

Поради изброените недостатъци при проверката на това допускане за валидността на линейните регресионни модели е добре да използваме аналитични методи, които да предоставят конкретни числени показатели. Един такъв подход ни дава тестът на Breusch-Godfrey-Bertolo [13] [14], който също може да бъде причислен към техниките, използващи статистическа проверка на хипотези, и позволява да бъде проверено наличието на зависимост между остатъчни стойности с желана от нас „дълбочина“. Ако се върнем отново към примерните данни от Фигура 13, това означава да проверим за връзка не само между $\varepsilon_t$ и $\varepsilon_{t-1}$, а да определим $p$ $(2 \leq p \leq k)$ на брой стъпки и да направим тест за остатъчните стойности $\{\varepsilon_t, \varepsilon_{t-1}, \dots, \varepsilon_{t-p}\}$. Ако липсва зависимост между тези $p$ елемента, то тогава би следвало регресията с тяхно участие да съдържа само нулеви коефициенти:

$$\varepsilon_t = \gamma_1\varepsilon_{t-1} + \gamma_2\varepsilon_{t-2} + \cdots + \gamma_k\varepsilon_{t-p} + v_t \text{, където } v_t \sim N(0, \sigma_v) \qquad (\text{V—17})$$

По този начин проверката за липса на автокорелация се превръща в проверка за валидността на нулева хипотеза от типа $\gamma_1 = \gamma_2 = \cdots \gamma_p = 0$. Самата проверка може да направим отново с множителите на Лагранж и показателя за значимост на регресионното уравнение за $\varepsilon_t$:

$$(n-p)R^2 \sim \chi^2(p)\,. \tag{V—18}$$

За разлика от проверката за хомоскедастицитет тук е необходимо да използваме множител $n-p$, тъй като $p$ на брой от наблюденията участват в регресионната зависимост ($\{\varepsilon_t, \varepsilon_{t-1}, \dots, \varepsilon_{t-p}\}$). Последният етап от проверката е сравняване на стойността от теоретичното $\chi^2$ разпределение с $p$ степени на свобода и желано ниво на съгласие и стойността $(n-p)R^2$, за да преценим дали може да отхвърлим нулевата хипотеза.

Необходимо е да припомним обаче, че тази проверка за липса на взаимна връзка между грешките в регресията се отнася само за определената „дълбочина" $p$. Това оставя отворен въпроса как да бъде определена стойността на $p$, като отговорът на този въпрос зависи както от размера на анализираната съвкупност и извадка, така и от теоретичните основи, върху които се изгражда регресионната зависимост. Например, ако финансовата теория предполага, че е налице зависимост само от предния период, достатъчно е да бъде направена проверка с „дълбочина" 1. В случаите, когато съществуват предположения за по-дългосрочни зависимости (например при изследване на процеси с изразен сезонен характер), е необходимо да анализираме по-голям обем от данни и да използваме по-големи стойности на $p$.

## 4. Проверка за зависимост между грешката и факторите

Проверката за връзка между остатъчните стойности и факторите е важна, за да бъдем сигурни, че получените оценки за параметрите на регресията не са изместени. В случай че това условие не е изпълнено, получените резултати могат да ни въведат в заблуждение и да надценим влиянието на част от факторите. За да си обясним защо това е така, нека предположим, че имаме регресионното уравнение:

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i\,, \tag{V—19}$$

като при това нека имаме зависимост между $x_1$ и $\varepsilon$ и $cov(x_1, \varepsilon) > 0$. Това означава, че ако $\varepsilon_i$ нараства, и стойностите едновременно на $x_{1i}$ и на $y_i$ ще нарастват. В случай че знаем предварително за връзката между грешката и променлива-фактор, можем да предвидим този ефект и да се опитаме да го неутрализираме. Ако обаче не ни е известно (или без основание сме допуснали, че няма зависимост), то тогава бихме могли да изтълкуваме увеличената стойност на $y_i$ като ефект от увеличената стойност на $x_{i1}$. Подобно тълкуване ще доведе до неправилно засилване на значимостта на фактора $x_1$ и може да стане причина за вземането на погрешни финансови решения.

Един от методите за проверка дали остатъчните стойности и факторите се намират във връзка е да използваме вече разгледаните подходи за тестване за хомоскедастицитет. Разликата в този случай ще бъде, че нулевата хипотеза ще бъде базирана на коефициентите на регресионно уравнение, при което целевата променлива е $\varepsilon_i$, а факторите съвпадат с тези от първоначалния модел.

## 5. Проверка дали грешките следват нормално разпределение

Допускането, че остатъчните стойности на регресията следват нормално разпределение, е последното от изискванията в Таблица 6, но то далеч не е маловажно. Това предположение не бива да оставя у нас впечатление, че всеки един от факторите $x_i$ в регресионното уравнение ( V—14 ) трябва да бъде нормално разпределен. Допускането се отнася само до остатъчните стойности – $\varepsilon_i$.

Допускането за нормално разпределение на остатъчните стойности е особено важно, когато работим с малки по обем съвкупности и извадки, тъй като нарушаването му в този случай ще доведе до грешки при проверката на хипотези и оценяването на получените резултати. Когато пък анализираме достатъчно големи по обем извадки, е по-лесно да се позовем на изпълнимостта на това допускане благодарение на централната гранична теорема.

Съществуват различни подходи при проверка за нормалност. Някои от тях са свързани с проверка до каква степен реално наблюдаваното разпределение се различава от предварително зададено теоретично такова. Други методи се използват не само за следене на отклонения от нормалното разпределение, а и за оценка на параметрите на това статистическо разпределение, което е най-близо до наблюдаваните данни. Ще анализираме две от използваните техники:

- Anderson-Darling [15] тестовете могат да се използват не само за следене на отклонения от нормалното разпределение, а и за оценка на параметрите на това статистическо разпределение, което е най-близо до наблюдаваните данни. За целта се изчислява разликата между реално наблюдаваните данни (т.е. разпределението на грешките) и теоретичното разпределение:

$$D_{A-D} = n \int_{-\infty}^{+\infty} \frac{(F_\varepsilon(x) - F_t(x))^2}{F_t(x)(1 - F_t(x))} dF_t(x), \qquad ( V—20 )$$

където $F_\varepsilon(x)$ е кумулативната вероятностна функция на остатъчните стойности, а $F_t(x)$ е кумулативната функция на теоретичното разпределение.

- Тестът на Jarque-Bera е основан на проверка за наличие на асиметрия (измерена с третия централен момент/коефициента на асиметрия) и ексцес (измерени с четвъртия централен момент).

За да видим каква е същността на теста на Jarque-Bera [16], нека припомним как се дефинират коефициентът на асиметрия и коефициентът на ексцес:

$$\begin{aligned} &\text{коеф. на асиметрия} = E\left[(\frac{x-\bar{x}}{\sigma})^3\right] = \frac{\mathrm{E}[(x-\bar{x})^3]}{\sigma^3} = \frac{\mathrm{E}[(x-\bar{x})^3]}{(E[(x-\bar{x})^2])^{\frac{3}{2}}}, \\ &\text{коеф. на ексцес} = E\left[(\frac{x-\bar{x}}{\sigma})^4\right] = \frac{\mathrm{E}[(x-\bar{x})^4]}{\sigma^4} = \frac{\mathrm{E}[(x-\bar{x})^4]}{(E[(x-\bar{x})^2])^2}. \end{aligned} \qquad (\text{V—21})$$

Дефиницията и на двата коефициента идва от определението съответно на трети и четвърти централен момент. Нека сега си представим, че вместо абстрактната променлива $x$, която сме използвали в горните уравнения, отнесем коефициентите на асиметрия и ексцес към остатъчните стойности на регресията. В този контекст освен трети и четвърти централен момент ще имаме и още една важна особеност, а именно че очакваната или средната стойност на грешките трябва да бъде равна на нула (в случай че е в сила първото от разглежданите от нас допускания):

$$\begin{aligned} &\text{коеф. на} \\ &\text{асиметрия остатъчни с}-\text{ти} = E\left[\left(\frac{\varepsilon-\bar{\varepsilon}}{\sigma_\varepsilon}\right)^3\right] = \frac{\mathrm{E}[(\varepsilon-\bar{\varepsilon})^3]}{{\sigma_\varepsilon}^3} = \frac{\mathrm{E}[(\varepsilon-\bar{\varepsilon})^3]}{(E[(\varepsilon-\bar{\varepsilon})^2])^{\frac{3}{2}}} \end{aligned} \qquad (\text{V—22})$$

$$\begin{aligned} &\text{коеф. на} \\ &\text{ексцес остатъчни с}-\text{ти} = E\left[(\frac{\varepsilon-\bar{\varepsilon}}{\sigma_\varepsilon})^4\right] = \frac{\mathrm{E}[(\varepsilon-\varepsilon)^4]}{{\sigma_\varepsilon}^4} = \frac{\mathrm{E}[(\varepsilon-\bar{\varepsilon})^4]}{(E[(\varepsilon-\bar{\varepsilon})^2])^2} \end{aligned} \qquad (\text{V—23})$$

$$\mathrm{E}(\varepsilon) = \bar{\varepsilon} = 0\,. \qquad (\text{V—24})$$

Като отчетем, че средната стойност на остатъчните стойности трябва да бъде нула, можем да получим два коефициента, свързани с асиметрията и ексцеса на разпределението на грешките:

$$As = \frac{E(\varepsilon^3)}{({\sigma_\varepsilon}^2)^{\frac{3}{2}}}, \qquad (\text{V—25})$$

$$Kr = \frac{E(\varepsilon^4)}{({\sigma_\varepsilon}^2)^2}, \qquad (\text{V—26})$$

с чиято помощ може да получим и самия показател на Jarque-Bera:

$$JB = N\left(\frac{As^2}{6} + \frac{(Kr-3)^2}{24}\right) \sim \chi^2(2)\,. \qquad (\text{V—27})$$

Тъй като показателят следва $\chi^2$ разпределение с две степени на свобода, това ни позволява да направим тест на нулевата хипотеза, че отсъстват асиметрия и ексцес, което е равносилно да твърдим, че грешките са нормално разпределени.

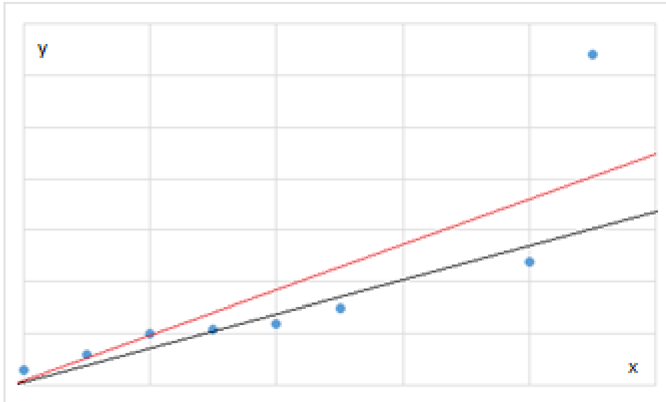
След като вече разполагаме с методи за проверка за нормалност, остава да дадем отговор на по-важния въпрос – какво да направим, в случай че резултатите посочат отклонения между разпределението на остатъчните стойност и нормалното. Можем да посочим два основни подхода за справяне със ситуацията:

- да използваме други методи за анализ и оценка на параметрите на регресионната зависимост, които не са чувствителни спрямо нормалното разпределение на остатъчните стойности

На практика това означава да търсим алтернативен алгоритъм за оценка, който не изисква толкова много и строги допускания. Но това е по-лесно да се каже, отколкото да се направи, поради две причини – няма гаранция дали използването на друг метод за оценка няма да доведе до друг тип неточности или да измести акцента от същността на решавания икономически проблем към справянето с един формален недостатък; използването на алтернативни методи е свързано обикновено с по-сложни изчисления, откъдето нараства и вероятността за грешка.

- да се обърнем за помощ към финансовите и икономическите модели и да се опитаме да направим анализ защо е нарушено допускането за нормално разпределение на остатъчните стойности

Този подход невинаги ни гарантира успех, но може да бъде полезен в случаите, когато допускането за нормалност е нарушено благодарение само на няколко крайни/екстремни стойности на целевата променлива. Както показва Фигура 14, наличието на такива крайни стойности може да доведе до получаването на неточни оценки за параметрите на регресията.




*Фигура 14. Влияние на екстремните стойности върху оценката на параметрите на регресията*

В някои частни случаи влиянието на тези малко на брой, но с висока стойност (поради голямата стойност на целевата променлива) случаи може да доведе не само до разместване в стойността на параметъра, но дори и до обръщане на знака му (с което пък се изкривява информацията за начина, по който факторът влияе върху целевата променлива). В случай че финансовият модел допуска това може да направим опит да „елиминираме“ крайните стойности с цел да подобрим общата точност на модела. Тук под отстраняване на влиянието им, нямаме предвид просто да забравим за тяхното съществуване (т.е. да не ги включваме в регресията), понеже това би нарушило целостта на анализираната съвкупност и евентуално нейния представителен характер (ако тя е част от по-голяма съвкупност). Вместо това можем да въведем допълнителни „служебни“ променливи в регресионната зависимост, които са нула във всички останали случаи освен в случаите с екстремни стойности на целевата променлива. По този начин с изчисляването на „служебните“ им параметри ще можем да елиминираме крайните стойности и ефекта, който те имат върху получените резултати.

Трябва да предупредим обаче, че самоволното изключване на определени стойности от анализа (без значение какъв е начинът, по който то се осъществява) крие немалко рискове. Те са свързани преди всичко с това дали имаме право да елиминираме определени случаи просто защото те не „пасват“ на изградения от нас модел. Има разлика дали става дума за сезонни колебания, които ние се опитваме да изгладим, или за изключително редки събития, за които предполагаме, че няма да възникнат отново скоро. В крайна сметка ние правим модел и използваме статистически и математически методи с цел да извлечем максималната полезна информация от наличните данни. Ако доброволно се лишаваме от важни части на пъзела, ние рискуваме никога да не добием точна представа за изследваните процеси или да бъдем неприятно изненадани от ефекти, които са останали извън фокуса на нашето внимание. Поради тази причина начините за справяне с отклоненията от нормалното разпределение при остатъчните стойности зависят до голяма степен от конкретните проучвани проблеми и ограниченията, с които се налага да се съобразяваме. В този смисъл и разглежданите тук методи трябва да се приемат по-скоро като указващи пътя, а не като инструменти, валидни за всяка една ситуация.

## 6. Задачи и въпроси за размисъл

1. Избройте всички допускания за ефективно и адекватно използване на линейните регресионни модели.
2. При нарушаването на кои от допусканията се получават изместени оценки за параметрите на регресията?
3. Обяснете коя част от твърдението е невярна „$\chi^2$ разпределението има два параметъра и при голям брой степени на свобода се доближава до нормалното разпределение“.
4. Начертайте графика на разпределение, което не е симетрично и такова с висок ексцес.

# VI. Специфични проблеми при линейните регресионни модели и подходи за решаването им

В предната глава разгледахме някои от методите с чиято помощ можем да проверим дали допусканията, свързани с използването на линейни регресионни модели и метода на най-малките квадрати, са налице. Едновременно с това акцентирахме върху проблемите, с които финансовите анализатори могат да се сблъскат, ако не обърнат достатъчно внимание върху нарушенията на тези предпоставки. За съжаление обаче, не всички възможни случаи се покриват от проверката на петте допускания от Таблица 4. Поради тази причина е необходимо да разгледаме няколко ситуации, при наличието на които е възможно да изтълкуваме погрешно резултатите.

## 1. Мултиколинеарност

Мултиколинеарността е свързана с анализ на факторите, които участват в регресионната зависимост. До момента те ни интересуваха основно поради влиянието, което оказват върху стойностите на целевата променлива, и факта, че с помощта на наблюдаваните им стойности можем да получим оценка на параметрите на регресията. Встрани от нашето внимание беше оставен въпросът дали има връзка между отделните фактори и как тя може да повлияе върху тълкуването на получените резултати и тяхната надеждност. В случаите, когато не е налице връзка между отделните $x_i$, то тогава казваме, че те са ортогонални помежду си и премахването или добавянето на някои от тях ще повлияе върху общата обясняваща сила на модела, но няма да доведе до промяна в оценките на коефициентите пред останалите обясняващи променливи. Нека да си припомним, че общият вид на регресионните модели, които разглеждахме дотук беше:

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i, \qquad \text{( VI—1 )}$$

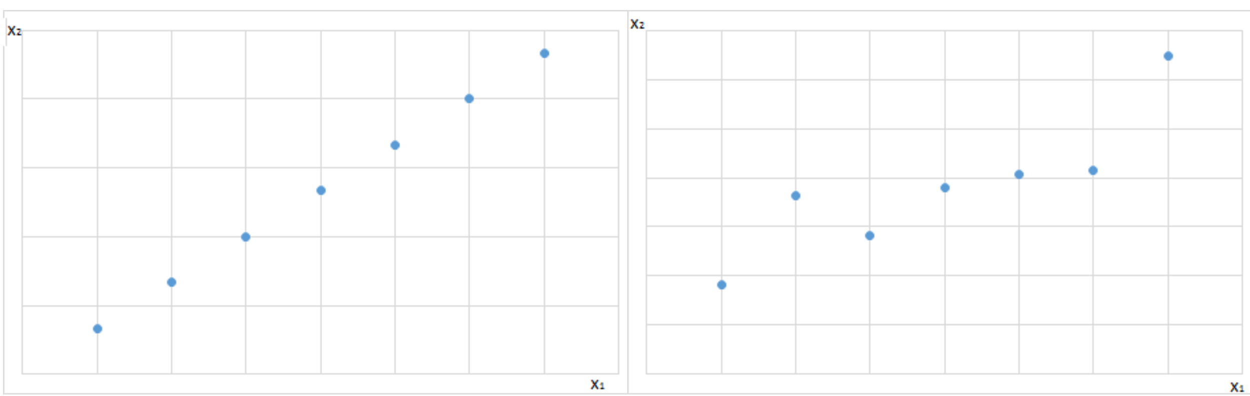
а допускането за липса на връзка означава, че корелацията между отделните фактори е нула. Обърнете внимание, че използваме понятието „корелация“ в по-широк смисъл като наличие на определена зависимост, която не е задължително да включва само две от променливите – т.е. $corr(x_i, x_j) = \frac{cov(x_i, x_j)}{\sigma_{x_i}\sigma_{x_j}} = 0$ е частен случай и необходимо но не и достатъчно условие за ортогоналност. Когато не можем да гарантираме независимост на променливите $x_1, x_2, \ldots, x_k$, а имаме силна връзка между тях, казваме, че е налице мултиколинеарност, която може да приема две форми:

- Перфектна (пълна) мултиколинеарност, при която имаме изразена точна и ясна връзка между някои от факторите

В тази форма е налице математическа връзка между някои от факторите в модела, която е изпълнена изцяло за всички техни наблюдавани стойности. Нека за момент да си представим, че е налице връзка между $x_1$ и $x_2$: $x_1 = \frac{1}{3}x_2$. Това означава, че ако се опитаме да изградим модел, в който да участват и двете променливи, няма да можем да намерим стойностите на всички коефициенти от него[14]. На практика е малко вероятно да попаднем в ситуация, при която е налице перфектна зависимост между подбраните фактори. Ако все пак това се случи, то проблемът може да бъде отстранен посредством преформулиране на модела, доколкото не е необходимо в него да се включват едновременно всички от участващите във връзката елементи.

- Почти пълна мултиколинеарност, при която имаме силна, но не перфектна връзка между някои от факторите в модела

В тази форма е налице висока степен на зависимост, която в рамките на нашия пример ще означава висока стойност на коефициента на корелация между $x_1$ и $x_2$, но без да е налице перфектна зависимост (респ. коефициент на корелация от единица). Разликата между двата случая е демонстрирана на Фигура 15.




*Фигура 15. Перфектна (вляво) и почти пълна (вдясно) мултиколинеарност*

В случаите, когато предполагаме, че е възможна връзка по двойки между факторите, можем да направим сравнително проста проверка дали попадаме в някой от двата случая на мултиколинеарност. Ако изградим една матрица с коефициентите на корелация, то тя ще

[14] Математическото доказателство на въпросното твърдение е свързано с това, че поради зависимостта на двете променливи ще имаме $k{-}1$ независими уравнения за коефициентите $\alpha, \beta_1, \beta_2, \dots, \beta_{k-1}$.

има размерност $k \times k$, като всяка клетка ще ни даде представа за степента на зависимост между съответните променливи.

Ще разгледаме един случай, при който мултиколинеарността може да окаже ефект върху получените крайни резултати. Нека предположим, че нашата цел е да анализираме модификация на известните ни от финансите единични индексни модели:

$$r - r_f = \alpha + \beta\left(r_i - r_f\right), \qquad \text{( VI—2 )}$$

където $r$ е възвръщаемостта на изследвания актив, $r_f$ е безрисковата възвръщаемост, а $r_i$ е възвръщаемостта на пазарния индекс, с чиято помощ се изгражда моделът. Ако използваме и други показатели, а не само стойността на пазарния индекс, можем да получим следния вариант на модела:

$$r - r_f = \alpha + \beta_1\left(r_{inf} - r_f\right) + \beta_2\left(r_{AAA} - r_f\right) + \beta_3\left(r_{large\ cap} - r_f\right) + \beta_4\left(r_{small\ cap} - r_f\right), \qquad \text{( VI—3 )}$$

където $r_{AAA}$ е индекс на възвръщаемостта на облигации с най-висок рейтинг, $r_{large\ cap}$ е индекс на възвръщаемостта на акциите на фирми с голяма пазарна капитализация, $r_{inf}$ е темпът на инфлация, а $r_{small\ cap}$ е индекс на възвръщаемостта на акциите на фирми с малка пазарна капитализация. Разбира се, ние можем да добавим и още фактори, но за демонстрация на ефекта от мултиколинеарността и тези са достатъчно. При проверката за корелацията по двойки между отделните фактори ще получим таблица, която прилича на посочената по-долу:

*Таблица 14. Примерна проверка за мултиколинеарност по двойки*

| Променливи | $r_{inf}$ / $r_{AAA}$ | $r_{inf}$ / $r_{large\ cap}$ | $r_{inf}$ / $r_{smal\ cap}$ | $r_{AAA}$ / $r_{large\ cap}$ | $r_{AAA}$ / $r_{small\ cap}$ | $r_{large\ cap}$ / $r_{small\ cap}$ |
|---|---|---|---|---|---|---|
| Коеф. на корелация | 0.82 | 0.2 | 0.25 | 0.4 | 0.2 | 0.85 |

Ако търсим подкрепа на твърдението за наличие на мултиколинеарност, то елементите, към които първо трябва да се насочим, са първата и последната двойка показатели, тъй като те имат най-високи стойности за коефициента на корелация. Разбира се, бихме могли да „филтрираме“ отделни фактори според теоретичните предпоставки за връзка помежду им. Но използването на реалните данни е винаги за предпочитане, за да можем да изградим нашия модел на базата на взаимодействие на теория и реално наблюдавани стойности. Въпреки че това не е универсално правило, много често като прагова стойност, която предполага допълнително проучване за мултиколинеарност, се приема коефициент на корелация, по-висок от 0.4.

Трябва да отбележим обаче, че посоченият в примера подход ще ни бъде от полза в случаите, когато предполагаме наличието на зависимост „по двойки“. Ако е налице основание да смятаме, че съществува зависимост между повече от една променливи, то тогава резултатите от Таблица 14 няма да ни бъдат от полза. В този случай се налага да погледнем на мултиколинеарността от различен ъгъл, а именно че в случая за нас е важно не просто дали има, или няма зависимост между отделните фактори, а каква е нейната сила (т.е. до каква степен те са зависими).

Разликата между „сила на зависимост“ и „проверка за зависимост“ изглежда малка на пръв поглед, но тя е много съществена при анализа на мултиколинеарността. За разлика от изследването за изпълнимост на допусканията на линейната регресия, които разглеждахме по-рано, тук ние не правим тест за „наличие“ или „липса“, а се стремим да определим дали силата на зависимост между отделните обясняващи променливи е толкова голяма, че да ни доведе до неправилни изводи. Тръгвайки от тази позиция, можем да дефинираме две стратегии за допълнителен анализ на мултиколинеарността:

- с помощта на коефициентите за частична корелация − ако получените с помощта на метода на най-малките квадрати оценки за параметрите на регресията показват висока обясняваща сила (т.е. висока стойност на $R^2$), но едновременно с това значимостта на всеки един от получените коефициенти е малка;
- с помощта на допълнителни регресионни зависимости (т.нар. външни регресии) − при тази стратегия освен основното регресионно уравнение ( VI—3 ) трябва да оценим още няколко допълнителни зависимости, при които всяка една от обясняващите променливи се представя като функция от останалите. В контекста на предния пример ще имаме:

$$r - r_f = \alpha + \beta_1(r_{inf} - r_f) + \beta_2(r_{AAA} - r_f) + \beta_3(r_{large\ cap} - r_f) + \beta_4(r_{small\ cap} - r_f)$$

$$r_{inf} = \alpha_1 + \beta_{21} r_{AAA} + \beta_{31} r_{large\ cap} + \beta_{41} r_{small\ cap}$$

$$r_{AAA} = \alpha_2 + \beta_{12} r_{inf} + \beta_{32} r_{large\ cap} + \beta_{42} r_{small\ cap}$$

$$r_{large\ cap} = \alpha_3 + \beta_{13} r_{inf} + \beta_{23} r_{AAA} + \beta_{43} r_{small\ cap}$$

$$r_{small\ cap} = \alpha_4 + \beta_{14} r_{inf} + \beta_{24} r_{AAA} + \beta_{34} r_{large\ cap}$$

В случай че обясняващата сила на някоя от последните четири зависимости е висока, това означава, че един от факторите може да бъде представен като линейна комбинация от останалите и трябва да се съобразяваме с мултиколинеарността и нейните ефекти върху получените крайни резултати.

Когато не можем да отхвърлим допускането за мултиколинеарност, остава да вземем решение за действията, с които да неутрализираме негативните ѝ ефекти. Съществуват няколко подхода, с чиято помощ можем да търсим решение на проблемите със зависимост между обясняващите променливи в линейния регресионен модел:

- Да не предприемаме никакви действия за елиминиране на ефектите от мултиколинеарността.

„Забравянето" за проблема обаче не води до неговото автоматично решение. Тъкмо напротив – в този случай се излагаме на опасност да бъдем подведени от получените високи оценки за общата обясняваща сила на регресионния модел (измерена с $R^2$ или модифицирания $R^2$).

- Ако предполагаме зависимост между променливите $x_i$ и $x_j$, тогава можем да използваме в уравнение ( VI—3 ) само една от тях.

Този подход е свързан с две принципни трудности: първо, да установим между кои фактори има връзка, и второ, да изберем кои от тях да останат в модела. Когато сме изправени пред почти пълна мултиколинеарност (което е и по-често срещаният случай), елиминирането на някои от обясняващите променливи може да повлияе на общата точност на модела. Освен това изпускането на обясняващи променливи трябва да има солидна теоретична основа, а не да бъде направено единствено с цел „напасване" на регресията към наличните данни.

- Да увеличим обема на анализираните данни.

Увеличаването на обема на анализираните данни, с чиято помощ оценяваме параметрите на линейната регресионна зависимост, би трябвало да доведе до снижаване на грешката в получените оценки. Разбира се, „би трябвало" не означава винаги, а освен това е възможно вече да сме използвали при оценките всички налични данни и физически да не е възможно да увеличим техния обем.

- Да преобразуваме регресионния модел, така че да включва съотношение между обясняващите променливи, за които можем да предполагаме връзка.

Този подход има слабости, които са аналогични на случая с използване само на една от променливите, които са „заподозрени", че се намират във връзка. Използването на коефициент освен всичко друго води и до промяна в смисъла (тълкуването) на получената оценка на съответния параметър от линейната регресионна зависимост.

В зависимост от естеството на анализираните проблеми можем да използваме и комбинация от изброените подходи за решаване на проблемите, произтичащи от наличието на мултиколинеарност.

## 2. Тест на Ramsey и проверка за адекватност на линейната зависимост

Линейната регресионна зависимост, както показва и самото име, предполага наличието на определен тип зависимост между целевата променлива и факторите. Въпреки че този момент рядко попада в списъка на изричните допускания, той е изключително важен за качеството на получените резултати. Казано по друг начин – ако ние не можем да проверим, че зависимостта е линейна, то всички последващи доказателства и проучвания за приложимост на метода на най-малките квадрати губят от своето значение и смисъл.

Съществуват редица методи за проверка дали линейната (т.е. от първа степен) функционална форма е подходяща, за да представи връзката между целевата променлива и факторите в един регресионен модел. За целите на иконометричното моделиране можем да обобщим подходите към проблема в две големи групи, които не са взаимно изключващи се, а е уместно да бъдат използвани съвместно:

- Подходи, използващи теоретичните модели, залегнали при изграждането на регресията

Водещите доводи за утвърждаване на линейната функционална форма (или за нейното отхвърляне) в този случай са характеристиките на използваните теоретични концепции, които дават обяснение на изследваните явления. Трябва да отбележим, че не е задължително теоретичните модели да предполагат строго линейна зависимост – достатъчно е да е налице функционална връзка, която с прости преобразувания да може да бъде сведена до линейна. Таблица 15 представя списък със зависимости, които биха могли да бъдат преобразувани до линейни и следователно ни дават основание да прибегнем до услугите на линейната регресия и метода на най-малките квадрати.

Както посочва и Brooks [3], промяната на функционалната форма е само една от стъпките при свеждането на зависимостта до линейна. По-важната от гледна точна на последващия финансов и икономически анализ задача е да определим коректно как да тълкуваме получените оценки за параметрите на модифицираното регресионно уравнение.

*Таблица 15. Примерни функционални форми, които могат да бъдат приведени до линейни зависимости*

| Функционална зависимост | Начин на преобразуване |
|---|---|
| Линеен модел<br>$y \;\;= \alpha + \beta_1 x_1 + \; \beta_2 x_2 + \cdots + \beta_k x_n$<br><br>Пример:<br>Модел за оценка на капиталови активи (МОКА):<br>$r = r_f + \beta(R_m - r_f)$<br>В този случай $x_1 = \left(R_m - r_f\right),\; y = r$. | Не е нужно преобразуване, тъй като формата поначало е линейна. Единица промяна в стойността на обясняващата променлива $x_i$ ще доведе до $\beta_i$ промяна в стойността на y, при условие че всички останали променливи запазят стойностите си. |
| Лог-линеен модел<br>$\ln(y) \;\;= \alpha + \beta_1 x_1 + \; \beta_2 x_2 + \cdots + \beta_k x_k$<br><br>Примерна начална форма:<br>Функции, описващи различни нараствания като експоненциална зависимост. | Обикновено този тип модели се получават след логаритмуване на първоначална функционална форма, при която обясняващите променливи участват като степенен показател. Тълкуването на променената форма е, че всяка промяна на стойността на $x_i$ с единица ще доведе до $100.\beta_i$ процента промяна в стойността на y (при условие че останалите стойности не се променят). |
| Линейно-логаритмичен модел<br>$\mathrm{y} = \alpha + \beta_1 \ln(x_1) + \; \beta_2 \ln(x_2) + \cdots + \beta_k \ln(x_k)$<br><br>Примерна начална форма:<br>Изследване на икономически процеси и явления, при които имаме големи разлики в отделните стойности и малко на брой, но силно отклоняващи се от останалите стойности (например разпределението на богатството). | Обикновено този тип модели се получават след логаритмуване на първоначална функционална форма при която целевата променлива участва като степенен показател. Тълкуването на променената форма е, че всяка промяна на стойността на $x_i$ с 1% ще доведе до $\frac{\beta_i}{100}$% промяна в стойността на y (при условие че останалите стойности не се променят). |
| Логаритмичен модел<br>$\ln(y) = \alpha + \beta_1 \ln(x_1) + \; \beta_2 \ln(x_2) + \cdots + \beta_k \ln(x_k)$<br><br>Примерна начална форма:<br>Производствена функция на Коб-Дъглас:<br>$y = AL^{\beta_1} K^{\beta_2}$<br>В този случай трансформацията ще има вида:<br>$\ln(y) = \ln(A) + \beta_1 \ln(L) + \beta_2 \ln(K)$,<br>а съответно $x_1$ и $x_2$ ще бъдат вложените труд и капитал. | Обикновено този тип модели се получават след логаритмуване на първоначална функционална форма, при която и целевата променлива, и факторите участват като степенен показател.<br>Тълкуването на променената форма е, че при равни други условия нарастването на стойността на $x_i$ с 1% ще доведе до $\beta_i$% промяна в стойността на $y$. |

- Подходи, базирани непосредствено на характеристиките на анализираните данни

В този случай подкрепата или основанието за отхвърляне на линейната функционална форма се базират на анализ на данните от изследваната съвкупност. Въпреки че съществуват различни конкретни методи, с чиято помощ се реализира проверката за коректност на функционалната форма, в крайна сметка това, което правим, е да проверим една статистическа хипотеза. А именно дали са налице достатъчно основания, за да отхвърлим използването на линейна връзка при изграждането на регресионния модел. Един от широко използваните методи за проверка на несъответствие на използваната функционална форма е тестът на Ramsey [15], наричан още RESET тест. RESET извършва статистическа проверка на хипотезата, че връзката между целевата променлива и факторите е линейна посредством построяването на една допълнителна регресионна зависимост, където всяка от променливите $x_i$ участва със степени, по-високи от първа. Ако „истинската“ функционална форма на зависимост не е линейна, то можем да очакваме, че допълнителната регресионна зависимост ще има по-добра (т.е. по-висока) обясняваща сила и съответно за нея ще получим по-високи стойности за показателя $R^2$. Ако обаче не станем свидетели на нарастване на $R^2$ или това нарастване не съответства на теоретичните изисквания, то тогава няма да имаме основание да отхвърлим хипотезата, че линейната зависимост отразява реалната връзка между изследваните променливи.

Ако имаме една типична линейна регресионна зависимост:

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i, \qquad \text{( VI—4 )}$$

то тогава, за да проведем теста на Ramsey, ще трябва да направим една допълнителна регресионна зависимост от вида:

$$y_i = {y_i}^2 + {y_i}^3 + \cdots + {y_i}^q + \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i\ , \qquad \text{( VI—5 )}$$

а нулевата ни хипотеза ще бъде, че „линейната форма на зависимост е правилна“. За да имаме основание да отхвърлим тази хипотеза, трябва да сравним получения за допълнителната регресия $R^2$ показател с теоретично очакваната му стойност. Определянето на показателя за обясняваща сила става по вече описания в Част I метод, а сравняването с теоретичната му стойност можем да направим, като имаме предвид, че $nR^2$ (където $n$ е броят на данните/наблюденията, въз основа на които изграждаме нашите оценки за параметрите на регресията) следва $\chi^2$ разпределение с (q−1) степени на свобода. Следователно крайното решение ще вземем съобразно, получените резултати:

- Ако $nR^2$ е по-голямо от теоретичната стойност на $\chi^2$ разпределение с (q−1) степени на свобода[15] (съобразно с избраното ниво на съгласие), то можем да отхвърлим

[15] q е най-високата степен, с която изграждаме допълнителната регресионна зависимост.

нулевата хипотеза и да заключим, че линейната функционална форма на зависимост не е подходяща за изследване на тези икономически и финансови процеси
- Ако $nR^2$ е по-малко или равно на теоретичната стойност на $\chi^2$ разпределение с (q-1) степени на свобода (съобразно избраното ниво на съгласие), то нямаме достатъчно основание да отхвърлим нулевата хипотеза. Т.е. нямаме достатъчно основание да приемем, че линейната функционална форма на зависимост не е подходяща за използване.

Тестът на Ramsey ни дава възможност да направим проверка дали линейната функционална форма на зависимост е адекватна за анализираните данни. Но RESET методът не ни дава директен отговор на два важни въпроса:

- От коя степен $q$ да бъде допълнителната регресионна зависимост и как можем да гарантираме, че използването на по-ниска или по-висока степен няма да повлияе на крайното решение.
- В случаите, когато имаме достатъчно основание да отхвърлим нулевата хипотеза, това не ни дава допълнителна информация коя е правилната функционална зависимост.

За да дадем отговор на въпроса коя е правилната форма на функционална връзка между целевата променлива и факторите, можем да използваме два общи подхода:

- Да се позовем на съществуващите икономически и финансови модели, като приемем, че типът на връзката трябва да съответства на това, което те налагат.

В този случай изборът на функционална връзка е улеснен от предписанията на теоретичните модели и зависимости. За съжаление, опростяването на нашата задача в този случай си има цена – приемането на един определен теоретичен постулат като правило е свързано с неявното съгласие с всички допускания и условия, които го съпътстват. Не можем да дадем обща рецепта за всички случаи кога е допустимо безрезервно да се възползваме от общите икономически модели, затова решението трябва да бъде направено в контекста на конкретно анализираните проблеми.

- Да използваме техники за полиномна интерполация, като например параболична, бикубична и др.

Съществуват редица проблемни ситуации, при които не е възможно да приведем първоначалната форма на функционална зависимост в линейна (или да използваме някоя от коментираните в Таблица 15 възможности). Тогава се налага да използваме решения, различни от метода на най-малките квадрати. Основните неудобствата тогава са свързани с допълнително усложняване на изчислителните процедури, което обаче в съвременните условия не е чак толкова сериозен проблем (обикновено нито едно сериозно изследване не се прави „на ръка“ и без помощта на компютър). Това, което представлява по-голямо предизвикателство при използването на различните методи за интерполация, е как да тълкуваме получените оценки за параметрите на зависимостта, за да не бъдем подведени да направим погрешни изводи. Допълнителен проблем, на който ще обърнем внимание и в това изложение, е как да определим в каква степен получените резултати са стабилни и валидни извън анализираната извадка или ограничена съвкупност от данни. Различните форми на полиномна интерполация се използват широко във финансите, като например при оценката на кривата на доходността, форуърдните лихвени проценти и оценката на облигации.

## 3. Проблеми свързани с избора на обясняващи променливи

Обикновено проблемите, свързани с избора на обясняващи променливи, биват два вида: от „технологично“ естество − свързани с мащабиране или набавяне на достатъчен обем от данни, и проблеми „по същество“ − които са свързани най-често с избора какви точно данни да включим в регресионния модел. Техническите проблеми, въпреки че изискват нужното внимание, имат относително по-ниска сложност и могат да бъдат разрешени сравнително лесно. Проблемите „по същество“ обаче могат да имат много силен ефект спрямо общото качество и използваемост на получените крайни резултати. Обикновено те са свързани с приложението на конкретни теоретични постановки и тяхното разрешаване често надхвърля мащабите само на конкретното проучване и обема от данни, с който оценяваме параметрите на регресионния модел. Ние ще обърнем внимание на два важни случая, които имат пряко отношение към основните характеристики и качества на изгражданите иконометрични модели и приложението на метода на най-малките квадрати за оценка на параметрите на регресионната зависимост:

- включването в анализа на по-малко от „необходимите“ променливи – т.е. изключването на определени обясняващи характеристики, които са важни за анализираните финансови процеси и явления;
- включването на повече от „необходимите“ променливи − т.е. включването на определени обясняващи характеристики, които не оказват съществено влияние върху целевата променлива и съответно анализираните финансови процеси и явления.

И двете възможности предполагат, че ще получим модел, който ще се отклонява от идеалния, но ефектите ще бъдат различни в зависимост от това дали пропускаме важен фактор, или просто добавяме фактор, който не влияе на целевата променлива в регресионния модел. Разбира се, най-добре би било да изградим анализа си въз основа на точно тези фактори, които могат да обяснят измененията в целевата променлива. На практика обаче тази възможност е по-скоро пожелателна, отколкото лесно изпълнима, защото:

- Дори и да изградим нашите модели въз основа на най-новите теоретични постижения на икономическата наука, рискуваме да не можем да обхванем напълно всички фактори, които влияят на целевата променлива. Теоретичните постулати представляват стилизиран и често опростен образ на реалността – съпътстван от множество допускания и опростявания. Ако искаме да „паснем“ изцяло на теоретичните рамки, ще трябва да обработваме предварително входните данни, с които работим, откъдето може да достигнем ситуация, при която да анализираме идеализиран образ на действителността, вместо самата действителност.
- Съществуват проблеми, които могат да бъдат обяснени от повече от един теоретичен модел, при това невинаги има достатъчно емпирични доказателства в полза на един от всички модели. Така качеството на получените резултати може да се окаже жертва още на първото решение за избор на теоретична „платформа“, с което да обезсмислим последващите проверки на хипотези и иконометрични тълкувания.

Ефектите от проблемите свързани с избора на обясняващи променливи са обобщени в Таблица 16. Наличието на изместване в оценката на параметрите на регресията нарушава третото свойство, описано в Таблица 5, и на практика е свързано със систематично надценяване (респ. подценяване) на влиянието на дадена факторна променлива.

*Таблица 16. Ефекти, породени от избора на обясняващи променливи*

<table>
<tr><th colspan="2" rowspan="2"></th><th colspan="2">Корелация между липсващите/допълнителните и останалите обясняващи променливи</th></tr>
<tr><th>Висока</th><th>Ниска</th></tr>
<tr><td rowspan="2">Избор на обясняващи променливи</td><td>Изпускане на важни за анализираните финансови процеси и явления променливи</td><td>Получените оценки за параметрите на регресионната зависимост ще бъдат изместени и стандартната грешка ще бъде по-висока, отколкото ако в модела участват и пропуснатите променливи. Получените оценки няма да отговарят на изискването за консистентност.</td><td>Получените оценки за параметрите на регресионната зависимост ще бъдат изместени, но негативният ефект ще бъде относително по-нисък спрямо случая с висока корелация. Получените оценки няма да бъдат консистентни, както и в случая с висока степен на корелация между изпуснатата и останалите обясняващи променливи.</td></tr>
<tr><td>Включване на повече от необходимите променливи</td><td>Получените стойности за грешка в оценките на параметрите на регресията ще бъдат по-високи, отколкото ако допълнителните променливи не участваха в модела. Получените оценки за параметрите обаче ще бъдат неизместени и консистентни.</td><td>Получените стойности на стандартната грешка ще бъдат по-високи, отколкото ако допълнителните променливи не участваха в модела. При ниска степен на корелация обаче, негативния ефект ще бъде по-слаб спрямо случая с висока корелация между обясняващите променливи.</td></tr>
</table>

Математическият израз на разглежданите две ситуации ще означава, че вместо регресионна зависимост от типа ( VI—6 ) ще имаме различно по вид уравнение.

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \qquad \text{( VI—6 )}$$

Да предположим, че ние използваме в първия случай:

$y_i = \alpha + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$ (изпуснали сме важната променлива $x_1$). ( VI—7 )

Или сме добави допълнителна обясняваща променлива $x_{k+1}$:

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \beta_{k+1} x_{i(k+1)} + \varepsilon_i \, . \qquad \text{( VI—8 )}$$

Внимателният читател може да се запита защо е необходимо да посочваме втората възможност като източник на грешки, след като, ако регресионната зависимост описва точно изследваните процеси, би трябвало автоматично да получим оценка за коефициента $\beta_{k+1} = 0$? Въпреки че подобен въпрос има своето основание трябва да припомним две важни характеристики на получените резултати:

- първо – дори и линейната зависимост да апроксимира добре изследваните процеси, то няма гаранция, че ние разполагаме с достатъчно точни и изчистени от „шум“ (респ. странични ефекти) данни;
- второ – не трябва да забравяме, че получените резултати за коефициентите в регресионното уравнение са преди всичко **оценки**. Т.е. те могат да съвпадат абсолютно точно с „верните“ стойности , но могат и просто да бъдат близо до тях. А стойност, която е малка, но все пак не е нулева (т.е. $\beta_{k+1}$ е малко по абсолютна стойност, но не е нулево) ще има ефект върху останалите получени стойности и при изчисляването на стандартната грешка.

Проблемите породени от избора на обясняващи променливи, се наричат често и грешки на спецификацията (specification error) поради факта, че те са заложени още при избора на модел, чиито параметри се опитваме да оценим. Поради сложния характер на изследваните финансови процеси и явления е много трудно да се предпазим изцяло от подобен тип грешки, но можем да се опитаме поне да намалим тяхното значение и общия им негативен ефект върху получените крайни резултати и изводи. За целта можем да използваме едно общо правило, че ако се налага да избираме между грешка. породена от изпускането на важна променлива, или такава, предизвикана от включването на несъществена променлива, е по-добре да се спрем на втората възможност. Този извод се подкрепя и от изложените в Таблица 16 разсъждения и стремежа да запазим получените оценки за параметрите на регресията консистентни и неизместени. Посоченото правило не означава обаче, че моделите трябва да се „разводняват“ или че можем произволно да добавяме нови обясняващи променливи, водени единствено от мисълта да не изпуснем важен обясняващ фактор. Таблица 17 съдържа четири важни критерия, систематизирани от Studenmunt [16], които могат да ни дадат насока дали определена променлива трябва да бъде включена в регресионното уравнение.

*Таблица 17. Критерии за включване/изключване на една факторна променлива в регресионния модел*

| Критерий | Обяснение |
|---|---|
| **Въпрос 1:** Вписва ли се обясняващата променлива в използваните теоретични модели? | Ако обясняващата променлива може да бъде съгласувана или участва в теоретичните модели, използвани при анализа, то тя би трябвало да бъде включена и в регресионното уравнение. |
| **Въпрос 2:** Издържат ли получените оценки за коефициента на дадена обясняваща променлива на t-тест за значимост? | Ако са налице достатъчно основания да смятаме, че факторната променлива е статистически значима, то тя трябва да бъде включена в регресионното уравнение. |
| **Въпрос 3:** Дали включването/изключването на дадена обясняваща променлива има ефект върху обясняващата сила на модела (обикновено измерена посредством модифицирания $R^2$)? | Съществената промяна в обясняващата сила на изградения модел като резултат от включването/изключването на факторната променлива обикновено е индикатор, че тя трябва да участва в регресионното уравнение. |
| **Въпрос 4:** Води ли включването или изключването на дадена обясняваща променлива до големи промени в оценките на останалите параметри на регресионното уравнение? | Съществената промяна в получените оценки на параметрите на регресионното уравнение е индикатор, че факторната променлива трябва да бъде включена в анализа. |

В случаите, когато имаме разнопосочни отговори на четирите въпроса от Таблица 17, изборът − дали дадена променлива да бъде включена, или не − трябва да бъде взет съобразно с характеристиките на конкретния анализиран проблем. Тогава обикновено решението се взема на базата на комплекс от съображения, които могат да включват допълнителен анализ или стремеж да се доближим до вече съществуващ теоретичен модел.

## 4. Проблеми свързани със стабилността на получените оценки и резултати

В последните две глави разгледахме редица допускания, които са свързани с приложението на линейните регресионни модели и оказват въздействие върху качеството и достоверността на получените крайни резултати. Ако един иконометричен анализ претендира за пълнота и адекватност, то той би трябвало да съдържа не само получените оценки на параметрите на регресионната зависимост, но и допълнителните проверки, с които да удостовери, че условията за приложимост на линейните регресионни модели са налице.

Има едно допускане обаче, на което не сме отредили достатъчно значима роля в изложението до момента. То е неявно, но е изключително важно и се отнася за това, до каква степен получените с помощта на ограничен обем от данни оценки са приложими, за да опишат изследваните финансови процеси и явления въобще. В по-тесен смисъл същото предположение е свързано с това, до каква степен получените оценъчни стойности за параметрите на регресионната зависимост са стабилни или променят значително стойността си в зависимост кои данни използваме, за да ги изчислим. Съществуват различни методи, с чиято помощ можем да оценим стабилността на получените параметри, или казано по друг начин − да преценим дали в регресионния модел:

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \qquad ( \text{VI—9} )$$

ще получим стойности за $\alpha, \beta_1, \beta_{2,} \dots, \beta_k$ , които не се влияят силно от факта, каква част от наличните данни използваме, за да ги оценим. Интуитивният подход, за да направим тази преценка, е да разделим наличните данни на групи и да сравним получените стойности за параметрите на регресионната зависимост за всяка група, както и за цялата съвкупност от данни. В този контекст ще разгледаме два от възможните тестове за стабилност на получените оценки – теста на Chow (Чоу) [19] и проверка на прогнозиращите качества на модела.

### 4.1. Тест на Chow

Тестът на Chow е свързан с разделянето на наличните данни на две групи и оценка на параметрите на регресионното уравнение отделно за всяка една от двете групи, плюс оценка, получена въз основа на всички налични данни. По този начин се получават общо три регресионни зависимости (m е разделителят и $1 < m < n$):

$$\begin{aligned} y_i &= \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i, i \in [1, \dots, n] \\ y_i &= \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i, i \in [1, \dots, m) \\ y_i &= \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i, i \in [m, \dots, n] \end{aligned} \qquad ( \text{VI—10} )$$

Всяка една от тях ще ни даде набор от оценки за параметрите $\alpha, \beta_1, \beta_{2,} \dots, \beta_k$ и съответно ще има различна сума от квадратите на остатъчните стойности – $RSS_{1..n}$, $RSS_{1..m}$ и $RSS_{m..n}$. След като разполагаме с изброените регресионни зависимости, можем да формулираме теста на Chow като специфична статистическа проверка на хипотеза. Нека дефинираме като нулева хипотеза, че стойностите на коефициентите не се променят и остават същите при трите регресионни зависимости. Ако това е изпълнено, то тогава би трябвало и сумата от квадратите на остатъчните стойности за всички анализирани данни (т.е. $i \in [1, \ldots, n]$) да не бъде много по-висока от сумата на квадратите на остатъчните стойности за двете съвкупности $i \in [1, \ldots, m)$ и $i \in [m, \ldots, n]$. Именно посредством показателите $RSS_{1..n}$, $RSS_{1..m}$ и $RSS_{m..n}$ можем да направим и проверка на нулевата хипотеза:

$$\text{показател на } Chow = \frac{RSS_{1..n} - (RSS_{1..m} + RSS_{m..n})}{RSS_{1..m} + RSS_{m..n}} \cdot \frac{n - 2(k+1)}{k+1} \quad \text{( VI—11 )}$$

Показателят на Chow прилича изключително много на F-теста, който разглеждахме в частта „Съвместна проверка на повече от една хипотези“, с тази разлика, че тук броят на факторите е *2(k + 1)*, въпреки че броят на ограниченията е $k + 1$. Това се дължи на факта, че неограничената регресионна зависимост (за която ограничението заложено от нас като нулева хипотеза, не важи) е тази, която обхваща всички налични данни ($i \in [1, \ldots, n]$), докато ограничените регресии са две (по една за двете групи $[1, \ldots, m)$ и $[m, \ldots, n]$), всяка от които съдържа данни за оригиналните фактори и константа – т.е. $k + 1$ на брой ограничения.

За да вземем окончателно решение дали са налице достатъчно основания да отхвърлим нулевата хипотеза, трябва да сравним стойността на показателя на Chow с теоретичната стойност на F разпределение с параметри *(k + 1)* и *(n - 2(k + 1))* и предварително определено ниво на съгласие. В зависимост от полученото съотношение можем да дефинираме, че:

Ако показателят на $Chow > F(k+1, n-2(k+1))$, *то са налице основания да отхвърлим нулевата хипотеза и следователно НЕ МОЖЕМ да твърдим, че получените оценки на параметрите на регресионната зависимост са стабилни.*

Ако показателят на $Chow \leq F(k+1, n-2(k+1))$, *то НЕ са налице достатъчно основания да отхвърлим нулевата хипотеза и ПРИЕМАМЕ, че получените оценки са стабилни.*

Основният проблем, с който можем да се сблъскаме при практическата употреба на теста на Chow, е свързан с разделянето на наличните данни на групи – по какъв начин да бъде взето решението за разделяне и колко голяма да бъде всяка една група (т.е. $m$ = ?). Обикновено разделянето на данните се подчинява преди всичко на икономическата логика и особеностите на разглежданите финансови процеси. Ако се върнем към примера с модела за оценка на капиталови активи, то логично би било да изберем за отправна точка тези данни, които са свързани със значими промени в общите пазарни условия или индивидуалните характеристики на компанията (например, ако изследваме фирмата Nokia, моментът на продажбата на подразделението за мобилни апарати е добър кандидат за „разделителна линия“, тъй като променя цялостния профил на фирмата). По отношение на обема на всяка една от групите е логично той да бъде избран така, че да позволява да получим оценки за всеки един от показателите (в нашите примерни три регресионни зависимости това означава да разполагаме с набор от $m$ вектора, съдържащи данни за факторите в модела, така че $m \geq k+1$, тъй като в противен случай сравняването на сумата от квадратите на остатъчните стойности няма да носи същия смисъл и да бъде оправдано. Това последно условие неявно предполага, че методът на Chow е приложим само в случай че разполагаме с поне $2(k+1)$ вектора с данни за факторите в модела, с чиято помощ получаваме оценки за параметрите на регресионната зависимост. Когато това не е изпълнено или поради теоретични съображения искаме да направим разделението по групи, така че $m < k+1$ или $n-m < k+1$ се налага да използваме други средства за проверка на стабилността на получените оценки.

## 4.2. Проверка на прогнозиращите качества на модела

Проверката на прогнозиращите качества на модела е начин да погледнем към стабилността на оценките на параметрите на регресионната зависимост от друг ъгъл. При този подход отново разделяме наличните данни на две части, но без да сме задължени да спазваме изискването за минимален брой входни вектори във всяка една част ($m \geq k+1$). Вместо три регресионни зависимости, както е при теста на Chow, тук имаме само две:

$$\begin{aligned} y_i &= \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i, i \in [1, \ldots, n] \\ y_i &= \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i, i \in [1, \ldots, m) \end{aligned} \qquad \text{( VI—12 )}$$

където интервалът [1,...,m) включва по-голямата част от наличните данни. Останалите данни ([m,...,n]) се използват в качеството си на тест за това, до каква степен оценените параметри на регресионната зависимост могат да ги прогнозират. Ако прогнозиращите качества на изградения модел са добри, то можем да очакваме, че няма да има големи разлики между тестовите стойности и тези, получени с помощта на регресионната зависимост. Аналогично на теста на Chow и тук дефинираме като нулева хипотезата, че получените стойности за параметрите на регресията са стабилни. Решението дали има достатъчно основания да отхвърлим тази хипотеза можем да вземем, като сравним два показателя:

$$\begin{aligned} &\text{показател за прогнозиращите качества на модела} = \frac{RSS_{1..n} - RSS_{1..m}}{RSS_{1.m}} . \frac{m-(k+1)}{n-m} \\ &\text{теоретична стойност на } F \text{ разпределение} = F\big(n-m, m-(k+1)\big)\,, \end{aligned} \qquad \text{( VI—13 )}$$

като при това са възможни два резултата:

- Ако $\frac{RSS_{1..n}-RSS_{1..m}}{RSS_{1..m}}.\frac{m-(k+1)}{n-m} > F(n-m, m-(k+1))$, то можем да отхвърлим нулевата хипотеза, т.е. НЕ МОЖЕМ да твърдим, че получените оценки на параметрите на регресията са стабилни.

- Ако $\frac{RSS_{1..n}-RSS_{1..m}}{RSS_{1..m}}.\frac{m-(k+1)}{n-m} \leq F(n-m, m-(k+1))$, то няма достатъчно основания да отхвърлим нулевата хипотеза, ПРИЕМАМЕ, че получените оценки на параметрите на регресията са стабилни.

При практическото използване на проверката за прогнозиращите качества на модела трябва да имаме предвид, че е възможно да подберем „проверяващата“ съвкупност от данните по различен признак и по различен начин. Когато анализираме съвкупност от данни, които са подредени по хронологичен ред (вж. Част VII), обикновено се говори за два стандартни начина за избор – екстраполиране „напред“ и „назад“. В първия случай $m$ се избира по такъв начин, че данните за оценка на параметрите на втората регресионна зависимост [1,...,$m$) включват наличната информация от началото до първите m вектора, а останалите $n - m$ данни се използват за оценка на прогнозните качества на изграждания модел. Когато говорим за екстраполиране „назад“ спрямо целия отрязък от време, за който имаме данни, оформянето на интервалите е в обратен ред – $m$ се избира по такъв начин, че стойностите, използвани за проверка на качествата на модела, са тези в началото на хронологичния ред. Трябва да отбележим, че показателят за качествата на модела е приложим и за двата случая, тъй като той отчита броя на елементите в отделните групи (доколкото те се отразяват на параметрите на теоретичното F разпределение), а не тяхното разположение във времето.

### 4.3. Стратегии за разделяне на наличните данни на групи

В зависимост от избрания метод за проверка за стабилност са налице определени условия при избора на интервали и разделянето на наличните данни на части. Невинаги обаче тези изисквания са достатъчни или отговарят на смисъла на изследваните финансови процеси. Като резултат от това може да се окаже, че използваните тестове за стабилност не се прилагат ефективно или водят до грешни изводи. За да осигурим съвместимост между използваните принципи за разделяне на наличните данни и характеристиките на анализираните финансови и икономически процеси, можем да използваме различни помощни средства при определяне на границите на отделните групи от данни, както и да маркираме тези събития и елементи, които могат да доведат до нестабилност на получените оценки.

*Таблица 18. Подходи за анализ на наличните данни и разделянето им на групи*

| Подход | Приложимост |
|---|---|
| Анализ на графичното представяне на наличните данни. | Графичният анализ не претендира за висока точност, но позволява лесно и бързо да се идентифицират елементи или събития, които водят до „скокообразно“ (т.е. „революционно“ вместо „еволюционно“) изменение в наблюдаваните финансови процеси. Недостатък на този подход е, че той не ни дава представа дали тези резки изменения са единствените предпоставки, водещи до нестабилност в оценките (т.е. възможно е да има и други фактори, които да не са така ясно разпознаваеми на графиките).<br>Този подход може да се използва успешно както с теста на Chow, така и по отношение на оценка на прогнозните качества на модела. Обикновено подходящи за отправна точка при разделяне на данните са критични събития (например пазарни сривове) или важни промени в структурата на анализираните обекти. |
| Подход на Куанд (QLR подход) и използване на множество възможни разделения. | Когато не е възможно да бъде определен точният момент (респ. елемент), който води до качествена промяна в изследвания финансов процес или обект, но е възможно да бъде определен интервал, в който е настъпила промяната, може да използваме т.нар. QLR подход. При него се правят множество разделения на наличните данни (съобразно с определения „интересен“ интервал) и за всяко едно разделение се прилага тестът на Chow. При отделните тестове се сравняват получените стойности на F разпределението и се избира най-голямата от тях (за провеждане на крайния тест и обосновка на крайния извод). |
| Използване на допълнителни тестове като CUSUM и CUSUMSQ. | И двата теста са свързани с използването на сумата от остатъчните стойности (които в първия случай са нормализирани, а при CUSUMSQ се използват остатъчните стойности, повдигнати на квадрат), получени при множество на брой рекурсивни[16] проверки. |

Таблица 18 показва някои от най-популярните методи за анализ на наличните данни и избор на под интервали. В зависимост от конкретния изследван проблем или ситуация

[16] Рекурсивни означава, че последователно разширяваме интервала [1..m) докато той достигне [1..n].

можем да отдадем предпочитание на един или друг метод, като обикновено графичният подход е най-лесен за прилагане и последващ анализ.

### 4.4. Примерен анализа за стабилност на получените оценки

Ще демонстрираме как анализът на стабилност на получените оценки за параметрите на регресионната зависимост може да се използва на практика. За целта ще се опитаме да оценим бета коефициента на Google в контекста на модела за оценка на капиталови активи. Основните характеристики на използвания пример са описани в Таблица 19.

*Таблица 19. Примерно приложение на теста за Chow и прогнозиращите качества на регресионния модел*

| Параметър | Описание |
|---|---|
| Обясняваща променлива | Възвръщаемостта от акциите на Google минус безрисковата възвръщаемост, изчислена на месечна база като разлика в цените на затваряне на борсовата търговия: $$r_{GOOG} = \frac{P_t - P_{t-1}}{P_{t-1}}$$ $$y = r_{GOOG} - r_f$$ |
| Целева променлива | Премията във възвръщаемостта[17] на портфейл, следващ поведението на индекса Nasdaq Composite 100, изчислена на месечна база като разлика в цените на затваряне на борсовата търговия: $$r_{\wedge NDX} = \frac{I_t - I_{t-1}}{I_{t-1}}$$ $$x = r_{\wedge NDX} - r_f$$ |
| Безрискова възвръщаемост | Като показател за безрискова възвръщаемост е използвана доходността на 10–годишните държавни облигации на САЩ. Това не е единственият възможен избор, както и по принцип валидността на МОКА/индексните модели е дискусионна в контекста на съвременните кризи, но за целите на нашия пример е приемливо за използваме този показател. |
| Брой данни, използвани в анализа | 114 месечни данни<br>(обхващат период от 1.септ.2004 до 2.дек.2013 г.) |

[17] Т.е. възвръщаемостта на портфейла минус безрисковата възвръщаемост.

Регресионната зависимост, чиито параметри ще се опитаме да оценим ще има вида:

$$r_{GOOG} - r_f = \alpha + \beta\left(r_{NDX} - r_f\right) + \varepsilon .$$

Преди да пристъпим към оценка на параметрите, трябва да бъдем сигурни, че използваме данни, които са съвместими и не носят различен смисъл от този, който ние залагаме в тях. Така например едно от първите преобразувания, които трябва да направим, е да преобразуваме годишната възвръщаемост за държавните облигации в месечна, тъй като факторът и целевата променлива са изчислени на месечна основа.

С помощта на Excel, можем да направим оценка на показателите на регресионната зависимост, като резултатите са посочени в Таблица 20.

*Таблица 20. Резултати от оценка на параметрите с помощта на Excel*

| Показател | Стойност | Показател | Стойност |
|---|---|---|---|
| α | 0.015464 | α | 0 |
| β | 1.14319 | β | 1.184306 |
| Стандартна грешка | 0.086327 | Стандартна грешка | 0.087305 |
| $R^2$ | 0.334818 | $R^2$ | 0.348247 |
| Модифициран $R^2$ | 0.328825 | Модифициран $R^2$ | 0.339319 |

Автоматизираното прилагане на метода на най-малките квадрати ни позволява да направим две оценки без особена трудност, като в лявата част на Фигура 16 сме наложили едно допълнително ограничение − константата в регресионната зависимост да бъде 0 (т.е. за да отговорим по-точно на функционалната форма, налагана от МОКА и индексните модели). Поради ниската стойност на α и в двете части на таблицата получаваме близки резултати по отделните характеристики. Съответно в двата случая регресионната зависимост ще има вида:

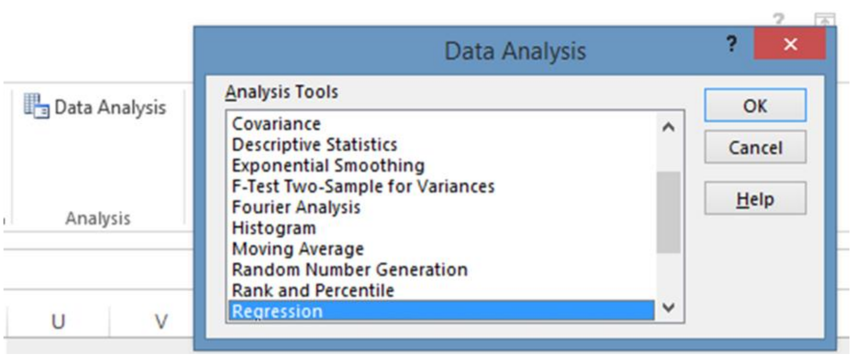
$$(r_{GOOG} - r_f)_i = 0.015464 + 1.14319(r_{^\wedge NDX} - r_f)_i + \varepsilon_i$$

и съответно при наложено условие за $\alpha = 0$ ще придобие следния облик:

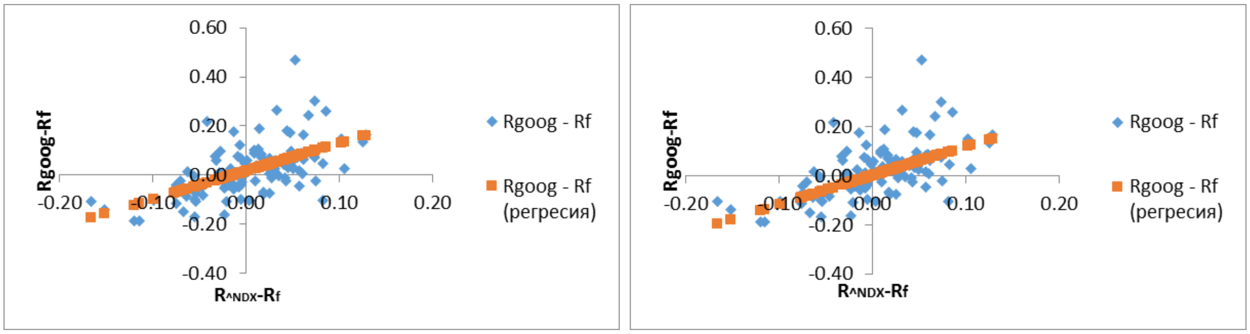
$$(r_{GOOG} - r_f)_i = 1.184306(r_{^\wedge NDX} - r_f)_i + \varepsilon_i \,.$$

## Метод на най-малките квадрати с Excel

Excel разполага с богат инструментариум от статистически функции, включително и възможността да прилага метода на най-малките квадрати. За да се възползваме от това, е необходимо да проверим, че допълнението (add-in) за анализ на данни е активирано. Ако това не е изпълнено, трябва да го активираме ръчно от менюто File->Options->Add-ins.



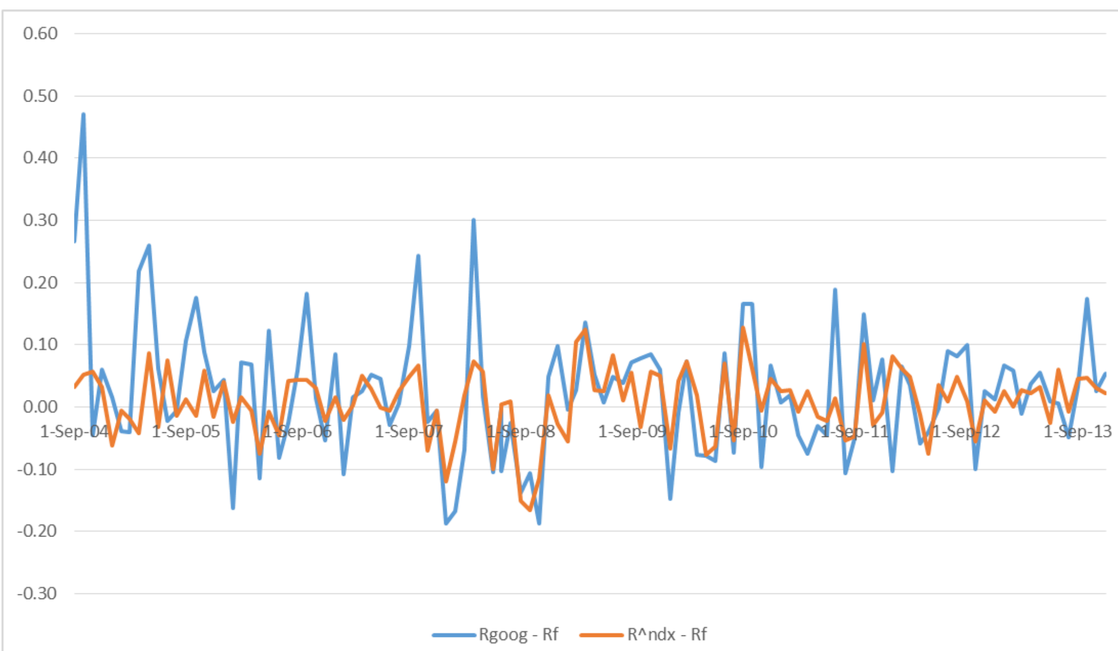

Функцията, която трябва да използваме за оценка на параметрите на регресионната зависимост, се нарича “Regression” и е достъпна като част от пакета за анализ на данни (Data Analysis Pack). Приложената снимка и начин за активиране са валидни за Excel 2013/2010.




*Фигура 16. Резултати от използването на метода на най-малките квадрати за оценка на параметрите на регресионната зависимост*

Оценките на параметрите на регресионната зависимост са получени с използването на всички налични данни. Все още стои въпросът дали в рамките на анализирания период не са настъпили някакви сериозни промени или изключителни събития, които да доведат до това, че получените резултати да описват добре възвръщаемостта на акциите в рамките на наличните данни, но да не са адекватни извън тях (например поради факта, че изключителното събитие вече не е налице или са се променили коренно икономическите условия).

За да проверим стабилността на получените оценки с теста на Chow и прогнозиращата сила на модела, ще използваме първо графиките на наличните данни спрямо времето – съобразно с предложения в първия ред на Таблица 18 подход.




*Фигура 17. Графика на измененията на фактора и целевата променлива във времето*

Това което можем да видим на пръв поглед от Фигура 17, е че като цяло възвръщаемостта на акциите на Google се изменя в по-големи граници от тази на портфейл, възпроизвеждащ пазарния индекс. Единствено от графичното представяне обаче не можем да посочим веднага специфични фирмени събития или промени, които да водят до резки разминавания между общите пазарни колебания и това, което се случва с акциите на фирмата.


*Фигура 18. Графика на доходността на 10−годишните държавни облигации на САЩ (на месечна база)*

Ако обърнем поглед към избрания показател, който да представя безрисковата възвръщаемост, можем да видим ясно последствията от глобалната финансова криза, разразила се през 2007 и 2008 година, и ефекта от действията на Федералния резерв. Въпреки че за официално начало на активната част от кризата се смята 9 август 2007 г., когато BNP Pariba спира тегленията от три хедж фонда, за нас е по-важно не формалното обявяване, а по-скоро ефектът който кризата е оказала върху средата, в която оперират фирмите. Поради тази причина, ще направим разделяне на данните на две части, смятано от месец ноември 2008 г., като това ще ни позволи да проверим дали ефектите от кризата са повлияли съществено върху съотношението риск/възвръщаемост за изследваната акция на Google. Избраната дата съвпада с активната намеса на пазара под формата на „количествени облекчения“ (Quantitative Easing) и едновременно с това е период, през който пазарът като цяло е свидетел на рязък спад (така например спадът на S&P 500 през ноември 2008 г. е 45% спрямо максималните му стойности от 2007 г.).

Общият обем на данни, с които разполагаме, е разположен в периода 1.септември.2004 г. до 2.декември.2013 г., или на месечна база това са 114 наблюдения, като всяко едно от тях включва данни за възвръщаемостта на Google, индекса и 10−годишните съкровищни бонове на САЩ. Времевите зависимости на данните няма да бъдат разглеждани самостоятелно за момента, поради което ще използваме времето само за да номерираме наблюденията от 1 до 114, започвайки от първото от тях. В съответствие с теста на Chow след избора на ноември 2008 за разделител ще имаме три регресионни зависимости:

$$(r_{GOOG} - r_f)_i = 0.015464 + 1.14319\left(r_{\wedge NDX} - r_f\right)_i + \varepsilon_i \,, i \in [1,\ldots,114]$$

$$(r_{GOOG} - r_f)_i = 0.032274092 + 1.14319\left(r_{\wedge NDX} - r_f\right)_i + \varepsilon_i \,, i \in [1,\ldots,49]$$

$$(r_{GOOG} - r_f)_i = 0.00409883 + 1.01812795\left(r_{\wedge NDX} - r_f\right)_i + \varepsilon_i \,, i \in [50,\ldots,114]$$

След пресмятане на съответните суми за квадратите на остатъчните стойности на регресионните зависимости ще получим:

$$RSS_{1..114} = 0.827211839$$

$$RSS_{1..49} = 0.563074821$$

$$RSS_{50..114} = 0.231884872$$

С така направените изчисления вече можем да пристъпим към проверка с помощта на показателя на Chow, който, ще припомним, е равен на:

$$\text{показател на } Chow = \frac{RSS_{1..114} - (RSS_{1..49} + RSS_{50..114})}{RSS_{1..49} + RSS_{50..114}} \cdot \frac{n - 2(k+1)}{k+1} \,,$$

и трябва да бъде сравнен с теоретичната стойност на $F(k+1, n-2(k+1))$. За конкретния пример ще имаме:

$$\text{показател на } Chow = \frac{0.827211839 - (0.563074821 + 0.231884872)}{0.563074821 + 0.231884872} \cdot \frac{114 - 2x2}{2} = 2.23139367$$

$$F(2, 114 - 2x2)_{5\%} = 3.07881949 \text{ [18]}.$$

Тъй като при избраното ниво на съгласие 5% (α=0.05) стойността на показателя на Chow е по-малка от стойността на F разпределение съответно с 2 и 110 степени на свобода, то

[18] Показателят може да бъде изчислен с помощта на Excel с функциите FINV или F.INV.RT .

можем да заключим, че *няма достатъчно основания да отхвърлим нулевата хипотеза и ПРИЕМАМЕ, че получените оценки са стабилни*.

Следващата стъпка в анализа е да проверим дали получените резултати имат икономически смисъл и отговарят на действителността. Безспорно глобалната финансова криза от 2007-2008 г. има сериозни последици както за компаниите, свързани с инвестиции в ценни книжа, така и за икономическото развитие на световната икономика като цяло. Получените от нас резултати също са в съзвучие с този факт – ние не установихме, че Google не е била под влиянието на финансовите турбуленции, а просто че няма достатъчно основание да твърдим, че проблемите са довели до фундаментална промяна в съотношението риск/възвръщаемост при акциите на компанията. Причините за това са както политиката на последователно реинвестиране на печалбите (което е индикатор, че ръководителите на компанията и инвеститорите са склонни да вярват във вътрешните източници на растеж), така и фактът, че рекламният бизнес на фирмата (който е основният източник на приходи по време на изследвания период) остава сравнително по-слабо засегнат от кризата.

Проверка за стабилността на получените оценки можем да направим и посредством оценка на прогнозиращата сила на модела. За целта избираме последните четири месеца (септември-декември 2013 г.) в ролята на показатели за изпробване на прогнозиращите качества на модела, а останалите наблюдения − за оценка на допълнителната регресия. В тази ситуация ще имаме две регресионни зависимости:

$$(r_{GOOG} - r_f)_i = 0.015464 + 1.14319\left(r_{^\wedge NDX} - r_f\right)_i + \varepsilon_i \, , i \in [1, \ldots, 114]$$

$$(r_{GOOG} - r_f)_j = 0.015129254 + 1.124309274\left(r_{^\wedge NDX} - r_f\right)_j + \varepsilon_i \, , j \in [1, \ldots, 109] \, .$$

След като разполагаме с двете регресии, може, да изчислим необходимите показатели, за да вземем решение относно нулевата хипотеза, която гласи, че получените оценки за параметрите на регресията са стабилни:

$$\text{показател за прогнозиращите качества на модела} = \frac{RSS_i - RSS_j}{RSS_j} . \frac{m - (k+1)}{n - m}$$
$$= \frac{0.827211839 - 0.81513559}{0.81513559} . \frac{109 - 2}{114 - 109} = 0.31704140$$

$$\text{теоретична стойност на } F \text{ разпределение} = F\big(n - m, m - (k+1)\big)$$
$$= F(114 - 109, 109 - 2)_{5\%} = 2.29923433 \, .$$

Крайният извод е аналогичен на този от теста на Chow, че не разполагаме с достатъчно основания, за да отхвърлим нулевата хипотеза − и ПРИЕМАМЕ, че получените оценки на параметрите на регресията са стабилни. Казано по друг начин − въз основа на наличната информация хипотезата, че получените оценки са стабилни, остава валидна.

## 5. Задачи и въпроси за размисъл

1. Какво ще направите, ако установите, че между два от факторите в модела е налице перфектна връзка?
2. Смятате ли, че мултиколинеарността ще се отрази на $R^2$ или преди всичко на значимостта на индивидуалните оценки на параметрите?
3. Какво бихте предпочели – да включите в модела променлива, която не е непосредствено значима за измененията на целевата променлива, или да пропуснете от модела променлива, която е значима за целевата величина? Обосновете ефектите в двата случая.
4. Обяснете каква е разликата в приложимостта на теста на Chow и проверката на прогнозиращите качества на регресионния модел? Кога бихте използвали втория подход?
5. Повторете проверката за стабилност на оценките на параметрите на регресионната зависимост от примера за Google, но този път използвайте регресия, при която по условие $\alpha = 0$. Какво ще се наложи да промените (освен изчисленията) при теста на Chow?

# VII. Въведение в анализа на времеви редове

В предишните части анализирахме различни аспекти от характеристиките и приложението на линейните регресионни модели. Освен използваната методика моделите, които изграждахме, имат една важна обща черта – те опитват да обяснят изменениата в една променлива с използването на няколко други показателя или фактора. Казано с професионални термини – те принадлежат към групата на т.нар. структурни модели. Тази група има широко приложение и се използва при анализа на различни икономически и финансови проблеми и особено, когато е необходимо и възможно да бъдат идентифицирани връзки между различните наблюдавани променливи. Съществуват обаче редица случаи, при които е или невъзможно да бъдат предварително идентифицирани факторите, влияещи на една променлива, или тяхното измерване е трудно за прилагане на практика. Обикновено невъзможността да дефинираме предварително кои са променливите, влияещи върху изследваните явления и процеси, може да произтича както от недостатъчното развитие на теорията, така и от съществуването на конкурентни теоретични възгледи, от които да не можем да изберем подходящ за нашия анализ. Практическите затруднения могат да бъдат от най-различно естество, като например неточност на измерванията на някои показатели или липса на данни с желаната продължителност (история) и честота. Във всички тези случаи е необходимо да предприемем допълнителни действия, за да отстраним възникналите трудности.

Когато възникналите проблеми не могат да бъдат надеждно преодолени, е възможно да опитаме да подходим към анализа на интересуващите ни явления и процеси по принципно различен начин – като използваме информация за наблюденията върху интересуващия ни процес и не търсим изрично кои са тези външни променливи, които му оказват влияние. Отказът от предварителното дефиниране на обясняващи променливи носи със себе си както предимства, така и недостатъци, като например:

- Дава ни възможност да се освободим от „оковите“ на предварително избран теоретичен модел.
- Не е необходимо да ограничаваме виждането си за анализираните финансови процеси само до определен набор от влияещи променливи. Когато анализираме данните сами по себе си, ние можем да уловим влиянието на много различни фактори, без дори да ги познаваме поотделно.
- Недостатък е, че е необходимо да ползваме други методи освен познатата ни вече линейна регресия. При това получените резултати не е задължително да ни дадат директно причинно-следствени връзки между различни икономически и финансови показатели – т.е. необходимо е допълнително усилие, за да оформим получените резултати като теоретичен модел (или да ги поставим в рамките на вече съществуващ такъв).

Когато използваме методи за иконометричен анализ, които използват миналите стойности на даден финансов показател с цел да изяснят неговите характеристики и свойства, говорим за анализ на времеви редове. Той намира широко приложение за изследване на различни проблеми и е една от важните съвкупности от инструменти, които трябва да познаваме.

## 1. Основни понятия, свързани с анализа на времеви редове

Преди да разгледаме кои точно алгоритми и техники се свързват с анализа на времеви редове, е необходимо да дадем по-тясна дефиниция на термини, които досега не сме срещали или сме използвали в различен контекст.

*Таблица 21. Основни понятия, свързани с анализа на времеви редове*

| Понятие | Дефиниция |
|---|---|
| Времеви ред | Последователност от данни, отнасящи се за определен финансов или икономически обект, които описват неговото състояние през N различни времеви контура (N момента или N периода, $N > 1$). Ще обозначаваме тези данни като $y_{t_i}$, където индексът $t_i$ показва времевия контур. Ще отбелязваме един такъв ред като: $$y(t) = \{y_{t_1}, y_{t_2}, \dots, y_{t_N}\}.$$ В зависимост от това дали стойностите на променливата y се отнасят до моменти, или до периоди говорим за моментни и хронологични (периодни) редове. Ние ще използваме основно периодни редове, а в случаите, когато разполагаме с моментни наблюдения, те могат предварително да бъдат преобразувани до хронологичен ред. |
| Процес | Въпреки че досега използвахме по-свободно понятието за процес, при анализа на времеви редове трябва да го поставим в по-тесни и строги рамки. А именно: *Под процес ще разбираме развитието на една величина или система във времето*. В случаите, когато става въпрос за развитие във времето на съвкупност от случайни величини, ще наричаме тези процеси *стохастични* или *случайни*. В зависимост от начина, по който отчитаме времето, различаваме *дискретни (времевият индекс може да взема дискретни стойности)* и *непрекъснати (времевият индекс може да взема всяка една стойност от определен интервал)* процеси, като предмет на нашето внимание в настоящата част ще бъдат първия тип. |
| Автоковариация | При разглеждането на структурните модели използвахме ковариацията (и съответно корелацията) като измерител за линейната зависимост на две променливи. Когато разглеждаме времеви редове обаче, ние разглеждаме само една променлива (т.е. тази, която |

| | |
|---|---|
| | изследваме) и поради тази причина е възможно да търсим зависимост само в рамките на нейните стойности през отделните времеви периоди. Автоковариацията измерва точно тази зависимост между стойностите на променливата в два различни периода и се дефинира като: $$C_{yy}(t_1, t_2) = E[(y_{t_1} - \mu_{t_1})(y_{t_2} - \mu_{t_2})]$$ Където $E(y_{t_i}) = \mu_{t_i}$ е очакваната стойност на изследваната величина в момента $t_i$. |
| Автокорелация | Подобно на връзката между ковариация и корелация, можем да дефинираме и връзка между автоковариация и автокорелация: $$R_{yy}(t_1, t_2) = \frac{E[(y_{t_1} - \mu_{t_1})(y_{t_2} - \mu_{t_2})]}{\sigma_{t_1}\sigma_{t_2}} = \frac{C_{yy}(t_1, t_2)}{\sigma_{t_1}\sigma_{t_2}},$$ където $\sigma_{t_i}$ е стандартното отклонение на изследваната променлива в момента $t_i$. Трябва да отбележим, че са възможни ситуации при които автокорелацията не може да бъде дефинирана (ако например $\sigma_{t_1}$ или $\sigma_{t_2}$са нула). |
| Лаг / Оператор за лаг | В анализа на времеви редове с оператор за лаг се дефинира действието при което от един елемент на реда $y_t$ получаваме предхождащата го стойност на показателя $y_{t-1}$. Операторът за лаг се бележи с L и се определя, като: $$L^p y_t = y_{t-p} .$$ Показателят $p$ може да има и отрицателни стойности, при което операторът за лаг води до получаването на следваща (а не предхождаща) стойност от времевия ред. |
| Функция на автоковариация | Функцията на автоковариация е: $$E[(y_{t_1} - \mu_{t_1})(y_{t_2} - \mu_{t_2})] = C_{yy}(t_1, t_2) = \gamma_{t_2 - t_1} .$$ Интересен частен случай е, когато времевите контури $t_1$ и $t_2$ съвпадат – тогава стойността на функцията ще съвпадне с квадрата на стандартното отклонение – $\sigma_{t_1}^2 = \sigma_{t_2}^2$. |
| Корелограма | Корелограма, или още известна като графика на автокорелацията (в англоезичните източници обозначавана като acf), е визуалното представяне на $R_{yy}(t_1, t_2)$ в зависимост от времето. При стационарните процеси, които ще дискутираме, корелограмата обикновено се визуализира спрямо различни стойности на лага. |

Изброените в Таблица 21 понятия са основата в представянето на анализа на времеви редове, което е причина да ги посочим изрично. Познаването им е необходимо условие, за да можем да разберем значението и характеристиките на прилаганите методи. За да получим обаче по-пълна представа защо отдаваме такова значение точно на тези концепции, е необходимо да разгледаме кои важни характеристики могат да отличат развитието на една променлива във времето (т.е. един процес) и кои са типовете процеси, които играят важна роля в анализа на времеви редове.

## 2. Стационарност на процесите и важни типове процеси

Въпреки че стационарните процеси под една или друга форма са ни познати както от финансите, така и от заобикалящата ни среда, е необходимо да им дадем по-точна дефиниция. Вероятно на внимателния читател му е направило впечатление, че при дефинирането на автоковариацията и автокорелацията използвахме индекси за средните величини ($\mu_t$) и стандартното отклонение ($\sigma_t$). Това показва, че ние допускаме възможността и двете величини да се променят във времето, а да не бъдат константи. Тъй като използваните методи (и съответно тяхната сложност) зависят от характеристиките на анализирания времеви ред, можем предварително да опростим нашата задача, като класифицираме начините, по които се развива във времето една променлива.

Стационарен процес е такъв тип развитие на една величина, при което вероятността стойностите да попаднат в определен интервал не се променя с времето. Записано формално, това означава, че ако функцията на кумулативната вероятност е F, то трябва да бъде изпълнено:

$$F\left(y_{t_1}, y_{t_2}, \dots, y_{t_k}\right) = F\left(y_{t_{1+i}}, y_{t_{2+i}}, \dots y_{t_{k+i}}\right), \text{за } \forall\, k \text{ и } i, \text{ имащи смисъл за реда.} \qquad (\text{ VII—1 })$$

Ако горното равенство е изпълнено наистина за всяка стойност на $k$ и $i$, то това ще доведе до две важни за анализа на такъв тип процеси следствия:

- функцията на кумулативната вероятност няма да зависи от времето, следователно, ако веднъж установим нейния вид, можем да я прилагаме за целия времеви ред;
- стойностите на средната величина (μ) и стандартното отклонение (σ) за признака у във всеки времеви контур ще бъдат константи и също няма да зависят от времето.

И двете следствия водят до значително опростяване при изследването на стационарни процеси и ще се връщаме към тях по-късно. Интуитивен пример за стационарен процес (който освен това отговаря и на изискването за дискретност във времето) е познатата ни от курса по математика схема на Бернули[19] – например редицата от резултати (ези/тура) от последователно хвърляне на една монета.

Освен посочената дефиниция се използва още една форма на стационарност, която, макар и с отслабени изисквания, позволява да се опрости анализът и намира важно практическо приложение:

[19] Схема на Бернули е последователност от случайно разпределени величини, всяка от които може да взема две стойности (1,0) с вероятности (p, 1−p).

- слаба форма на стационарност, при която е необходимо да бъдат изпълнени само изискванията за константна стойност на средната величина (μ) и стандартното отклонение (σ).

Освен че разширява обхвата на процесите, които могат да бъдат наречени стационарни, тази форма има още едно важно следствие, а именно − че автоковариацията зависи само от лага във времето и е функция само на този лаг. Това е резултат от използването на функцията на автоковариация от Таблица 21 при условие на константна средна величина:

$$E\left[\left(y_{t_{i_1}} - \mu_{t_{i_1}}\right)\left(y_{t_{i_2}} - \mu_{t_{i_2}}\right)\right] = E\left[\left(y_{t_{i_1}} - \mu\right)\left(y_{t_{i_2}} - \mu\right)\right] = C_{yy}\left(t_{i_1} - t_{i_2}\right). \qquad \text{( VII—2 )}$$

Обърнете внимание, че в общата дефиниция на ковариацията от Таблица 21 тя беше функция от две променливи − $t_{i_1}$и $t_{i_2}$, докато при слабата форма на стационарност тя е функция само на тяхната разлика. Тъй като стандартното отклонение също е константа от връзката:

$$R_{yy}\left(t_{i_1}, t_{i_2}\right) = \frac{C_{yy}\left(t_{i_1}, t_{i_2}\right)}{\sigma_{t_{i_1}}\sigma_{t_{i_2}}} = \frac{C_{yy}\left(t_{i_1} - t_{i_2}\right)}{\sigma^2}, \qquad \text{( VII—3 )}$$

следва, че автокорелацията също е функция само на лага. Този факт ще използваме по-късно при съкратено записване на операциите с процеси, отговарящи на условията на слабата форма на стационарност. В редица текстове като синоним на слабата форма на стационарност се използва „стационарност от втори ред“, което може да доведе до объркване, тъй като терминология за „втори ред“ се използва в различен контекст при описание на случайни процеси. При анализа на времеви редове ще използваме термина за „слаба форма на стационарност“ или ще се позоваваме на конкретните изисквания, за да избегнем проблеми с наименуването.

Защо стационарните процеси са важни за анализа на финансови проблеми? Ще се опитваме да дадем отговор на този въпрос с всеки един от примерите в нашето изложение. Но като обобщение можем да кажем, че ако един процес е стационарен, то това може да се тълкува, че в основата му стоят едни и същи движещи сили – т.е. посредством анализа на времеви редове можем да се надяваме да установим кои са те и да използваме практически получените резултати. Например, ако цената на една ценна книга следва стационарен процес (за съжаление, на практика това не е толкова често срещано, колкото ни се иска), то можем да търсим неговите характеристики и оттам да установим кое всъщност определя цената на ценната книга и кара тази цена да се променя.

## 2.1. Бял шум – пример за случаен процес

Основната ни цел при изучаването и употребата на анализа на времеви редове ще бъде да установим дали на базата на наличните данни и техните характеристики можем да изградим модел, който да описва изследваните финансови проблеми. Преди да представим някои от методите и инструментите, които ни помагат да намерим решение, е добре да дефинираме една отправна точка – а именно, процес, за който предварително е ясно, че липса връзка между отделните стойности и няма скрита вътрешна структура. Стойностите на променливата, следваща този процес, ще бъдат случайни и ще ни послужат като основа върху която да стъпим при последващите си разсъждения. Важен процес, който отговаря на тези условия и е широко използван, е т.нар. бял шум. Както подсказва името му, той може да бъде представен и в звукова форма като сигнал (с фиксирана мощност и ширина), а любознателният читател може да намери подходящи аудио файлове онлайн. (Ако нямате достъп до мрежата, може да настроите своето радио на неизползвана честота. Ако например настоящият текст не ви е приспал, то генератор на бял шум с подходяща мощност със сигурност ще има по-добър ефект.)

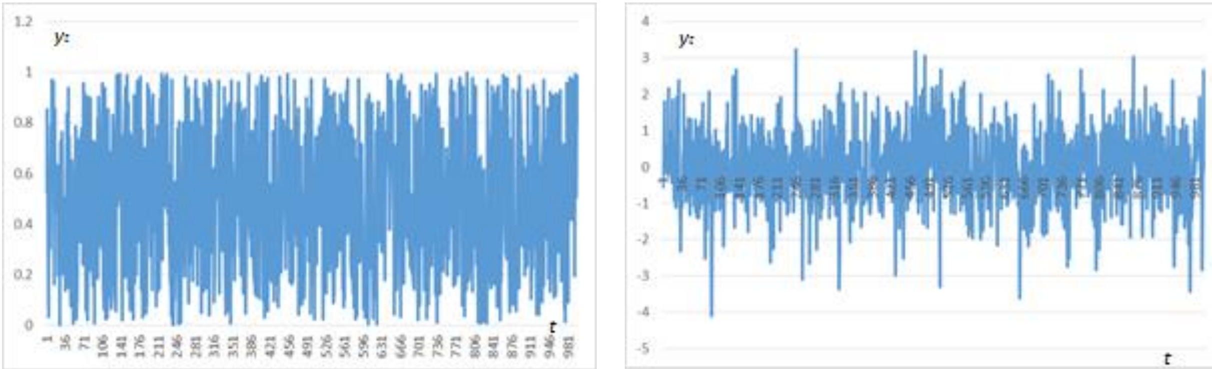
За да бъде класифициран един процес (респ. стойностите на една променлива във времето) като бял шум, то трябва да бъдат изпълнени следните три условия:

- да имаме константна средна стойност, т.е. $E\left(y_{t_i}\right) = \mu$;
- да имаме константна стойност на стандартното отклонение, т.е. $E[\left(y_{t_i} - \mu\right)\left(y_{t_i} - \mu\right)] = \sigma^2$;
- автоковариацията да бъде нула, освен за лаг 0 (когато, както помним тя ще бъде равна на квадрата на стандартното отклонение), т.е. функционалната зависимост за $C_{yy}(t_{i_1}, t_{i_2})$ ще бъде:

$$C_{yy}\left(t_{i_1}, t_{i_2}\right) = \begin{cases} 1, & t_{i_1} = t_{i_2} \\ 0, & t_{i_1} \neq t_{i_2} \end{cases} . \qquad \text{( VII—4 )}$$

Както се вижда от условията, всеки един процес, който отговаря на изискванията за бял шум, едновременно с това е и стационарен. Тази особеност определя и част от важните приложения на този тип процеси.

Трябва изрично да отбележим, че под наименованието „бял шум“ се крият различни процеси, които обаче отговарят на горните три изисквания. Така например променливата величина, чиито стойности във времето наблюдаваме, може да има различно разпределение. Фигура 19 представя графично 1000 генерирани стойности на променлива, която следва равномерно и стандартно нормално (със средна стойност 0 и стандартно отклонение 1) разпределение.




*Фигура 19. Генериран бял шум с равномерно (вляво) и стандартно нормално (вдясно) разпределени стойности на изследваната величина*

Ако е известно, че един процес отговаря на изискванията за бял шум, то ние можем да се възползваме от вече дефинираните общи характеристики и да опростим нашия анализ. Ако обаче разполагаме само с данните от времевия ред и желаем да проверим дали наистина автоковариацията между отделните данни е нула, можем да използваме следните методи:

- Да изчислим (въз основа на наличните данни) всички възможни коефициенти на автокорелация и да проверим дали те са статистически значими

На практика този метод се свежда до статистическа проверка на хипотези за всеки един автокорелационен коефициент. Ще демонстрираме това с пример върху вече разглежданата възвръщаемост на акциите на Google (изчисленията са направени на база 114-те месечни данни от предходния пример).

*Таблица 22. Пример за проверка за значимост на индивидуални автокорелационни коефициенти*

| **Период/Лаг** | **1** | **2** | **3** | **4** | **5** | **6** |
|---|---|---|---|---|---|---|
| Възвръщаемост | 0.0532 | 0.0259 | 0.1744 | 0.0319 | −0.0483 | 0.0063 |
| Автокорелация ( лаг = период ) | 0.14003 | −0.10058 | −0.07132 | −0.08839 | 0.03847 | 0.07445 |

Ако възвръщаемостите използвани за изчисляване на данните от таблицата, са резултат от процес, тип „бял шум“, и $y_t$ следва нормално разпределение, то тогава, както посочва Brooks [3], разпределението на автокорелационните коефициенти би трябвало да се приближава до нормално разпределение със средна стойност 0 и дисперсия $\frac{1}{N}$ ($N$ е броят на данните във времевия ред). Това ни позволява да направим бърза проверка за значимостта на всеки един от коефициентите, като имаме предвид, че 95% от площта, заключена от кривата на нормалното разпределение се намират приблизително в интервала $0 \pm 2\sqrt{\frac{1}{N}}$ (ако трябва да бъдем точни, интервалът е $0 \pm 1.95996\sqrt{\frac{1}{N}}$). За конкретния пример границите, при които ще имаме основание да отхвърлим хипотезата, че автокорелационният коефициент е нула, ще бъдат:

$$\left(0 - 1.95966\frac{1}{\sqrt{114}}; 0 + 1.95966\frac{1}{\sqrt{114}}\right) = (-\,0.183539;\ +0.183539)$$

Тъй като нито един от изчислените в Таблица 22 коефициенти не попада извън този интервал, то нямаме основание да отхвърлим нулевата хипотеза, че всички автокорелационни коефициенти са нула.

- Да използваме методите за статистическа проверка на хипотези, като дефинираме нашата нулева хипотеза, че „всички коефициенти на автокорелация са равни на нула“

За разлика от първия подход тук не е нужно да изчисляваме предварително всеки един корелационен коефициент и след това да разглеждаме дали той попада в избрания интервал на значимост. Въпреки че е достатъчно само един от коефициентите да бъде статистически различен от нула, за да отхвърлим цялата нулева хипотеза, то ако трябва да анализираме много на брой елементи във времевия ред, ще трябва да направим много индивидуални тестове. Вместо това може, да направим проверка на общата хипотеза, че всички автокорелационни коефициенти са нулеви. Съществуват няколко различни теста, с помощта на които можем да изследваме така поставената хипотеза. За горния пример ще използваме Q-показателя на Ljung-Box [17], който сравняваме с теоретичната стойност на $\chi^2$ разпределение със степени на свобода $q$ (където $q$ е максималният лаг в анализирания случай):

$$Q = N(N-2)\sum_{j=1}^{q-1}\frac{R_{yy}\left(t_{i_q}, t_{i_{q-j}}\right)^2}{(N-j)} \sim \chi^2(q)\,. \qquad \text{( VII—5 )}$$

При данните от Таблица 22 стойностите на показателя Q и теоретичното $\chi^2$ разпределение с шест степени на свобода ще бъдат:

$$Q = 114(114-2)\left(\frac{0.14003^2}{114-1} + \frac{-0.10058^2}{114-2} + \frac{-0.07132^2}{114-3} + \frac{-0.08839^2}{114-4} + \frac{0.03847^2}{114-5} + \frac{0.07445^2}{114-6}\right) = 5.689419;$$

$\chi^2(6)$ за $\alpha = 5\%$ ще бъде $CHISQ.INV(0.95, 6) = 12.59159$.

Тъй като стойността на Q е по-малка от теоретичната стойност на $\chi^2$ разпределение с шест степени на свобода (пресметната при ниво на съгласие от 5%), то нямаме достатъчно основание да отхвърлим нулевата хипотеза.

## 2.2. Процеси с движещи се средни

Процесите с движещи се средни, или както още се наричат, плъзгащи се средни, намират широко приложение в техническия анализ на финансовите пазари, както и при изследването на широк спектър от икономически показатели. Освен всичко друго те са и относително прости за представяне в математическа форма, което ги прави подходящи за редица изследвания. Тук ще изложим накратко основните характеристики на процесите, които в основата си включват плъзгащи се средни, както и ще коментираме техните основни характеристики.

Движещите се средни са метод за анализ, при който на базата на първоначалната информация се изгражда последователност от средни величини, всяка една от които включва определен сегмент от първоначалните данни. От горната дефиниция следва, че е възможно да имаме различни начини, по които се изчислява средната величина, а следователно − различни типове плъзгащи се средни (които споделят едно и също име, но имат различни характеристики). При изследването на времеви редове и икономически проблеми изобщо се използват най-често три изчислителни метода:

- Обикновени плъзгащи се средни

Обикновените плъзгащи се средни се изчисляват, като сумата от първоначалните данни се раздели на техния брой:

$$\mathrm{M}_{t,n} = \frac{y_{t_n}+y_{t_{n-1}}+y_{t_{n-2}}+\cdots+y_{t_1}}{n}, \qquad 1 < n < N . \qquad \text{( VII—6 )}$$

Този подход е „интуитивен“ и най-лесен за възприемане, но има един съществен недостатък – отдава едно и също значение на всички стойности, участващи в израза за $\mathrm{M}_{t,n}$. Когато става въпрос за анализ на времеви редове, тази характеристика води до това, че всички стойности, независимо през кой времеви контур са отчетени, имат едно и също значение. В условията на една динамична икономическа обстановка подобно неявно допускане невинаги е оправдано и може да доведе до получаването на неточни резултати. Отделен въпрос, който стои пред приложението на плъзгащата се средна, е „прозорец“ с каква точно дължина да се използва (т.е. колко да бъде стойността на *n* от горната формула). Например, ако изберем твърде голяма или твърде малка стойност за *n*, рискуваме да пропуснем (като ги „изгладим“ или те се загубят сред множеството други стойности) или

„разкъсаме“ (т.е. те не могат да се проявят изцяло) част от характеристиките на изследваните проблеми.

- Претеглени плъзгащи се средни

Претеглените плъзгащи се средни ни предоставят начин да коригираме еднаквата тежест на данните, използвани при изчисляването на средната величина. В зависимост от избора на теглови коефициенти можем да моделираме различни проблеми, като отдаваме по-голямо значение на „старите“ или „новите“ (по-скорошните) стойности на времевия ред. По този начин формулата за изчисляване на претеглената средна ще бъде:

$$M_{t,n}^{MA} = \frac{\theta_0 y_{t_n} + \theta_1 y_{t_{n-1}} + \theta_2 y_{t_{n-2}} + \cdots + \theta_{(n-1)} y_{t_1}}{\theta_0 + \theta_1 + \theta_2 + \cdots + \theta_{(n-1)}}, \quad 1 < n < N. \qquad (\text{VII—7})$$

Обикновената средна по този начин се явява частен случай на претеглената средна, при която всеки един от тегловите коефициенти $\theta_{t_i}$ е равен на 1. Начинът, по който избираме теглата на отделните стойности, е от ключово значение за качеството на постигнатите резултати и би трябвало да отразява теоретичните виждания за изследваните финансови проблеми. Във връзка с широкото използване на претеглените плъзгащи се средни за анализ на финансовите пазари са се наложили и някои „стандартни“ начини за определяне на значимостта на отделните елементи, като например използването на проста аритметична прогресия:

$$M_{t,n}^{MA} = \frac{n y_{t_n} + (n-1) y_{t_{n-1}} + (n-2) y_{t_{n-2}} + \cdots + y_{t_1}}{n + (n-1) + (n-2) + \cdots + 1}. \qquad (\text{VII—8})$$

Независимо дали използваме „готов“ вариант за изчисляване на средната, или избираме тегловите коефициенти за конкретния случай, добре обоснованият анализ трябва да съдържа аргументация за техните стойности.

- Експоненциални плъзгащи се средни

Експоненциалните плъзгащи се средни са специален клас от претеглени плъзгащи се средни, при които тегловите коефициенти намаляват по експоненциален закон, като могат да бъдат много малки, но не и нулеви. Експоненциалните плъзгащи се средни са свързани с предложената от Brows [18] и Holt [19] рекурсивна форма за определяне на $\mathrm{M}_{t,n}$:

$$M_{t,n}^{EXP} = \alpha y_{t_n} + (1-\alpha) M_{t_{n-1},n}^{EXP}, \text{където } \alpha \in (0,1), \qquad 1 < n < N-1. \quad (\text{VII—9})$$

Т.е., за да определим една експоненциална плъзгаща се средна, ние трябва да знаем стойността на средната (определена по същия метод) за предходния период. Интервалът за стойността $\alpha$ неслучайно е отворен, тъй като, ако $\alpha = 0$ то и средната величина ще бъде равна на нула, а ако $\alpha = 1$, то тогава средната величина ще бъде просто сума от стойностите на елементите от изследваната съвкупност от данни. Експоненциалните плъзгащи се средни намират широко приложение при анализа на финансови проблеми заради двете си основни характеристики:

- поради рекурсивната форма на задаването си всяка една от изчислените средни величини включва всички данни до текущия момент, като това е едно възможно решение на проблема с избора на големина на „прозореца“ при плъзгащите средни;
- тъй като тегловите коефициенти намаляват експоненциално, то „старите“ данни в един времеви ред ще имат относително по-ниска тежест спрямо „новите“ данни, което позволява да преодолеем един от проблемите при обикновените движещи се средни.

След като разгледахме общоприетите начини за изчисляване на движещи се средни, можем да анализираме как те могат да ни бъдат от полза при изследване на конкретни икономически проблеми. Процес с движещи се средни се нарича развитието във времето на една променлива, чиито стойности се определят като линейна комбинация от стойностите на случайна променлива – „бял шум“:

$$y_{t_n} = c + \theta_0 v_{t_n} + \theta_1 v_{t_{n-1}} + \cdots + \theta_q v_{t_{n-q}}\,, \qquad (\text{VII—10})$$

където $v_t$ отговаря на изискванията за „бял шум“ с константни средна и стандартно отклонение, а именно:

$$\mathrm{E}(v_{t_n}) = E(v_{t_{n-1}}) = \cdots = \mathrm{E}\left(v_{t_{n-q}}\right) = 0 \qquad (\text{VII—11})$$

$$E\left[\left(v_{t_n} - E(v_{t_n})\right)\left(v_{t_n} - E(v_{t_n})\right)\right] = \sigma_v{}^2. \qquad (\text{VII—12})$$

Освен с параметрите на случайната компонента, процесите с плъзгаща се средна се характеризират и с големината на „прозореца“ или в нашия случай – $n$, което дава начало и на съкратеното им обозначаване като MA($n$)[20].

Трябва да отбележим, че дори и когато средната стойност на $v_{t_n}$ не е нула, ние бихме могли да използваме този процес като основа за получаване на стойностите на $y_{t_n}$. Нека да предположим, че имаме един бял шум $v'_{t_n}$ с ненулева средна, т.е.:

$$\mathrm{E}(v'_{t_n}) = E(v'_{t_{n-1}}) = \cdots = E\left(v'_{t_{n-q}}\right) = \mathrm{C} \qquad (\text{VII—13})$$

[20] MA е от съкращението на moving average.

$$E\left[\left(v'_{t_n} - E\left(v'_{t_n}\right)\right)\left(v'_{t_n} - E\left(v'_{t_n}\right)\right)\right] = \sigma_{v'}{'^2}. \qquad ( \text{VII—14} )$$

Тогава на всяка една стойност на променливата $v'_{t_n}$ ние можем да съпоставим една друга променлива $v_{t_n} = v'_{t_n} - \mathrm{C}$, при което средната стойност вече ще бъде:

$$\mathrm{E}\left(v_{t_n}\right) = E\left(v'_{t_n} - \mathrm{C}\right) = E\left(v_{t_n}\right) - \mathrm{C} = \mathrm{C} - \mathrm{C} = 0\,. \qquad ( \text{VII—15} )$$

Разбира се, това ще се отрази и върху числовата стойност на стандартното отклонение, но в нашите допускания ние искаме то да бъде просто константа. Много често в различни учебници по иконометрия същата последователност от действия се прилага и към процеса с движещи се средни, но вече с цел да елиминираме константата *c* от ( VII—10 ):

$$y'_{t_n} = y_{t_n} - \mathrm{c}\,. \qquad ( \text{VII—16} )$$

Това позволява да се използва съкратен запис на процеса с плъзгащи се средни, който след отпадането на константата ще бъде:

$$y'_{t_n} = \theta_0 v_{t_n} + \theta_1 v_{t_{n-1}} + \cdots + \theta_q v_{t_{n-q}} = \sum_{j=0}^{q} \theta_j L^j v_{t_n}\,, \qquad ( \text{VII—17} )$$

където L е оператор за лаг и $L^j v_{t_n} = v_{t_{n-j}}$. Разбира се, това е само различна форма, която позволява по-кратко да се опишат някои операции и преобразувания на основния израз за процеса $y'_{t_n}$. Смисълът и икономическата логика не се променят от начина, по който ги формализираме, но за по-голяма яснота ще използваме пълния запис. По-важно е да опишем характеристиките на процесите, базирани на плъзгащи се средни и по какъв начин те могат да се използват за практически цели. При описанието им ще използваме първоначалната форма:

$$y_{t_n} = c + \theta_0 v_{t_n} + \theta_1 v_{t_{n-1}} + \cdots + \theta_q v_{t_{n-q}}\,. \qquad ( \text{VII—18} )$$

Основните характеристики на процесите с движещи се средни, които трябва да познаваме, са:

- Средна стойност на $y_{t_n}$

За да намерим израз за средната стойност на $y_{t_n}$, нека предположим, че *y* зависи само от текущата и предишната стойности на случайната променлива $v_{t_n}$. Тогава:

$$y_{t_n} = c + \theta_0 v_{t_n} + \theta_1 v_{t_{n-1}}\,, \qquad ( \text{VII—19} )$$

а очакваната стойност ще бъде равна на:

$$\begin{aligned} \mathrm{E}(y_{t_n}) &= \mathrm{E}\big(c + \theta_0 v_{t_n} + \theta_1 v_{t_{n-1}}\big) = \mathrm{E}(\mathrm{c}) + \mathrm{E}(\theta_0 v_{t_n}) + \mathrm{E}\big(\theta_1 v_{t_{n-1}}\big) \\ &= \mathrm{c} + \theta_0 \mathrm{E}\big(v_{t_n}\big) + \theta_1 \mathrm{E}\big(v_{t_{n-1}}\big) = \mathrm{c}\,. \end{aligned} \qquad \text{( VII—20 )}$$

Горният резултат е следствие от факта, че по дефиниция променливата $v_{t_n}$ отговаря на условията за бял шум и има средна стойност нула, т.е. $\mathrm{E}\big(v_{t_n}\big) = E\big(v_{t_{n-1}}\big) = 0$. Полученият израз за средната стойност ще бъде валиден и в случаите, когато участват повече от две стойности на $v_{t_n}$, което означава, че за такъв тип процес с плъзгащи се средни можем да посочим, че:

$$E\big(y_{t_n}\big) = c\,. \qquad \text{( VII—21 )}$$

- Стандартно отклонение на $y_{t_n}$

За да намерим израз за стандартното отклонение на $y_{t_n}$, нека отново предположим, че $y$ зависи само от текущата и предишната стойности на случайната променлива $v_{t_n}$. Това означава, че тръгваме от същия частен случай, както при определяне на средната стойност.

Дисперсията по дефиниция е:

$$\sigma_{t_n}^2 = E\left[\Big(y_{t_n} - E\big(y_{t_n}\big)\Big)^2\right]. \qquad \text{( VII—22 )}$$

Като имаме предвид, че $E\big(y_{t_n}\big) = c$, и заместим израза за $y_t$, ще получим, че:

$$\begin{aligned} \sigma_{t_n}^2 &= E\big[\big(c + \theta_0 v_{t_n} + \theta_1 v_{t_{n-1}} - \mathrm{c}\big)\big(c + \theta_0 v_{t_n} + \theta_1 v_{t_{n-1}} - \mathrm{c}\big)\big] = \\ &= \mathrm{E}\big[\big(\theta_0 v_{t_n} + \theta_1 v_{t_{n-1}}\big)\big(\theta_0 v_{t_n} + \theta_1 v_{t_{n-1}}\big)\big]\,. \end{aligned} \qquad \text{( VII—23 )}$$

Като разкрием скобите, в последния израз, ще получим:

$$\begin{aligned} \sigma_{t_n}^2 &= \mathrm{E}\big[{\theta_0}^2 {v_{t_n}}^2 + \theta_0 v_{t_n} \theta_1 v_{t_{n-1}} + \theta_1 v_{t_{n-1}} \theta_0 v_{t_n} + {\theta_1}^2 {v_{t_{n-1}}}^2\big)\big] = \\ &= {\theta_0}^2 E\big({v_{t_n}}^2\big) + \theta_0 \theta_1 E\big(v_{t_n} v_{t_{n-1}}\big) + \theta_1 \theta_0 E\big(v_{t_{n-1}} v_{t_n}\big) \\ &\quad + {\theta_1}^2 E({v_{t_{n-1}}}^2) \end{aligned} \qquad \text{( VII—24 )}$$

За да опростим още повече горния израз, нека да си припомним още две от условията при които съществува разглежданият процес за $y_{t_n}$:

$$\mathrm{E}\big(v_{t_n}\big) = \mathrm{E}\big(v_{t_{n-1}}\big) = 0, \text{следователно}\ E\big({v_{t_n}}^2\big) = \mathrm{E}\big[\big(v_{t_n} - 0\big)\big(v_{t_n} - 0\big)\big] = {\sigma_v}^2$$

$$cov\big(v_{t_n}, v_{t_{n-1}}\big) = E\left[\Big(v_{t_n} - \mathrm{E}\big(v_{t_n}\big)\Big)\Big(v_{t_{n-1}} - \mathrm{E}\big(v_{t_{n-1}}\big)\Big)\right] = E\big(v_{t_n} v_{t_{n-1}}\big) = 0\,.$$

Втората зависимост идва от дефиницията на бял шум като такъв процес, при който автоковариацията е нула (освен за лаг 0) – т.е. няма зависимост между отделните стойности на променливата във времето.

Като заместим двете условия в израза за $\sigma_{t_n}^2$, ще получим окончателно, че за MA(1) процес:

$$\begin{aligned}\sigma_{t_n}^2 = \mathrm{E}[{\theta_0}^2{v_{t_n}}^2 + \theta_0 v_{t_n}\theta_1 v_{t_{n-1}} + \theta_1 v_{t_{n-1}}\theta_0 v_{t_n} + {\theta_1}^2{v_{t_{n-1}}}^2] =\\ = {\theta_0}^2 E\left({v_{t_n}}^2\right) + \theta_0\theta_1 E\left(v_{t_n}v_{t_{n-1}}\right) + \theta_1\theta_0 E\left(v_{t_{n-1}}v_{t_n}\right) + {\theta_1}^2 E\left({v_{t_{n-1}}}^2\right)\\ = {\theta_0}^2{\sigma_v}^2 + \theta_0\theta_1 .0 + \theta_1\theta_0 .0 + {\theta_1}^2{\sigma_v}^2\end{aligned} \qquad (\text{ VII—25 })$$

$$\sigma_{t_n}^2 = \left({\theta_0}^2 + {\theta_1}^2\right){\sigma_v}^2. \qquad (\text{ VII—26 })$$

В случай че имаме по-голям прозорец ($n$) за процеса, можем да запишем общия вид на израза за $\sigma_{t_n}^2$:

$$\sigma_{t_n}^2 = \left({\theta_0}^2 + {\theta_1}^2 + \cdots + {\theta_n}^2\right){\sigma_v}^2. \qquad (\text{ VII—27 })$$

- Автоковариация и автокорелация на $y_{t_n}$

При определяне на автоковариацията вече имаме една от стъпките – при стандартното отклонение ние практически направихме необходимото за оценка на автоковариация с лаг 0. Нека накратко проследим действията за определяне на този показател, но с лаг 1:

$$\begin{aligned}\gamma_{t_n - t_{n-1}} = \gamma_1 = E\left[\left(y_{t_n} - E\left(y_{t_n}\right)\right)\left(y_{t_{n-1}} - E\left(y_{t_{n-1}}\right)\right)\right] =\\ = E\left[\left(y_{t_n} - c\right)\left(y_{t_{n-1}} - c\right)\right] =\\ = E\left[\left(c + \theta_0 v_{t_n} + \theta_1 v_{t_{n-1}} - \mathrm{c}\right)\left(c + \theta_0 v_{t_{n-1}} + \theta_1 v_{t_{n-2}} - \mathrm{c}\right)\right] =\\ = \mathrm{E}\left[\left(\theta_0 v_{t_n} + \theta_1 v_{t_{n-1}}\right)\left(\theta_0 v_{t_{n-1}} + \theta_1 v_{t_{n-2}}\right)\right].\end{aligned} \qquad (\text{ VII—28 })$$

След като разкрием скобите и като отчетем, че променливата $v_{t_n}$ отговаря на изискванията за бял шум, получаваме:

$$\begin{aligned}\gamma_1 = \mathrm{E}\left[\theta_0 v_{t_n}\theta_0 v_{t_{n-1}} + \theta_0 v_{t_n}\theta_1 v_{t_{n-2}} + \theta_1 v_{t_{n-1}}\theta_0 v_{t_{n-1}} + \theta_1 v_{t_{n-1}}\theta_1 v_{t_{n-2}}\right] =\\ = {\theta_0}^2 E\left(v_{t_n}v_{t_{n-1}}\right) + \theta_0\theta_1 E\left(v_{t_n}v_{t_{n-2}}\right) + \theta_0\theta_1 E\left({v_{t_{n-1}}}^2\right)\\ + {\theta_1}^2 E\left(v_{t_{n-1}}v_{t_{n-2}}\right) = \theta_0\theta_1{\sigma_v}^2\end{aligned} \qquad (\text{ VII—29 })$$

Като отчетем връзката между автокорелация и автоковариация, която посочихме в Таблица 21, ще получим:

$$R_{yy}(t_n, t_{n-1}) = \frac{\gamma_1}{\sigma_{t_n}^2} = \frac{\theta_0\theta_1{\sigma_v}^2}{\left({\theta_0}^2 + {\theta_1}^2\right){\sigma_v}^2} = \frac{\theta_0\theta_1}{\left({\theta_0}^2 + {\theta_1}^2\right)}. \qquad (\text{ VII—30 })$$

Трябва да отбележим, че в случаите, когато лагът на тези коефициенти е по-голям от лага, заложен в изследвания процес, то тогава и стойността на автоковариацията, и коефициентът на автокорелация е 0. Ако в горния пример търсим автоковариация с лаг 2 (или по-голям), то резултатът ще бъде 0 (тъй като в нашия случай $q=1$ и $y_{t_n}$ зависи само от $v_{t_n}$ и $v_{t_{n-1}}$ (лаг 1)).

## 2.3. Авторегресивни процеси

Авторегресивните процеси описват измененията на една променлива във времето, при условие че стойностите ѝ зависят единствено и само от:

- линейна комбинация от предходните стойности на променливата;
- случаен фактор, който има характеристиките на процес от типа бял шум.

Формално изразено, горното твърдение означава, че променливата $y_{t_n}$, която следва авторегресивен процес от ред $q$ (т.е. AR(q)), ще има вида:

$$y_{t_n} = \mathrm{c} + \rho_1 y_{t_{n-1}} + \rho_2 y_{t_{n-2}} + \cdots + \rho_q y_{t_{n-q}} + \varepsilon_{t_n} \,, \qquad \text{( VII—31 )}$$

където $\rho_i$ са коефициентите, с които се измерва приносът на всяка една от предходните стойности за формирането на стойността на променливата в момента $t_n$, а $\varepsilon_{t_n}$ изразява ефекта от случайния фактор.

Основният израз на авторегресивните процеси прилича на линейните регресионни уравнения, с които работихме в предните части. Тази прилика обаче е по-скоро формална и не трябва да забравяме, че са налице съществени разлики:

- при линейните регресионни модели имаме $k$ фактора и една целева променлива, докато при авторегресивните процеси става въпрос за стойностите на една и съща целева променлива, но в различни времеви контури $t_1, \dots, t_N$;
- линейните регресионни модели предполагаха нормално разпределение на грешките, докато тук $\varepsilon_{t_n}$ е израз на случайните колебания и не изисква определено разпределение (а само да отговаря на условията за бял шум);
- константният термин в израза е по-скоро показател за начална стойност, отколкото за определен минимум или стабилно състояние (при отсъствие на външни въздействия);
- методите и инструментите за анализ са различни (което за целите на този текст е от съществено значение);
- общият вид на зависимостта не съдържа гаранции, че авторегресивният процес е стационарен, и съответно не дава представа какви ще бъдат ефектите от използване на различните аналитични методи и инструменти.

Изброените разлики са важни и аргументират необходимостта от специално внимание към авторегресивните процеси. За да може да дадем отговор на въпроса доколко успешни ще бъдат опитите ни за тяхното изследване, трябва първо да можем да определим дали конкретният процес е стационарен, или не. Както коментирахме в началото на тази глава, доказателството за стационарност може да опрости значително анализа и да облекчи нашите усилия.

Проверката за стационарност изисква да запишем авторегресивния процес по друг начин, с използването на оператора за лаг. Това се получава, като оставим от дясната страна на уравнението само константата и белия шум:

$$y_{t_n} - \rho_1 y_{t_{n-1}} - \rho_2 y_{t_{n-2}} - \cdots - \rho_q y_{t_{n-q}} = \mathrm{c} + \varepsilon_{t_n} \,. \qquad ( \text{VII—32} )$$

И като си припомним, дефиницията на оператора за лаг, можем да получим, че:

$$y_{t_{n-i}} = L^i y_{t_n} \qquad ( \text{VII—33} )$$

$$\rho(L) y_{t_n} = \mathrm{c} + \varepsilon_{t_n}, \text{където } \rho(L) = \left(1 - \rho_1 L - \rho_2 L^2 - \cdots - \rho_q L^q\right) \qquad ( \text{VII—34} )$$

Ако представим последния израз като един полином и намерим неговите корени, се получава условието за стационарност на авторегресивния процес:

$$0 = \left(1 - \rho_1 q - \rho_2 q^2 - \cdots - \rho_q q^q\right), \qquad ( \text{VII—35} )$$

като всички корени на $q$ трябва да имат абсолютна стойност по-малка от единица.

Ако условията не са изпълнени, коефициентите на автокорелация няма да намаляват заедно с увеличаване на лага (например зависимостта между текущата стойност и все по-отдалечените във времето стойности няма да намалява), откъдето и случайните отклонения от предишните периоди ще имат също незатихващ ефект върху последващите стойности на $y_{t_n}$. За да илюстрираме принципа на проверката за стационарност, ще приложим дискутираното условие върху един от широко известните във финансите процеси – т.нар. случайна разходка (random walk).

Терминът „случайна разходка“ е използван за първи път още в началото на миналия век от Pearson [20]. Една от ярките илюстрации на процеса е свързана с поведението на човек, който избира да прави крачка наляво или надясно в зависимост от последователни подхвърляния на монета. Тогава изминатото от него разстояние ще се представя за всяко едно хвърляне като:

$$y_{t_n} = y_{t_{n-1}} + \varepsilon_{t_n} \,, \qquad ( \text{VII—36} )$$

т.е. на всяка стъпка ще имаме изминатото разстояние в момента на предходното хвърляне на монетата плюс текущия резултат (който е случаен процес и ще определи дали ще пристъпим наляво, или надясно). Преди да продължим с изложението, трябва да направим едно важно уточнение. Хвърлянето на монета е схема на Бернули и ние вече използвахме такава схема при дефинирането на случайни процеси, тип бял шум. За да не изпадаме в заблуждение, трябва да отбележим, че има разлика дали изследваме съвкупността (подредбата) на получените ези/тура, или тяхната сума (както е в случая на случайната разходка). Т.е. в случая няма противоречие, че схема на Бернули е в основата и на двата варианта, тъй като в различните случаи изследваме различни неща.

За да проверим дали случайната разходка описва стационарен процес, трябва да извършим описаните по-горе преобразувания и да получим полинома, чиито корени ни интересуват. Като прехвърлим $y_{t_{n-1}}$ в лявата част на уравнението, ще получим:

$$y_{t_n} - y_{t_{n-1}} = \varepsilon_{t_n} \qquad ( \text{VII—37} )$$

$$\rho(L) y_{t_n} = \varepsilon_{t_n}, \text{където } \rho(L) = 1 - L\,. \qquad ( \text{VII—38} )$$

Следователно полиномът, на който търсим решения, е от първа степен и се получава като:

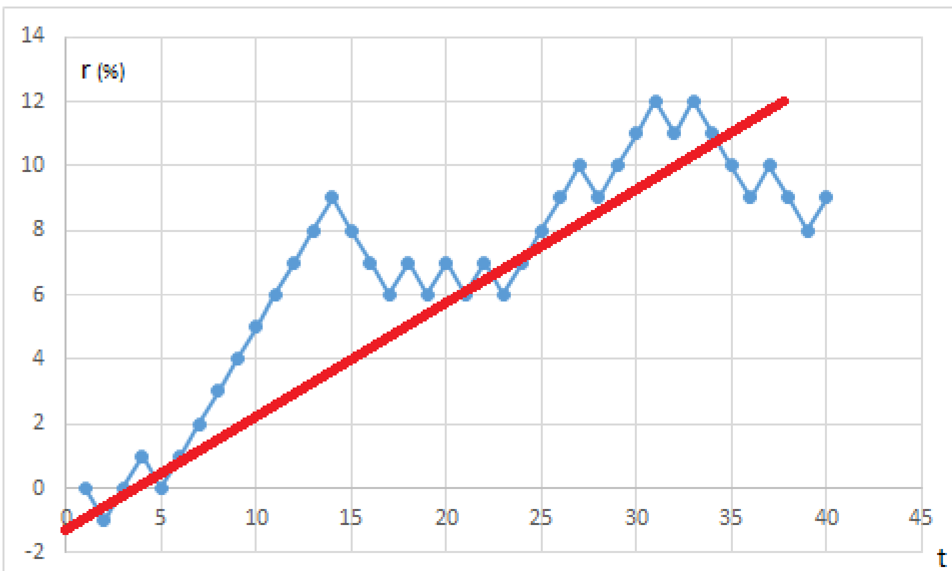
$$0 = 1 - q, \text{откъдето следва, че } q = 1\,. \qquad ( \text{VII—39} )$$

Условието за стационарност беше абсолютната стойност на всяко едно от решенията да бъде по-малка от единица, тук то не е изпълнено и следователно случайната разходка не е стационарен процес. По аналогичен начин бихме могли да направим проверка за авторегресивни процеси от по-висок ред, като, разбира се, в тези случаи ще трябва да се потрудим повече при намирането на корените за $q$ в ( VII-35 ).

*Таблица 23. Основни характеристики на авторегресивните процеси*

| Характеристика | Формула |
|---|---|
| Средна / очаквана стойност | $E(y_{t_n}) = \dfrac{c}{1 - \rho_1 - \rho_2 - \cdots - \rho_q}$ |
| Стандартно отклонение за AR(q) | $\sigma^2_{y_{t_n}} = \dfrac{\sigma^2_\varepsilon}{1 - {\rho_1}^2}$ |
| Автоковариация за AR(q) | $\gamma_1 = \dfrac{\rho_1 \sigma^2_\varepsilon}{1 - {\rho_1}^2}, \gamma_2 = \dfrac{{\rho_1}^2 \sigma^2_\varepsilon}{1 - {\rho_1}^2}, \gamma_3 = \dfrac{{\rho_1}^3 \sigma^2_\varepsilon}{1 - {\rho_1}^2}$ |
| Автокорелация за AR(q) | $\gamma_0 = \sigma^2_{y_{t_n}}, \quad \dfrac{\gamma_1}{\gamma_0} = \rho_1, \quad \dfrac{\gamma_2}{\gamma_0} = {\rho_1}^2, \quad \dfrac{\gamma_3}{\gamma_0} = {\rho_1}^3$ |

Авторегресивните процеси имат широко приложение при изследването на различни икономически и финансови проблеми. Всъщност една много съществена част от използваните теоретични модели стъпва в различна степен на допускането, че при наличието на стойности от миналото ние можем да „дестилираме“ основните характеристики на анализираното явление и следователно да прогнозираме неговите бъдещи форми и прояви. За една група от икономически показатели подобно допускане е оправдано – например можем да очакваме, че потреблението, нивата на безработица, миграцията по икономически причини и др. ще демонстрират инертност и могат да бъдат моделирани на базата на вече известните им стойности. Както отбелязахме при анализа на случайната разходка обаче, не всички процеси, при които имаме изменение на базата на „промяна“ спрямо предишното равнище, са подходящи за авторегресионен анализ. Но дори и в подобни случаи познаването на AR и MA процесите може да ни предпази от предварителни изводи като този, който „се набива на очи“ и е изобразен на Фигура 20.




*Фигура 20. Примерно развитие на цената на акция с времето*

На фигурата е показано развитието на цената на една акция с времето (t), като е нанесена дневната доходност, изчислена на базата на разликите в цената на затваряне. Ако използваме линеен регресионен модел (или ако просто гледаме достатъчно дълго графиката), можем да направим извод, че доходността на акцията бележи съществен ръст, а следователно може да очакваме и нейната цена да нараства.

На практика обаче, горната графика е генерирана с помощта на Excel и представлява част от симулация точно на случайна разходка. Ако ние не разполагаме с познания за същността на авторегресивните процеси и нямаме метод да проверим дали е налице стационарност, то тогава бихме могли да изпаднем лесно в състояние, при което да стигнем до погрешни изводи и в крайна сметка да загубим пари.

## 2.4. ARMA процеси

ARMA моделите се получават като комбинация на авторегресивен (AR) и процес с движещите се средни (MA). При това не е задължително размерността на двата процеса да е една и съща, поради което ARMA моделите имат два показателя за лаг – комбинирането

### Случайна разходка с Excel

Excel разполага с функция за генериране на равномерно разпределени случайни числа в интервала от 0 до 1 (RAND()). Можем да използваме тази функция в комбинация с функцията за условно изпълнение и да генерираме случайна разходка:

В клетката A0 записваме 0.
В клетката A1 записваме = A0 + IF (RAND()>0.5, 1, -1)
След това можем да копираме формулата от A1 в A2, A3 и т.н. – колкото стойности желаем.

на AR($q$) и MA($l$) се обозначава като ARMA($q$, $l$). От дефиницията за авторегресивен процес в част IV.2.3, ще имаме:

$$\text{AR}(q): y_{t_n} = \mathrm{c}_{AR} + \rho_1 y_{t_{n-1}} + \rho_2 y_{t_{n-2}} + \cdots + \rho_q y_{t_{n-q}} + \varepsilon_{t_n} \qquad (\text{VII—40})$$

а от дефиницията на процес с движещите се средни от част IV.2.2, ще имаме:

$$\text{MA}(l): y_{t_n} = c_{MA} + \theta_0 v_{t_n} + \theta_1 v_{t_{n-1}} + \cdots + \theta_l v_{t_{n-l}}, \qquad (\text{VII—41})$$

където $v$ е случаен процес с характеристики на бял шум.

Като комбинираме двата процеса за ARMA(q, $l$) ще получим израз от вида:

$$y_{t_n} = c + \rho_1 y_{t_{n-1}} + \rho_2 y_{t_{n-2}} + \cdots + \rho_q y_{t_{n-q}} + \theta_0 v_{t_n} + \theta_1 v_{t_{n-1}} + \cdots + \theta_l v_{t_{n-l}} + \varepsilon_{t_n} \qquad (\text{VII—42})$$

Горният израз означава, че полученият процес ще притежава комбинация от характеристиките на авторегресивната си компонента и на съставната си част с движещи се средни. Това допринася значително за популярността и приложимостта на ARMA процесите, тъй като позволява да се моделират и изследват редица икономически и финансови проблеми, при които имаме изразен тренд (включително и зависимост от предните стойности на изследваните променливи и инертност), но едновременно с това са налице предпоставки за резки изменения под въздействието на промяна в икономическата обстановка.

За да можем да се възползваме в пълна степен от гъвкавостта, която ни предоставят ARMA моделите, трябва да можем да си отговорим на три основни въпроса:

- Как да разпознаем кога е уместно за използваме ARMA процес?;
- С какво ARMA процесите са по-различни от простите авторегресионни и с движещи се средни процеси?
- Какви са основните характеристики на ARMA процесите?

За да отговорим точно на въпроса за разликата между ARMA и обикновените AR и MA процеси, е необходимо да използваме още един показател, който не беше коментиран до момента – частичната автокорелация. В Таблица 21 дефинирахме автокорелацията като стандартизиран измерител на зависимостта между стойностите на една променлива през два времеви периода (когато я използваме в контекста на времевите редове). Частичната автокорелация измерва корелацията между стойността $y_{t_n}$ и $y_{t_{n-q}}$, без да се отчита влиянието на „междинните“ стойности, които са $y_{t_{n-1}}, y_{t_{n-2}}, \dots y_{t_{n-q+1}}$. За първоначалното ни запознаване с частичната автокорелация е необходимо да познаваме две от нейните свойства:

- За лаг 1 стойността на частичната автокорелация съвпада с тази на показателя за автокорелация, което е следствие на това, че между $y_{t_n}$ и $y_{t_{n-1}}$ няма „междинни“ стойности.
- За авторегресивен процес от ред $q$ стойността на частичната корелация е нула за лаг, по-голям от $q$.

С помощта на автокорелацията и частичната автокорелация можем да направим съпоставка между разглежданите вече авторегресивни процеси, процеси с движеща се средна и ARMA процесите. Таблица 24 демонстрира основните различия.

*Таблица 24. Сравнение между AR(k), MA(l) и ARMA(k, l)*

| **AR(q)** | **MA(l)** | **ARMA(q, l)** |
|---|---|---|
| Експоненциално намаляваща автокорелация. | Броят на ненулевите стойности на автокорелацията е равен на $l$. | Експоненциално намаляваща автокорелация |
| Броят на ненулевите стойности на частичната автокорелация е $q$. | Експоненциално намаляваща частична автокорелация. | Експоненциално намаляваща частична автокорелация. |

Таблица 24 ни дава основните характеристики с чиято помощ можем да направим разлика между последните три вида разгледани процеси. Но тя не съдържа в себе си насоки по какъв начин да изберем тип процес, който да моделира успешно данните, с които работим. Ако сме сигурни, че изследвания процес съдържа времева зависимост между отделните наблюдения, то тогава бихме могли да използваме подхода предложен от Box и Jenkins [23], с чиято помощ да изградим нашият модел.

### 2.4.1. Подход на Box-Jenkins

Стъпките които предлагат Box и Jenkins при анализа на времеви редове, за които предполагаме наличие на авторегресивен, с плъзгащи се средни или ARMA процес са три:

- Установяване на основните характеристики данните и избор на вид процес, който да описва разполагаемите данни;
- Оценка на параметрите на идентифицираната на предния етап зависимост;
- Проверка на качествата на изградения модел.

Редът на посочените стъпки е важен, тъй като резултатите от всеки един етап се използват като основа за следващия. Например, установяването на основните характеристики на данните и избор на вид процес е свързано с търсенето на отговор на два важни въпроса:

  - Кой от разглежданите до момента видове процеси (ако предложим, че се ограничаваме в избора само до тях) „подхожда“ най-добре на информацията с която разполагаме?

За да отговорим на този въпрос, можем отново да използваме информацията систематизирана в Таблица 24, като за целта предварително изчислим коефициентите на автокорелация, и частична автокорелация. В зависимост от това, кой от посочените в таблицата случаи наблюдаваме можем да се насочим към избор на съответния тип процес.

  - Каква да бъде размерността/дълбочината на модела, като това решение обикновено не може да се базира само и единствено на получените стойности за автокорелацията и частичната автокорелация?

Изборът на модел с определена размерност е по същество компромис между сложността на изграждания модел и степента в която той може да опише добре наличните данни. Окончателното решение зависи както от характеристиките на анализирания проблем, така и от другите ограничения (например време, изисквания за сложност и точност) на модела. Един от важните проблеми, който трябва да бъде разрешен на този етап е до каква степен предположението за стационарност е приложимо към анализираните финансови и икономически проблеми. Както посочват Commandeur и Koopman [24] на практика това допускане може да не е изпълнено, или да е много трудно да се докаже с достатъчно висока гаранционна вероятност. Отклоненията от изискването за стационарност може да доведе до сериозни отклонения и неточности в получените резултати, които да обезсмислят анализа. Добра отправна точка за методите и стъпките за анализ на нестационарни времеви редове е книгата на Clements и Hendry [25].

Втората стъпка е свързана с получаването на оценка за параметрите на модела, след като вече е установен неговия вид и характеристики. За целта може да използваме метода на най-малките квадрати, който разгледахме подробно в Част III.4 или някоя от неговите модификации в Част III.5. Оценката на параметрите е важна и има непосредствено отношение към качеството на крайните резултати, но нейната ефективност е ограничена от взетото вече решение за вида и размерността на модела. Поради тази причина основното внимание към практическата употреба на подхода на Box-Jenkins е насочена преди всичко към първата стъпка.

Последната стъпка е свързана с оценката на качествата на получения модел и до каква степен той представя адекватно изследваните финансови и икономически явления. Обикновено на този етап се използват два, допълващи се метода:

- Сравняване с модели с по-голяма размерност;
- Анализ на остатъчните стойности (грешките).

Сравняването с модели, чиято размерност изкуствено е увеличена позволява да преценим дали усложняването на модела води до пренебрежимо малки (или никакви) подобрения в качеството на получените резултати. Ако това е така, то избраните в първата стъпка размерност и сложност са достатъчни. При сравняването можем успешно да използваме различните методи за статистическа проверка на хипотези, които бяха разгледани в предните части.

Анализът на остатъчните стойности е проверка за наличието на линейна зависимост между тях, която би означавала, че използвания модел няма необходимата дълбочина и не включва всички фактори, които указват влияние върху $y_{t_n}$. Един от възможните подходи при анализа на остатъчните стойности е да проверим хипотезата, че те отговарят на изискванията за бял шум, като за целта може да използваме разглеждания в Част VII.2.1 Q тест на Ljung-Box.

## 3. Приложение на анализа на времеви редове във финансите

Анализът на времеви редове намира широко приложение във финансите. Неговите превъплъщения са толкова разнообразни и интересни, че пълното им изброяване заслужава текст, който да е посветен само на това. Целта на тази част е по-скромна, а именно да обърне внимание на малко на брой, но съществени, характеристики от тези приложения.

Първата важна характеристика е, че не винаги е възможно ясно да бъде направено точно разграничение между структурните модели, чийто представител е обикновената линейна регресия разгледана в Част III и анализа на времеви редове под формата на авторегресивни процеси и процеси с движещи се средни, които разгледахме в Част VII. При изследването на реални финансови проблеми се налага описването на сложни обекти, които се движат от няколко променливи и едновременно с това всяка от тези променливи зависи от предходните си стойности. Въпреки че изборът на тип модел зависи от особеностите на изследвания проблем можем да използваме едно просто помощно правило:

*Структурните модели са за предпочитане когато разполагаме с ясна представа за взаимоотношенията между факторите и целевата ни променлива, както и когато са налице достатъчно данни за стойностите на всеки от факторите.*

За да демонстрираме ефективността на това правило, нека предположим, че сме избрали за използваме структурен модел за да обясним измененията във времето на променливата $y$, по начина показан в ( VII—43 ):

$$y_{t_n} = \alpha + \beta_1 x_{1t_n} + \beta_2 x_{2t_n} + \varepsilon_{t_n} \,. \qquad \text{( VII—43 )}$$

Тогава, дори да сме уверени във връзката на целевата променлива и факторите и да разполагаме с достатъчно данни за да получим стабилни оценки за $\alpha$ , $\beta_1$ и $\beta_2$ това все още не е достатъчно за да имаме един добър модел за $y_{t_n}$. Това е така, защото за да получим стойностите за целевата променлива за даде времеви период то ние трябва да знаем стойностите на факторите през този период. Или ако става въпрос за прогнозиране (каквато е много често целта на иконометричните модели) то ние просто прехвърляме тежестта от прогнозиране на достатъчно точни стойности за $y_{t_n}$, към получаване на добри прогнозни стойности за $x_{1t_n}$ и $x_{2t_n}$. Когато информация за стойностите на факторите не е налична или не можем да разчитаме на това, че тя е достатъчно точна, това е индикатор че трябва да разгледаме възможността да подходим към решаване на проблема с помощта на анализа на времеви редове.

Втората важна характеристика е свързана с представянето на целевата променлива, а именно дали моделът да бъде изграден на база на нейните стойности, на интервали в които променливата може да попадне или дори само на база посоката на нейното изменение.

Например, ако става въпрос за моделиране измененията на един пазарен индекс дали да изберем за целева променлива стойността на индекса, възможността той да попадне в определени интервали (например повишение/понижение с 5%, 10%, 15% и т.н.) или просто да проследим дали той ще се повиши или понижи, без да се интересуваме от мащаба на промяната. Кое представяне ще изберем зависи преди всичко от целите на изграждания модел, но то предопределя и за какъв кръг от проблеми може да се използват нашите резултати (например, модел базиран на само на посоката на изменение няма да може да се използва ако искаме да получим точна оценка за целевата променлива).

Третата важна характеристика е свързана с проверката на ефективността на изграждания модел. При това проверката може да се осъществи въз основа на данни, които са използвани за оценка на параметрите на модела, така и въз основа на данни, които не са използвани за получаване на оценки за параметрите на модела. За да оценим максимално точно до каква степен нашият модел е адекватен на изследвания проблем трябва да използваме и двата „типа“ данни, а не да се ограничаваме само с грешката получена в процеса на оценяване на параметрите му. Ако разполагаме с достатъчен обем данни един от предпочитаните подходи е да не използваме всички от тях за оценка на параметрите на модела, а част да бъдат оставени в „резерв“ и впоследствие да бъдат използвани за да преценим до каква степен получените резултати са адекватни и използваеми на практика. Не е без значение, кои точно от разполагаемите данни да бъдат използване за проверка на модела и какъв да бъде техният относителен дял спрямо цялата информация с която разполагаме. Отговорът на този въпрос зависи до голяма степен от конкретния решаван проблем и от количеството информация с която разполагаме.

Четвъртата важна характеристика е свързана с оценка на грешките на модела, но гледани от малко по-различен ъгъл. Тя е свързана с това, че когато анализирахме грешките, ние винаги до сега работехме със сумата от квадратите между получените от модела и реално наблюдаваните стойности на целевата променлива. Този подход не позволява грешките с противоположен знак да се „компенсират“ взаимно и дава по-точна представа за ефективността на получените резултати. Но той носи със себе си и две особености, които не са без значение за решаването на финансови проблеми:

- При редица финансови задачи отношението към надценяването или подценяването (т.е. грешка с различен „знак“) на целевата променлива не е симетрично.
- Тъй като оценяваме грешката на база сравнение на множество съпоставки между оценка/реална стойност то това означава, че не можем да съдим за качествата на модела на база само едно наблюдение.

В заключение ще споменем, че един от често пренебрегваните но важни етапи при прилагането на анализа на времеви редове за решаване на финансови проблеми е проверката дали са изпълнени допусканията за приложимост на отделните процеси (например дали изследваните процеси са стационарни). Крайните резултати винаги са по „интригуващи“ и изглеждат „по-важни“ от тази проверка, но без да сме сигурни че условията при които работи правилно анализа на времеви редове са налице, не можем да бъдем сигурни и до каква степен получените от нас резултати са валидни въобще.

## БИБЛИОГРАФИЯ


[1] O. Bjerkholt, Ragnar Frisch and the Foundation of Economic Society and Econometrica, Oslo: Statistics Norway, Research department, 1995.

[2] P. Wilmott, Quantitative Finance, том 1, NY: Wiley, 2006.

[3] C. Brooks, Introductory Econometrics For Finance, Cambridge: Cambridge University Press, 2008.

[4] I. J. Myung, „Tutorial on maximum likelihood estimation“, *Journal of Mathematical Psychology,* том 47, 2003, pp. 90‒100.

[5] C. R. Hill, G. G. Judge и W. Griffiths, Undergraduate Econometrics Second (2nd) Edition, 2000.

[6] E. F. Fama и K. R. French, „Common risk factors in the returns on stocks and bonds“, *Journal of Financial Economics,* том 33, № 3, 1993.

[7] D. H. Johnson, „The Insignificance of Statistical“, *Journal of Wildlife Management,* том 3, № 63, 1999, pp. 763‒772.

[8] И. Съйкова, А. Стойкова - Къналиева и С. Съйкова, Статистическо изследване на зависимости, София: УНСС, 2002.

[9] А. Велева, *Съвкупностният подход в статистиката,* Пловдив: Макрос, 2013.

[10] C. D. Schunn и D. Wallach, Evaluating Goodness-of-Fit in Comparison of Models to Data, University of Pittsburg, University of Applied Sciences Kaiserslautern, 2005.

[11] C. R. Hill, W. E. Griffiths и C. L. Guay, Principles of Econometrics, 4th Edition, John Wiley & Sons, 2010.

[12] H. White, „A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity, *Econometrika,* том 4, № 48, 1980.

[13] T. S. Breusch, „Testing for Autocorrelation in Dynamic Linear Models,“ *Australian Economic Papers,* том 17, 1979, pp. 334–355.

[14] L. G. Godfrey, „Testing Against General Autoregressive and Moving Average Error Models when the Regressors Include Lagged Dependent Variables", *Econometrica*, том 47, 1978, pp. 1293–1302.

[15] M. A. Stephens, „EDF Statistics for Goodness of Fit and Some Comparisons", *Journal of the American Statistical Association,* том 69, 1974, pp. 730−737.

[16] C. M. Jarque и A. K. Bera, „Efficient tests for normality, homoscedasticity and serial independence of regression residuals", *Economics Letters,* том 6, № 3, 1980, pp. 255–259.

[17] J. B. Ramsey, „Tests for Specification Errors in Classical Linear Least Squares Regression Analysis", *Journal of the Royal Statistical Society,* том 31, № 2, 1969, pp. 350–371.

[18] W. Studenmund, Using econometrics: A practical guide 4th ed., Addison-Wesley, 2000, p. 167.

[19] G. C. Chow, „Tests of Equality Between Sets of Coefficients in Two Linear Regressions", *Econometrica,* том 28, № 3, 1960, pp. 591–605.

[20] G. M. Ljung и G. E. P. Box, „On a Measure of a Lack of Fit in Time Series Models", *Biometrika,* том 65, № 2, 1978, pp. 297–303.

[21] R. G. Brown, Exponential Smoothing for Predicting Demand, Cambridge, Mass.: Arthur D. Little Inc, 1956.

[22] C. C. Holt, „Forecasting Trends and Seasonal by Exponentially Weighted Averages", том 56, 1957.

[23] K. Pearson, „The Problem of the Random Walk", *Nature,* том 65, 1905.

[24] G. Box и G. Jenkins, Time series analysis: Forecasting and control, San Francisco: Holden-Day, 1970.

[25] J. Commandeur и S. Koopman, Introduction to State Space Time Series Analysis, Oxford University Press, 2007.

[26] M. P. Clements и D. F. Hendry, Forecasting Non-Stationary Economic Time Series, The MIT Press, January 2001.

[27] P. Artzner, F. Delbaen, J.-M. Eber и D. Heath, „Coherent Measures of Risk", том 9, 1999.

[28] B. P. Kaare и M. S. Pedersen, „The Matrix Cookbook“, *Intelligent Signal Processing,* № Technical University of Denmark, 2012.