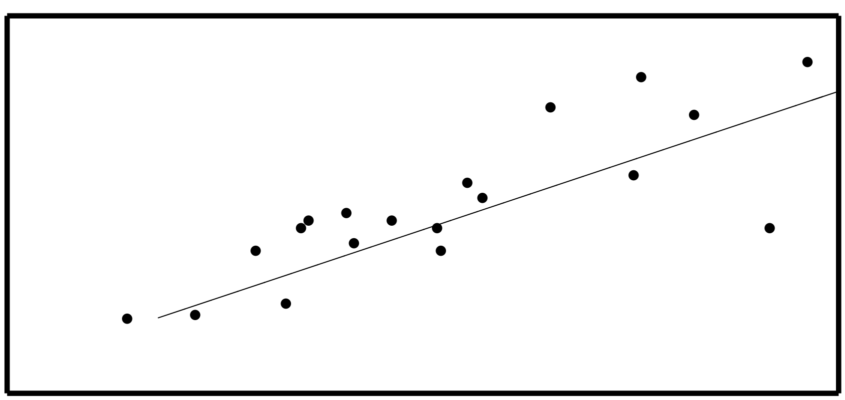
Alfredas Račkauskas

EKONOMETRIJOS ĮVADAS



2003 Vilnius

Įžanga

Teorija be praktikos – sausa
Praktika be teorijos – akla
(Emanuelis Kantas)

Ekonometrija – tarpkryptinė disciplina – ekonomikos, matematikos ir statistikos bendras vaisius. Jos, kaip savarankiško mokslo, pradžia skaičiuojama nuo 1930 metų, kai buvo įkurta ekonometrų draugija ir pradėtas leisti žurnalas "Econometrics". Dabar, mokslinių žurnalų skirtų ekonometrijai priskaičiuojama per dešimt. Tarp jų, "Journal of Econometrics", "Journal of Applied Econometrics", "Econometric theory", "Econometric reviews", "Economie et statistique" ir kt. Moksliniai ekonometrijos straipsniai randami daugelyje ekonominių ir matematinių žurnalų. Kasmet vyksta po keletą stambių konferencijų, skirtų ekonometrijos problemoms. Visa tai rodo augančią šio mokslo svarbą. O jo pripažinimą parodo net kelių ekonomikos Nobelio premijų paskyrimas ekonometrams.

Ekonometrijos kursai dėstomi daugelyje universitetų. Tam tikslui išleista nemažai vadovėlių, kurie skiriasi tiek pagal reikalingą matematinį pasiruošimą, tiek pagal teorijos ir praktikos santykį. Štai Gryno [3] vadovėlis pargįstai laikomas ekonometrijos enciklopedija. Goldbergeris [2] siūlo labiau matematizuotą požiūrį į ekonometriją. Džonstono ir Dinardo vadovėlis [4] – puikus teorijos ir praktikos derinimo pavyzdys. Neblogas standartinių vadovėlių papildymas yra Kenedžio ekonometrijos vadovas [6], o Hamiltono laiko eilučių knyga [5] – tikras šios srities perlas.

Lietuvoje sistemiškas ekonometrų ruošimas pradėtas tik prieš trjetą metų, patvirtinus tam skirtą statistikos bakalauro programą. Studentams ši programa suteikia galimybę gauti gerą matematinį ir ekonominį išsilavinimą bei įsisavinti šiuolaikines informacines technologijas. Ekonometrijos kursai prasideda nuo įvado, apžvelgiančio ekonometrijos tikslus ir priemones jiems pasiekti. Jam skirta ši mokymo priemonė, atitinkanti ekonometrijos įvado vieno semestro kursą. Ji paruošta sekant Hilo, Grifito, Judžo vadovėliu [1] ir Melenbergo mokymo priemone [7] bei atsižvelgiant į studentų matematinį pasiruošimą.

Įvadiniame skyriuje trumpai apžvelgiama ekonometrijos metodologija, aptariant stochastinius modelius bei jų įvertinimą ir diagnostiką. Antrasis ir trečiasis skyriai skirti regresinei analizei. Įvadas į tokią plačią ir labai svarbią temą pradedamas sąryšio tarp namų ūkių pajamų ir išlaidų maistui paieška. Šiuo paprastu pavyzdžiu išaiškinami pagrindiniai regresinės analizės principai. Įvadui į daugelio kintamųjų vienos lygties regresiją pasitelkiamas firmos pajamų sąryšis su produkcijos kaina ir išlaidomis reklamai. 4-6 skyriuose supažindinama su įvairiais regresinės analizės aspektais: heteroskedastiškumu, autokoreliacija, atsitiktinių regresorių atveju ir kitais. Simultaninių lygčių modeliai aptariami 9 skyriuje, pasitelkiant patį paprasčiausią uždaros rinkos be užsienio prekybos modelį. Paskutiniame, 10 skyriuje supažindinama su pagrindiniais laiko eilučių modeliais ir baziniais finansų ekonometrijos principais. Prieduose pateikiamos reikalingos tikimybių teorijos ir algebros žinios.

Turinys

1 skyrius. Ekonometrijos principai

1.1 Kas yra ekonometrija

Tai klausimas į kurį trumpai ir vienareikšmiškai vargu ar kas atsakys. Žodis „ekonometrija“ yra dviejų graikiškų žodžių $o\iota\kappa o\nu o\mu\iota\alpha$ (ekonomika) ir $\mu\epsilon\tau\rho o\nu$ (metrika) junginys. Gi ekonometrijos mokslas turi labai platų išaiškinimo spektrą: nuo buitinio ("... tai ką daro ekonometrai ") iki enciklopedinio ("... pagal empirinius duomenis vertina ir analizuoja ekonominių objektų ir procesų sąryšius"). Ekonometras vienu metu privalo būti ekonomistas (norint empiriškai analizuoti ekonomiką, reikia gerai suprasti ekonomikos teoriją), statistikas (testuojant ekonominių procesų sąryšius reikia taikyti egzistuojančius statistinius testus bei kurti naujus) ir matematikas (ekonominę problemą ir jos sprendimą reikia išreikšti matematiniais terminais). Dar pridėjus būtinumą išmanyti saskaitybą, taikomąją ir ekonominę statistikas bei sugebėjimą pasinaudoti naujausiomis informacinėmis technologijomis, gaunamas gana įvairiaspalvis ekonometro paveikslas.

Ekonometrijos metodologijoje skiriamos šios dalys:

- modelio specifikavimas;
- įvertinimas ir interpretavimas;
- diagnostika;
- modelių palyginimas.

Ekonomikos teorija kartu su matematine ekonomika nustato įvairių ekonominių objektų bei procesų tarpusavio sąveikas ir perteikia jas matematiniais simboliais[1] Tai sudaro teorines prielaidas ekonometrinio modelio kūrimui (modelio specifikavimui). Paprastai nagrinėjami trijų tipų modeliai: regresiniai, simultaninių lygčių ir laiko eilučių.

Regresiniais modeliais aprašoma ekonominių dydžių priklausomybė nuo vieno ar kelių ekonominių faktorių. Pavyzdžiui, trumpalaikių paskolų palūkanų normos priklausomybė nuo infliacijos kitimo ir bendrojo nacionalinio produkto augimo tempų, namų ūkių išlaidų maistui priklausomybė nuo pajamų ir t.t.

[1] Ypatingo skirtumo tarp ekonomikos teorijos ir matematinės ekonomikos nėra. Abu šie mokslai aprašo ekonominius sąryšius tik pirmasis tai padaro žodine (verbaline) forma, o antrasis – matematinių simbolių pagalba.

Simultaninių lygčių modeliais aprašomi dydžiai, susieti funkcine priklausomybe tiek su kitais ekonominiais dydžiais, tiek ir tarpusavyje. Pavyzdžiui, Lietuvos tekstilės pramonės ekonometriniame modelyje būtų lygtys, aprašančios tekstilės paklausą, eksporto apimtis, darbo vietų skaičių, investicijas į šią ūkio šaką, tekstilės kainą. Visi tie dydžiai yra priklausomi tarpusavyje ir, be to, priklauso nuo kitų rodiklių, kaip antai: vidutinės palūkanų normos, vartojimo prekių indekso, nacionalinio produkto augimo tempų ir pan.

Laiko eilučių modeliai skirti ekonominių objektų tyrimams pagal jų elgesį praeityje. Jais patogu tirti, pavyzdžiui, akcijų kainas, paskolų palūkanų normą ir pan. Laiko eilučių modeliai labai naudingi, kai apie tiriamą arba prognuozuojamą ekonominį objektą beveik nieko, išskyrus jo elgesį praeityje, nežinoma.

Kai modelis jau sukurtas, tada, remiantis empiriniais duomenimis, kuriuos dažniausiai reikiamai apdorotus pateikia ekonominė statistika, įvertinami modelio parametrai ir patikrinamas modelio atitikimas ekonomikos teorijai.

Ekonometrinio modelio vertinimui pritaikomi matematinės statistikos metodai (mažiausių kvadratų, didžiausio tikėtinumo, kvantiliniai ir pan.) bei vystomi specialūs metodai. Prie tokių priskirtini dvigubas mažiausių kvadratų metodas, trigubas mažiausių kvadratų metodas ir pan.

Įvertinto modelio diagnostikai pasitelkiami matematinės statistikos metodai (Stjudento testas, F-testas ir pan.) bei vystomi savi (Dickey-Fulerio, Durbino-Watsono testai ir pan.)

Ekonometrijos rezultatai taikomi:

norint patikrinti ekonomikos teorijos teiginius arba pirmines sąlygas (pavyzdžiui, parametrų ženklus, *a priori* galimas reikšmes ir t.t.);

analizuojant ekonominių reiškinių struktūrą (pavyzdžiui, BVP struktūrą);

prognozuojant ekonominius rodiklius (marketingas) (globalūs rodikliai BVP, bedarbystės-užimtumo lygis, prekių indeksai ir pan.);

formuojant ekonominės politikos principus.

Pažymėtina, kad ekonometriniai modeliai išvystyti beveik visoms ekonomikos sritims, įskaitant:

- vartojimo prekių rinką,
- energetiką,
- namų ūkio ekonomiką,

- industrijos ekonomiką,
- tarptautinę prekybą,
- finansus,
- darbo ekonomiką,
- transportą,
- regioninę ekonomiką.

Pastaraisiais metais ypač sparčiai vystoma finansų ekonometrija ir namų ūkio ekonometrija.

Beje, ekonometriniai modeliai bei ekonometrijos metodai taikomi ir kituose moksluose, pavyzdžiui demografijoje (taikomi regresiniai modeliai prognozuojant gimstamumą, mirtingumą), politiniuose moksluose, sociologijoje (analizuojant apklausų rezultatus).

1.2 STOCHASTINIAI MODELIAI PRIEŠ DETERMINISTINIUS

Modeliavimas yra neatsiejama bet kurio mokslo dalis - tiek socialinio, tiek gamtos. Realaus pasaulio sistemos paprastai yra labai sudėtingos. Norint šias sistemas suprasti, prognozuoti jų elgesį ar kontroliuoti, būtina jas supaprastinti, t.y. sukurti modelį. Egzistuoja daug modelio formų: pavyzdžiui, *verbaliniai/logistiniai* (sistemų veiklos aiškinimas paradigmomis, pavyzdžiui, "nematomos rankos" paradigma), *fizikiniai* (sumažinto mastelio ir supapra-tintos veiklos modeliai), *geometriniai* (lentelės, diagramos), *algebriniai* (algebrinės lygtys) ir pan.

> *Modelis – originalo atvaizdas, tapatus pasirinktu struktūros lygmeniu arba pasirinktomis funkcijomis.* (TŽŽ) [2];
>
> *– pavyzdys, pagal kurį kas nors gaminama, kuriama, tiriama.* (Ten pat.)

[2]Tarptautinių žodžių žodynas.

Grafikas 1.1: Geras ir labai geras modeliai (aut. *Matisse*)

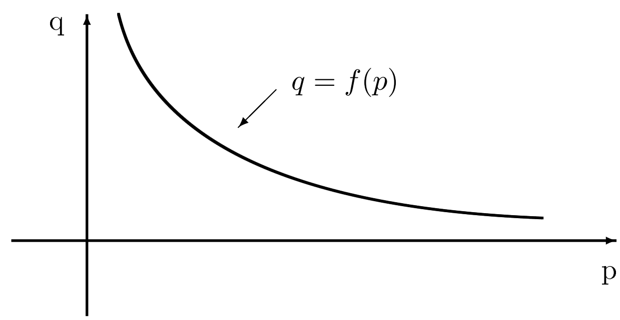
Sukurti *matematinį modelį* reiškia nagrinėjamam objektui suteikti matematinę išraišką. Čia gali pasireikšti du kraštutinumai: realistinis ir idealistinis. Realistinis modelis paprastai gana tiksliai aprašo tiriamą objektą, bet būna toks sudėtingas, kad neįmanoma jo nei ištirti, nei įvertinti. Kita vertus, idealistinis modelis, su kuriuo lengva dirbti, gali būti perdaug nutolęs nuo realaus tiriamo ekonominio fenomeno. Todėl geras modelis yra tam tikras kompromisas tarp realaus ir idealaus. Rasti tinkamą kompromisą – menas, kurio rezultatus didžia dalimi nulemia ekonometristo pasiruošimas ir talentas.

> ***Visi modeliai yra klaidingi. Tačiau kai kurie – naudingi.*** *(G. Box)*
>
> ***Modeliai yra tam kad juos naudotum, bet ne tam, kad jais tikėtum.*** *(H. Theil)*

Daugelis tradicinių ekonominių teorijų, ypač jei ryšiai tarp ekonominių objektų ir procesų išreiškiami diagramų ar algebrinių lygčių pagalba, postuluoja tikslią nagrinėjamų procesų funkcinę priklausomybę. Pavyzdžiui, paklausos teorija nustato, kad prekės paklausa priklauso nuo jos kainos:

$$q = f(p); \tag{1.1}$$

čia q – kurios nors prekės paklausos kiekis; p – tos prekės kaina.
Tačiau akivaizdu, kad faktorių, įtakojančių prekės paklausą, yra kur kas daugiau: tai komplektuojančių dalių ar komponenčių kainos, vartotojo pajamos bei jų teikiamos pirmenybės (prioritetai) ir t.t.. Pavyzdžiui, vaisvandenių



Grafikas 1.2: Hipotetinė paklausos priklausomybė nuo kainos

paklausa priklauso nuo oro sąlygų, vietos ir pan. Gi norėdami aprėpti absoliučiai visus faktorius įtakojančius prekės paklausą, turėtume nagrinėti modelį

$$q = f(p, p_0, Y, T; x_1, x_2, \dots), \tag{1.2}$$

kai

p_0 – kitų prekių kainos;

Y – vartotojo pajamos;

T – dydis, įvertinantis vartotojo prioritetus;

$x_1, x_2, \dots$ – kiti faktoriai, darantys įtaką paklausai.

Bet tokį modelį sunku tirti jau vien dėl neapibrėžto kintamųjų skaičiaus. Tai koks gi galimas kompromisas tarp (1.1) ir (1.2) modelių? Labai paprastas. Galima išskirti tik pačius svarbiausius, arba didžiausią įtaką darančius kintamuosius, o likusius aprašyti vienu faktoriumi, sakykime, ε. Dėl natūralaus neapibrėžtumo, ε patogu interpretuoti kaip atsitiktinį faktorių. Taigi, vietoje (1.1) ir (1.2) galima nagrinėti, pavyzdžiui, modelį:

$$q = \psi(p, p_0, Y, T; \varepsilon) \quad \text{arba} \quad q = \psi(p, p_0, Y, T) + \varepsilon; \tag{1.3}$$

Jame atsiranda atsitiktinis (stochastinis) faktorius ε, kuris kartu reiškia, kad paklausa q taip pat traktuojamas kaip atsitiktinis dydis, o (1.3) lygtimi aprašome jo skirstinį (lygybė tarp atsitiktinių dydžių suprantama kaip jų skirstinių lygybė).

Modeliai aprašomi lygtimis, į kurias įeina atsitiktiniai faktoriai, vadinami stochastiniais.

> *Stochastiniai modeliai, aprašantys ekonominių reiškinių bei procesų ryšius, vadinami ekonometriniais.*

Taigi modelis, aprašomas (1.1) formule, – deterministinis, o (1.3) formule, – ekonometrinis. Klasikiniai ekonominiai modeliai, kaip taisyklė, yra deterministiniai, o ekonometriniai modeliai – stochastiniai. Situaciją, kai deterministiniai modeliai keičiami stochastiniais, nėra nauja. Fizikoje klasikinės Niutono mechanikos modeliai yra deterministiniai, o kvantinės mechanikos – stochastiniai. Pastebėjimas, kad neįmanoma tiksliai nustatyti elementariosios dalelės buvimo vietos, bet galima nustatyti jos buvimo vietos tikimybinį skirstinį, fizikoje sukėlė revoliuciją. Ekonometriniai modeliai aprašo ne konkrečias ekonominių procesų ir įvairių ekonominių rodiklių reikšmes, bet tų reikšmių tikimybinius skirstinius. Taip (1.3) modeliu aprašome ne tikslias paklausos q reikšmes, bet tų reikšmių tikimybinį skirstinį.

Be minėtojo, yra ir kitų argumentų stochastinio-ekonometrinio modelio naudai. Vieni jų siejami su matavimų paklaidomis, kiti – su ekonominės aplinkos objektų stochastiniu elgesiu. Tiksli priklausomybė $Y = f(X)$ tarp ekonominių faktorių X ir Y dažnai negalima todėl, kad vietoje dydžio X (pvz., gaunamų pajamų) stebimas $Z = X+u$, kai u – atsitiktinė paklaida (kai kas nuslėpė tikrąsias pajamas, ne visi respondentai buvo atviri ir pan.) Be to, dydis X pats gali būti susijęs su stochastiniu ekonominių objektų elgesiu. Pavyzdžiui, realusis banko kapitalas aprašomas tokiu tiesiniu stochastiniu modeliu:

$$BK^* = a + \xi BK + u;$$

čia BK^* – realusis banko kapitalas, BK – balansinis kapitalas, a – dydis, atspindintis nebalansinės veiklos kapitalą, ξ – banko akcijų rinkos kurso ir jų nominaliosios vertės santykis (šis dydis yra atsitiktinis), u – atsitiktinė paklaida. Šiuo modeliu aprašomos ne tikslios banko realaus kapitalo reikšmės, bet tų reikšmių tikimybinis skirstinys, kuris gaunamas nustačius bendrą paklaidos u ir parametro ξ skirstinį.

1.3 Modeliavimo tikslai ir priemonės

Kaip taisyklė ekonometrinis modelis kuriamas turint konkretų tikslą. Net pats kūrybinis procesas didžia dalimi priklauso nuo to, ar siekiama paaiškinti duomenų generavimo mechanizmą (modeliuojami duomenys), ar bandoma padėti politikams atsakant į klausimą "kas jeigu" (pavyzdžiui, kas nutiks padidinus akcizą; kas bus, jei vienaip ar kitaip pertvarkysime mokesčių politiką ir t.t.), ar norima prognozuoti ekonominių rodiklių elgesį (pavyzdžiui, BVP augimo tempus, palūkanų normą ir t.t.), ar patikrinti ekonominės teorijos kurį nors teiginį konkrečiai ekonominei aplinkai (pavyzdžiui, Engelo ar Filipso kreivių geometriją Lietuvos ekonominei aplinkai).

Bet kuriuo atveju, pradiniai žingsniai kuriant ekonometrinį modelį yra tokie:

- nustatomi kintamieji;
- parenkama matematinė modelio forma (modelis specifikuojamas);
- išaiškinamos nežinomų parametrų *a priori* galimos reikšmės.

Vienas iš svarbiausių žingsnių kuriant ekonometrinį modelį yra kintamųjų parinkimas[3]. Jų yra įvairių, kartais tie patys vadinami skirtingais vardais. Paprastai ekonometriniai kintamieji skirstomi į egzogeninius ir endogeninius, paaiškinančiuosius ir paaiškinamuosius, regresorius ir regresantus.

Endogeniniai kintamieji yra tie, kurių reikšmės nustatomos modelio pagalba.

> *Endogeninis – vidinės kilmės, sukeliamas vidinių priežasčių.* (TŽŽ)

Apskritai, modelis kuriamas norint paaiškinti endogeninių kintamųjų elgesį. Todėl tie kintamieji dar vadinamai *paaiškinamaisiais.* (1.1), (1.2) ir (1.3) modeliuose endogeninis kintamasis yra tik vienas – paklausa. Regresiniuose modeliuose paaiškinamieji kintamieji dar vadinami regresantais.

Egzogeniniais vadinami tie modelio kintamieji, kurių reikšmės nustatomos už modelio ribų.

> *Egzogeninis – sukeliamas išorinių priežasčių, veiksnių, kilęs iš išorės.* (TŽŽ)

(1.3) modelyje kintamieji p, p_0, T, G – egzogeniniai. Egzogeniniais kintamaisiai gali būti ir pavėlinti laike endogeniniai kintamieji. Tada jie vadinami *paaiškinančiaisiais.* Regresiniuose modeliuose egzogeniniai kintamieji dar vadinami *regresoriais.*

Tiek deterministiniai, tiek stochastiniai modeliai gali būti *statiniai* arba *dinaminiai.* (1.1) lygtimi aprašomas modelis yra statinis, nes jame laikas

[3]Labai įdomiai šią problemą aprašo S.J.C. Granger (1999)

nevaidina jokio vaidmens. Modelis, kuriame laikas vaidina svarbų vaidmenį, vadinams *dinaminiu*. Dinaminiais modeliais tiriamas ekonominių reiškinių kitimas laike. Pavyzdžiui, remiantis dinamine paklausos teorija, įtaką prekės paklausai daro vartotojo pajamos ankstesniais laiko periodais (žymėkime Y_{t-1}, Y_{t-2}) bei vyriausybės vykdoma kreditavimo politika (išreiškime ją dydžiu G_t). Todėl paklausos modelis galėtų būti ir toks:

$$q_t = f(p_t, p_{0,t}, Y_t, T_t, Y_{t-1}, Y_{t-2}, G_t) + \varepsilon_t. \tag{1.4}$$

Čia indeksas t išreiškia atitinkamo dydžio reikšmę laiko momentu t. (1.4) lygtimi aprašomas modelis yra dinaminis. Labai dažnai dinaminiame modelyje egzogeniniais yra *pavėlinti endogeniniai* kintamieji. Taip, pavyzdžiui, (1.4) modelį galima modifikuoti, kintamąjį Y_{t-2} pakeitus kintamuoju q_{t-1}. Modeliai, kuriuose yra pavėlinti endogeniniai kintamieji, visada yra dinaminiai.

Ekonometrijoje iš esmės nagrinėjami trijų tipų modeliai: *regresiniai, simultaninių lygčių* ir *laiko eilučių*. Trumpai tuos modelius apžvelgsime.

1.3.1 REGRESINIAI MODELIAI

Žodžio „regresija" kilmė aiškinama švedų mokslininkų tyrimais, atliktais siekiant nustatyti tėvų ūgio nuokrypio nuo vidutinio ir jų suaugusių vaikų ūgio nuokrypio nuo vidutinio statistinį sąryšį. Tyrimo metu buvo patikrinta natūrali hipotezė: aukštų tėvų sūnūs aukšti, ir atvirkščiai. Kartu buvo pastebėta regresijos tendencija – sūnų ūgio regresija prie vidutinio, t.y. labai aukštų tėvų sūnūs vidutiniškai yra aukšti, bet jau ne tokie aukšti kaip tėvai ir atvirkščiai, žemų tėvų sūnūs yra vidutiniškai žemaūgiai bet didesni nei jų tėvai.

Regresiniais modeliais aprašomas endogeninių kintamųjų (paaiškinamųjų, regresantų), sakykime tie kintamieji yra $Y_1, \ldots, Y_d$, elgesys egzogeninių kintamųjų (paaiškinančiųjų kintamųjų arba regresorių), tarkime, $X_1, \ldots, X_q$ atžvilgiu. Bendrasis regresinis modelis su adytyvia[4] paklaida aprašomas lygčių sistema

$$Y_j = f_j(X_1, \ldots, X_q) + \varepsilon_j, \quad j = 1, \ldots, d. \tag{1.5}$$

Čia $f_j : R^q \to R, j = 1, \ldots, d$ yra nežinomos funkcijos, $\varepsilon_j \in R$ – atsitiktinės modelio paklaidos. Jų prigimtis, kaip jau minėta, gali būti įvairi: gali

[4]Analogiškai galima nagrinėti ir modelius su multiplikatyvia paklaida, $Y_j = f_j(X_1, \ldots, X_q)\varepsilon_j$.

atspindėti praleistus faktorius, netikslius matavimus, prigimtinę kintamųjų stochastiką ir pan.

(1.5) modelį patogu užrašyti vektoriniu būdu:

$$\boldsymbol{Y} = \boldsymbol{f}(\boldsymbol{X}) + \boldsymbol{\varepsilon}. \tag{1.6}$$

Čia $\boldsymbol{Y} = (Y_1, \ldots, Y_d)'$ – endogeninių kintamųjų vektorius[5], $\boldsymbol{X} = (X_1, \ldots, X_q)'$ – egzogeninių kintamųjų vektorius, $\boldsymbol{f} : R^q \to R^d$ – nežinoma funkcija, $\boldsymbol{f}(\boldsymbol{x}) = (f_1(x_1, \ldots, x_q), \ldots, f_d(x_1, \ldots, x_q))'$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_q)'$ – modelio paklaidų vektorius.

Aprašant modelį nustatomas ir atitinkamas duomenų generavimo mechanizmas. Tai yra, jei $Y_{jt}, t = 1, \ldots, T$, $X_{it}, t = 1, \ldots, T$ yra imtys, atitinkančios kintamuosius $Y_j, j = 1, \ldots, d$ ir $X_i, i = 1, \ldots, p$, kurie susieti (1.5) modeliu, tai tų imčių generavimo modelis yra

$$Y_{jt} = f_j(X_{1t}, \ldots, X_{qt}) + \varepsilon_{tj}, \quad j = 1, \ldots, d; t = 1, \ldots, T. \tag{1.7}$$

Čia $\varepsilon_{tj}, t = 1, \ldots, T; j = 1, \ldots, d$ – atsitiktiniai dydžiai. Jie atspindi ne tik bendrą neapibrėžtumą, bet ir galimus stochastinius sąryšius tarp imties elementų. Tie sąryšiai gali kilti dėl įvairių priežasčių. Labai dažnai tai susiję su duomenų struktūra, jų agregavimu, surinkimo metodu ir pan. Paprastai (1.7) sistema užrašoma vektoriniu būdu:

$$\boldsymbol{Y}_t = \boldsymbol{f}(\boldsymbol{X}_t) + \boldsymbol{\varepsilon}_t, \quad t = 1, \ldots, T. \tag{1.8}$$

Tolesnį modelio specifikavimą apsprendžia prielaidos funkcijai $\boldsymbol{f}$ bei modelio paklaidoms $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_T$. Pagal funkcijos $\boldsymbol{f}$ formą skiriami parametriniai ir neparametriniai modeliai.

Parametrinis modelis reiškia, kad funkciją $\boldsymbol{f}$ pilnai aprašo baigtinis parametrų skaičius. Tai yra, $\boldsymbol{f}(\cdot) = \boldsymbol{f}(\cdot; \boldsymbol{\beta})$ yra žinoma funkcija, priklausanti nuo nežinomo parametro $\boldsymbol{\beta} \in R^p$. Taigi bendrasis parametrinio vienos lygties regresinio modelio pavidalas yra

$$\boldsymbol{Y} = \boldsymbol{f}(\boldsymbol{X}; \boldsymbol{\beta}) + \boldsymbol{\varepsilon}, \quad \boldsymbol{\beta} = (\beta_1, \ldots, \beta_p), \tag{1.9}$$

kai $f(\cdot; \boldsymbol{\beta})$ – žinoma funkcija. Atskiru atveju, kai $\boldsymbol{f}$ yra tiesinė funkcija parametrų atžvilgiu, gaunamas tiesinis modelis. Pavyzdžiui, šis modelis yra tiesinis:

$$\boldsymbol{Y} = \boldsymbol{B}\boldsymbol{X} + \boldsymbol{\varepsilon}. \tag{1.10}$$

[5] Vektorius visada reiškia vektorių-stulpelį. Jei x yra vektorius, tai x' – transponuotas vektorius arba vektorius-eilutė.

Čia $\boldsymbol{B} = (b_{ij}, i = 1, \ldots, d; j = 1, \ldots, q)$ yra $d \times q$ matrica. Šiuo atveju imčių generavimo mechanizmas yra

$$\boldsymbol{Y}_t = \boldsymbol{B}\boldsymbol{X}_t + \boldsymbol{\varepsilon}_t, \quad t = 1, \ldots, T. \tag{1.11}$$

Tiesiniu taip pat vadintume modelį

$$\boldsymbol{Y} = \boldsymbol{B}\log(\boldsymbol{X}) + \boldsymbol{\varepsilon}, \tag{1.12}$$

kai $\log(\boldsymbol{x}) = (\log x_1, \ldots, \log x_p)'$. Vėliau matysime, kad tokie modeliai pasitaiko gana dažnai.

Jei funkcijos $\boldsymbol{f}$ parametrizuoti negalima, modelis vadinamas *neparametriniu*[6].

Modelio paklaidoms dažniausiai nagrinėjami įvairūs koreliaciniai sąryšiai. Su jais susipažinsime vėliau.

1.3.2 Simultaninių lygčių modeliai

Jei dauguma kitų modelių, naudojamų ekonominių duomenų analizei, paprastai yra adaptuoti statistikiniai modeliai, tai simultaninių lygčių modeliai yra saviti. Tą savitumą sąlygoja tai, kad nagrinėjami ekonominiai dydžiai nustatomi visi vienu metu per pusiausvyros sąlygas. Pavyzdžiui, panagrinėkime regresinį modelį, kuris aprašo nupirkto prekės kiekio ir jos kainos sąryšį laiko momentu t:

$$q_t = a + bp_t + cZ_t + \varepsilon_t. \tag{1.13}$$

Kokia šio modelio ekonominė prasmė? Ar tai paklausos, ar pasiūlos lygtis? Ekonomijos teorija teigia, kad tiek paklausa, tiek kaina nusistato kartu rinkoje ir juos reikia nagrinėti kartu kaip endogeninius kintamuosius. Tokiu būdu (1.13) modelis, nagrinėjamas izoliuotai, nepakankamas norint statistikiniams sąryšiams suteikti ekonominę prasmę.

Simultaninių lygčių modeliai tai sąryšiai tarp ekonominių rodiklių, aprašomi lygčių sistemomis, į kurias įeina endogeniniai bei paaiškinamieji kintamieji ir modelio paklaidos.

Paprastai lygtys sudarančios simultaninių lygčių modelius yra trijų tipų: ekonominės tapatybės, techninės (įstatyminės) lygtys ir elgsenos lygtys.

---

[6]Dar nagrinėjami taip vadinamieji pusiauparametriniai modeliai, bet tai neįeina į mūsų kursą.

- *Ekonominės tapatybės.* Pavyzdžiui, pajamos apibrėžiamos kaip suvartotų ir sutaupytų pinigų suma; darbo užmokesčio fondą sudaro darbo laiko ir norminio (valandinio) užmokesčio sandauga ir t.t.

- *Techninės arba įstatyminės lygtys.* Jų prigimtį apsprendžia besikeičiančios technologijos arba besikeičianti ekonominė aplinka. Pavyzdžiui, produkcijos gamybos lygtys susieja gamybinę medžiagą ir galutinį produktą atsižvelgiant į technologiją.

- *Elgsenos lygtys.* Jomis aprašoma tam tikrų individų elegsena ekonominėje aplinkoje. Pavyzdžiui, konkretaus produkto vartojimo funkcija priklauso nuo to, kaip firmos atžvilgiu elgsis vartotojas arba koks yra bedarbystės lygis ir pan.

Paprastai atrojo ir trečiojo tipų lygtys yra stochastinės, o pirmojo – ne. Simultaniniai modeliai turi tris formas: *struktūrinę, redukuotą ir galutinę.*

Bendras struktūrinis simultaninių lygčių modelis yra

$$\boldsymbol{F}(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\varepsilon}) = 0; \tag{1.14}$$

čia

$\boldsymbol{Y} = (Y_1, \ldots, Y_d)'$ – endogeninių kintamųjų vektorius,

$\boldsymbol{X} = (X_1, \ldots, X_q)'$ – paaiškinančiųjų kintamųjų vektorius,

$\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_d)'$ – paklaidų vektorius, $\boldsymbol{F} = (F_1, \ldots, F_d) : R^{2d+q} \to R^d$ – nežinoma vektorinė funkcija.

Funkcija $\boldsymbol{F}$ aprašo bendrą, simultaninį[7] kintamųjų elgesį sistemoje. Todėl tie modeliai ir vadinami simultaninių lygčių modeliais. Reikia dar kartą atkreipti dėmesį į tai, kad (1.14) yra lygčių sistema ir joje lygčių būtinai turi būti tiek, kiek yra endogeninių kintamųjų.

Jei $\boldsymbol{Y}_t, \boldsymbol{X}_t, t = 1, \ldots, T$ yra atitinkamai endogeninių kintamųjų $\boldsymbol{Y}$ ir paaiškinančiųjų kintamųjų $\boldsymbol{X}$, susietų (1.14) modeliu, imtys, tai jų generavimo mechanizmas aprašomas modeliu

$$\boldsymbol{F}(\boldsymbol{Y}_t, \boldsymbol{X}_t, \boldsymbol{\varepsilon}_t) = 0, \quad t = 1, \ldots, T. \tag{1.15}$$

Tolesnė specifikacija atitinka funkcijos $\boldsymbol{F}$ parinkimą bei daromas prielaidas atsitiktiniams vektoriams $\boldsymbol{\varepsilon}_t, = 1, \ldots, T$, kurie, kaip ir regresinio modelio

[7] Angl. *simultaneous* reiškia vykstantis, egzistuojantis vienu metu.

atveju, aprašo ir galimus duomenų sąryšius. Kaip regresiniai, taip ir simultaninių lygčių modeliai yra parametriniai ir neparametriniai. Prielaidos paklaidoms $\boldsymbol{\varepsilon}_t, t = 1, \ldots, T$ taip pat yra analogiškos.

Jei funkcija $\boldsymbol{F}$ yra tiesinė parametrų atžvilgiu, tai gaunamas simultaninių lygčių tiesinis modelis. Pavyzdžiui,

$$\Gamma \boldsymbol{Y}_t + \boldsymbol{B}\boldsymbol{X}_t = \boldsymbol{\varepsilon}_t, \quad t = 1, \ldots, T. \tag{1.16}$$

Čia, $\Gamma = (\gamma_{jl})$ yra endogeninių kintamųjų koeficientų neišsigimusi $d \times d$ matrica, o $B = (b_{ij})$ – egzogeninių (arba pavėlintų endogeninių) kintamųjų $d \times q$ matrica.

Redukuota modelio forma gaunama išsprendus (1.15) lygtį endogeninių kintamųjų atžvilgiu. Jei modelis yra dinaminis, tai jos sprendinys endogeninių kintamųjų atžvilgiu yra galutinė modelio forma. Abi minėtos simultaninių lygčių formos naudingos vertinant parametrus, skaičiuojant įvairius ekonominius rodiklius ir pan.

Simultaninių lygčių modeliai daugiausia naudojami makroekonomikos poreikiams, todėl dažnai dar vadinami makroekonometriniais. Iš esmės visi makroekonometriniai modeliai turi panašius bazinius elementus: nacionalinių pajamų apibrėžimą (arba keletą apibrėžimų); vartojimo funkciją (arba grupę tokių funkcijų); investicijų funkciją (arba grupę tokių funkcijų):

$$\begin{aligned} Y &= C + I + G \\ C &= C(Y, \ldots, u), \\ I &= I(Y, \ldots, v). \end{aligned}$$

Kuo daugiau makroekonomikos aspektų atspindi modelis, tuo daugiau jame lygčių bei kintamųjų. Pirmieji pokariniai (vadinamieji JAV tarpukario ekonomikos (1921–1941 metų) Kleino modeliai) turėjo tris stochastines ir tris nestochastines lygtis, kuriose buvo šeši endogeniniai ir keturi egzogeniniai kintamieji. Didelę įtaką ekonometrijai turėjo Kleino-Goldbergerio JAV ekonomikos, apimančios 1921–1941 ir 1946–1952 metus, modelis, kuriame buvo 15 stochastinių ir 5 nestochastinės lygtys su 20 endogeninių ir 14 egzogeninių kintamųjų. Prie rekordinių (kintamųjų skaičiaus atžvilgiu) priskiriamas 1960 metų JAV ekonomikos modelis, kuriame buvo 176 endogeniniai ir 89 egzogeniniai kintamieji. Aišku, šiuo modeliu norėta taip pat pademonstruoti išaugusį tuo laikotarpiu modelių kūrimo meną.Yra ir dar didesnių modelių, pavyzdžiui, 1972 metų industrijos modelį sudarė 346 endogeniniai ir 90 egzogeninių kintamųjų.

1.3.3 Laiko eilučių modeliai

Regresiniai ar simultaninių lygčių modeliai susieja tiriamus egzogeninius kintamuosius su paaiškinamaisiais dydžiais. Žinant paaiškinamųjų dydžių elgesį, galima prognozuoti ir tiriamųjų dydžių reikšmes. Visai kita ideologija remiasi laiko eilučių modeliai.

Laiko eilute vadinama kokio nors dydžio, tarkime, y stebėjimų laike seka $y(1), y(2), \ldots, y(T)$. *Stochastinių laiko eilučių modeliai* remiasi prielaida, kad tie stebėjimai yra generuoti stochastinio proceso[8] $(Y(t), t \in T)$. Čia T yra laiką atitinkanti parametrų aibė, ir tai gali būti, pavyzdžiui, Z, N ar pan.

Laiko eilutės labai naudingos trumpalaikiam prognozavimui. Vietoje modelio, surišančio tiriamo ekonominio kintamojo reikšmes su paaiškinamaisiais dydžiais, vienamačiai laiko eilučių modeliai susieja tiriamo dydžio reikšmę dabartiniu laiko momentu su jo reikšmėmis ankstesniais laiko momentais ir triukšmo dabartine bei ankstesnėmis reikšmėmis. Prie tokių modelių priskiriami, pavyzdžiui, autoregresiniai, slenkančio vidurkio, autoregresiniai integruoti slenkančio vidurkio modeliai ir t.t.

Pavyzdžiui, sakoma, kad laiko eilutė $(y(1), \ldots, y(T))$ aprašoma ARMA(p, q) modeliu, jei

$$y(t) = \phi_1 y(t-1) + \cdots + \phi_p y(t-p) + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q},$$

kai $(\varepsilon_t, t \in Z)$ (vadinamasis baltasis triukšmas) yra nepriklausomi standartiniai normaliniai atsitiktiniai dydžiai.

1.4 Ekonominiai duomenys

Paprastai ekonometrinė analizė naudoja neeksperimentinius duomenis. Jie būna įvairių tipų ir gali aprašyti tiek kiekybinius, tiek kokybinius ekonominius rodiklius. Kiekybinai rodikliai išreiškiami skaičiais.

Laiko eilutės (chronologiniai arba laiko aplinkybės) duomenys aprašo konkretų kintamąjį skirtingais laiko momentais. Laiko intervalai ekonomikoje dažniausiai būna metai, ketvirčiai, mėnesiai. Žemo, vidutinio ir aukšto dažnumo laiko eilutės skiriasi laiko intervalo ilgiu. Žemo dažnumo duomenis

[8] Stochastiniu procesu vadinamas atsitiktinių dydžių, apibrėžtų vienoje tikimybinėje erdvėje, rinkinys.

atitinka ilgi laiko intervalai – metai, ketvirčiai. Vidutinio dažnumo – mėnuo, savaitė, diena. O aukšto dažnumo – valandos, minutės. Aišku, kuo duomenų dažnumas didesnis, tuo didesni jų masyvai, tuo sudėtingesni jų apdorojimo metodai.

Vietos aplinkybės duomenys aprašo konkretų kintamąjį fiksuotu laiko momentu, bet skirtingoms vietovėms ar skirtingoms rūšims. Vietovės gali būti miestai, šalys, rajonai, regionai ir t.t., o rūšys priklauso nuo duomenų tipo. Pavyzdžiui, tiriant išlaidas vartojimui, galima išskirti maisto produktus, aptarnavimo sferą ir pan.

Kartais šių dviejų tipų duomenys yra sumaišyti (vadinamieji mišrūs duomenys) arba *paneliniai*. Vietos aplinkybės laiko eilutė aprašys kintamąjį tiek skirtingoms vietovėms, tiek ir skirtingais laiko momentais. Pavyzdžiui, vienam šalies gyventojui tenkanti nacionalinių pajamų dalis rinkos kainomis 1969–1972 metų laikotarpiu.

Reikia pastebėti, kad to paties modelio parametrų įvertinimai gali skirtis naudojant skirtingo tipo duomenis. Pavyzdžiui, žinoma, kad įvertintas pajamų elastingumas paklausai naudojant vietos aplinkybės duomenis yra didesnis nei įvertinimas, gautas naudojant laiko eilutes.

Kur gauti duomenų? Tam naudojami biuleteniai, knygos, laikmenos, internetas.

Duomenys dar yra: inžineriniai (duomenys apie techninius reikalavimus), įstatyminiai (pavyzdžiui, aprašantys muitų politiką) bei duomenys, sukurti ekonometristo (fiktyvūs). Pastarieji dažnai gaunami aprašant kokybinius ekonominius rodiklius.

Yra keletas problemų, susijusių su duomenimis, į kurias ekonometristas visada privalo atsižvelgti. Pirmoji – *laisvės laipsnių problema*. Gali atsitikti taip, kad surinkti duomenys neturės pakankamai informacijos, kurios pakaktų modelio įvertinimui. Ši problema atsiranda todėl, kad ekonominiai duomenys nėra eksperimentiniai. Tokiais atvejais dažnai padeda modelio supaprastinimas (mažiau paaiškinamųjų kintamųjų, paprastesnė forma ir t.t.).

Nemažai problemų tiek vertinant modelį, tiek jį taikant kelia taip vadinamasis dydžių agregavimas.

- *Individualių dydžių agregavimas.* Pavyzdžiui, visuminės pajamos yra individualių pajamų suma; bendros gamybos apimtys yra atskirų apimčių suma ir t.t.

- *Prekių agregavimas.* Galima apjungti įvairias plataus vartojimo prekes (naudojant indeksavimą) arba prekių grupes pagal kainą. Pavyzdžiui, vertinant maisto prekių paklausos priklausomybę nuo bendrųjų pajamų, prekių kainų ir kitų prekių kainų, visi dydžiai yra agreguoti.

- *Agregavimas laike.* Stebimi duomenys dažnai atitinka kitą laikotarpį nei tiriamasis periodas. Pvz., daugelio vartojimo prekių gamintojų produkcija sukomplektuojama per trumpesnį nei metai periodą. Todėl naudojant dar ir metinius duomenis gali atsirasti paklaidos.

- *Erdvinis agregavimas.* Pavyzdžiui, miesto, šalių gyventojai; produkcija regionais, šalimis ir t.t.

Bet koks duomenų suvidurkinimas gali duoti paklaidas, kurias labai svarbu mokėti įvertinti ir tirti. Tam būtina išsiaiškinti agregavimo pobūdžius, galimus poveikius ir ieškoti būdų, kaip išvengti kylančių problemų.

Multikolinearumo problemos. Tam tikras multikolinearumo lygis iš anksto užprogramuotas tarp ekonominių kintamųjų. Pavyzdžiui, pajamos, darbo vietos, vartojimas, investicijos, eksportas, importas ir pan. yra priklausomi dydžiai. Kainos ir darbo užmokestis turi tendenciją didėti kartu. Todėl įtraukus į kintamųjų sąrašą abu šiuos kintamuosius neišvengsime kolinearumo problemos.

Serijinės koreliacijos problemos atsiranda naudojant laiko eilutes. Jos dažniausiai atsiranda todėl, kad ekonominių sistemų pasikeitimai laike vyksta labai lėtai.

Matavimo paklaidos. Duomenys gaunami netiksliai atlikus matavimus. Atsiranda matavimų paklaidos ir tai iš esmės keičia modelio interpretavimą.

Struktūriniai pasikeitimai. Duomenys gali būti surinkti iki ir po akivaizdaus struktūrinio pasikeitimo. Pavyzdžiui, iki karo ir po karo. Labai svarbu mokėti atskirti struktūrinius pasikeitimus.

Labai dažnai duomenis, prieš jais naudojantis, reikia apdoroti. Interpoliacija, ekstrapoliacija naudojama norint papildyti imtis, kai yra praleistų stebėjimų arba jei prognozuojame.

Duomenų suglodinimas naudojamas siekiant pašalinti trendus, sezoniškumus ir pan. Paprastai bet kurią laiko eilutę x_t galime išskaidyti į keturias pagrindines komponentes: trendą T (aprašo ilgalaikę tendenciją), ciklą C (aprašo sinusinį pasikeitimą), sezoniškumą S (aprašo ciklišką pasikeitimą per vieną laiko periodą, paprastai metus), I – nereguliarią komponentę;

$$x = T \cdot C \cdot S \cdot I.$$

Pavyzdžiui, jei trendas $T_t = e^{at}$, o cikliškumas – $C_t = \cos(\theta t + \phi)$, tai laiko eilutėje

$$\widehat{x}_t = \frac{x_t e^{-at}}{\cos(\theta t + \phi)}$$

tiek trendas, tiek cikliškumas yra pašalinti. Kitas galimas būdas – nagrinėti logaritmų eilutę:

$$y_t = f(t) + u_t;$$

čia $f(t)$ aprašo reguliarią duomenų dalį $(\log T_t + \log C_t + \log S_t,)$ o u_t – stochastinę $(u_t = \log I_t.)$

2 skyrius. Paprasčiausias tiesinis regresinis modelis

2.1 Ekonominis uždavinys

Norėdami išsiaiškinti pagrindinius regresinio modelio principus, nagrinėsime paprastą, bet labai svarbų ekonominį uždavinį.

Tarkime, reikia išsiaiškinti sąryšį tarp *namų ūkio pajamų* ir *išlaidų maistui*. Pavadinkime tai „pajamų-išlaidų maistui (PIM)" uždaviniu. Tyrimo tikslas – atsakyti į tokius svarbius klausimus kaip, pavyzdžiui, šie:

- Kiek vidutiniškai padidėtų išlaidos maistui, jei pajamos padidėtų, tarkime, 100 Lt/mėn.?
- Ar gali vidutinės išlaidos maistui sumažėti didėjant pajamoms?
- Kokias vidutines išlaidas maistui galėtume prognozuoti namų ūkiams, kurių mėnesio pajamos siekia 800 Lt/mėn.?

Atsakymai į šiuos ir panašius klausimus suteiktų reikšmingos informacijos tiems, kurie priima sprendimus užsakant maisto produktus prekybos tinkluose, planuojant prekybos taškų išdėstymą kuriame nors mikrorajone ir pan.

Uždaviniui spręsti pirmiausia turime sudaryti ekonominį-matematinį modelį. Tiriant dviejų ar daugiau dydžių tarpusavio priklausomybę nekyla jokių problemų pasirenkant kintamuosius. Dažniausiai jie yra savaime aiškūs. Taip pat ir "pajamų-išlaidų maistui uždavinyje". Endogeninis kintamasis (regresantas, paaiškinamasis kintamasis) – namų ūkio išlaidos maistui. Pažymėkime jį Y. Egzogeninis kintamasis (regresorius, paaiškinantysis kintamasis) – namų ūkio pajamos X. Taigi norime nustatyti ryšį tarp kintamųjų Y ir X. Pirmiausia pažiūrėkime, ką šiuo klausimu sako ekonomikos teorija. Daugelis ekonominių vadovėlių rekomenduoja tiesinį sąryšį tarp pajamų ir vidutinių išlaidų maistui. Pasiaiškinkime, ką šiame kontekste reiškia vidutinės išlaidos. Norėdami susigaudyti situacijoje tarkime, kad mus domina tik tie namų ūkiai, kurių pajamos yra 700 Lt./mėn., arba 9 400 Lt. per metus. Atsitiktinai atrinkę tokius namų ūkius, kiekvieno iš jų turime paklausti: „Kiek jūsų namų ūkis išleidžia maistui ¿' Apklausę visus atrinktus namų ūkius (tarkime, jų atrinkome n), gausime imtį $y_1, \dots, y_n$. Tai bus dydžio Y reikšmės, kai $X = 700$. Akivaizdu, kad dydis Y yra atsitiktinis. Kol neuždavėme klausimo ir neišgirdome atsakymo, tol nežinome jo reikšmės. Ką galime daryti su imtimi $y_1, \ldots, y_n$? Galime suskaičiuoti vidutinę reikšmę $\overline{y} = n^{-1}(y_1 + \cdots + y_n)$. Ką ji reiškia? Nesunku suprasti, kad taip gausime sąlyginę vidutinę išlaidų maistui reikšmę su sąlyga, kad $X = 700$. Tai teorinio sąlyginio vidurkio, kurį teorijoje priimta žymėti $\mu_{Y|700} = E(Y|X = 700)$, statistinis įvertinimas $\widehat{\mu}_{Y|700}$. Galime nusipaišyti histogramą ir tai bus sąlyginė histograma su sąlyga, kad $X = 700$. Galime suskaičiuoti statistinę dispersiją arba statistinį kvadratinį nuokrypį. Bet kuri suskaičiuota statistinė charakteristika bus sąlyginė su sąlyga, kad $X = 700$.

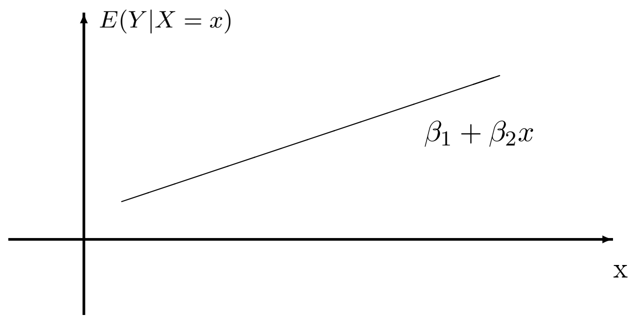
Toliau tarkime, kad tiriame tik tuos namų ūkius, kurių pajamos siekia 800 Lt. per mėnesį. Analogiškai elgdamiesi surasime sąlyginio vidurkio $\mu_{Y|800} = E(Y|X = 800)$ statistinį įvertį $\widehat{\mu}_{Y|800}$ ir t.t. Apibendrinę tuos samprotavimus matome, kad bendru atveju galime kalbėti apie sąlyginį vidurkį $\mu_{Y|x} = E(Y|X = x)$ – vidutines išlaidas maistui, kai fiksuotos pajamos $X = x$. Ekonomikos teorija kaip tik rekomenduoja pabandyti tiesinę vidutinių išlaidų maistui, kai fiksuotos pajamos, priklausomybę nuo pajamų. Mūsų žymėjimais

$$E(Y|X = x) = \beta_1 + \beta_2 x. \tag{2.1}$$

Tačiau visada reikia prisiminti, kad tai tik rekomendacija. Tais atvejais, kai turimi duomenys blogai atspindi šią prielaidą, reikia ieškoti kitokiu funkcinių priklausomybių. Dažnai tai galima nesunkiai pastebėti atvaizdavus duomenis. Nors duomenų vizualizavimas labai svarbus etapas parenkant modelį, tačiau labai piktnaudžiauti tuo negalima. Negalima ieškoti modelio tik turimiems duomenims.

(2.1) modelis patrauklus pirmiausia tuo, kad jame yra tik du nežinomi išreikštiniai parametrai. β_1 vadinamas *laisvuoju nariu* („intercept") ir aprašo vidutines išlaidas maistui tų namų ūkių, kurių mėnesio pajamos yra nulinės ($x = 0$) (apie tą parametrą dar bus kalbama). Parametras β_2 vadinamas *marginaliniu polinkiu išleisti maistui* ir aprašo vidutinių išlaidų maistui pasikeitimą pajamoms pakitus vienu litu:

$$\beta_2 = \frac{\Delta E(Y|X = x)}{\Delta x} = \frac{dE(Y|X = x)}{dx}.$$



Grafikas 2.1: Tiesinė vidutinių išlaidų priklausomybė nuo pajamų

2.2 Ekonometrinis modelis

Priminsime, kad $E(Y|X)$ yra atsitiktinis dydis, sąlyginis atsitiktinio dydžio Y vidurkis dydžio X atžvilgiu. Klasikinė regresijos teorija remiasi tuo, kad bet kurį atsitiktinį dydį Y (kurio dispersija yra baigtinė) galima išskaidyti

$$Y = E(Y|X) + \varepsilon. \tag{2.2}$$

Čia ε yra toks atsitiktinis dydis, kad $E\varepsilon = 0$ ir $E\varepsilon X = 0$. Paaiškinsime tiksliau. Jei Y yra atsitiktinis dydis su baigtine dispersija $\sigma^2 = EY^2 - (EY)^2$, tai $Y = E(Y|X) + (Y - E(Y|X)) = E(Y|X) + \varepsilon$. Atsitiktinis dydis $\varepsilon = Y - E(Y|X)$ yra ortogonalus X ta prasme, kad $E\varepsilon X = 0$, nes $E\varepsilon X = E(Y - E(Y|X))X = EYX - EXE(Y|X) = EYX - EXY = 0$. (2.1) sąryšį galime perrašyti taip:

$$E(Y|X) = \beta_1 + \beta_2 X. \tag{2.3}$$

Taip specifikavę tiesinę vidutinių išlaidų maistui priklausomybę nuo pajamų ir atsižvelgę į (2.2) sudarome regresinį modelį:

$$Y = \beta_1 + \beta_2 X + \varepsilon. \tag{2.4}$$

Tai ir yra paprasčiausias tiesinis regresinis modelis. Paprasčiausias todėl, kad šiame modelyje yra tik vienas regresantas X. Tačiau tuo ekonometrinio modelio kūrimas dar nesibaigia. Toliau reikės įvertinti parametrus. Tam reikalingi duomenys. Kaip surenkami duomenys, čia nediskutuosime. Tam bus skirtos atskiros paskaitos. Čia gi tarkime, kad imtis $(Y_t, X_t), t = 1, \ldots, T$ gauta stebint porą (Y, X). Grubiai šnekant, kažkokiu būdu atrinkti namų ūkiai, kuriems buvo užduodami du klausimai.

- Kiek jūsų namų ūkis išleidžia maistui per mėnesį?
- Kokios yra jūsų namų ūkio mėnesio pajamos?

Pirmojo namų ūkio atsakymai yra (Y_1, X_1), antrojo – (Y_2, X_2) ir t.t.

Kitu žingsniu reikia aprašyti duomenų generavimo mechanizmą. Tardami, kad išlaidų maistui ir pajamų sąryšis yra aprašomas (2.4) formule, kartu teigiame, kad atitinkamas imties generavimo modelis yra

$$Y_t = \beta_1 + \beta_2 X_t + \varepsilon_t, \quad \text{su visais} \quad t = 1, \ldots, T. \tag{2.5}$$

Čia $\varepsilon_1, \ldots, \varepsilon_T$ yra atsitiktiniai dydžiai. Jie atspindi duomenyse slypinčius neapibrėžtumus, surinkimo metodiką ir pan. Kaip jau minėta, jie dar vadinami modelio triukšmu, inovacijomis, paklaidomis. Kitaip tariant, kiekvieną imties $Y_1, \ldots, Y_T$ narį traktuojame kaip atsitiktinį dydį. Tolesnę modelio specifikaciją apsprendžia prielaidos, daromos paklaidoms (ε_t). Tradiciškai, jos yra tokios.

- *Nulinių vidurkių prielaida.* Ji reiškia, kad
$$E\varepsilon_1 = \cdots = E\varepsilon_T = 0. \tag{2.6}$$

(2.5) modelio atveju ši prielaida ekvivalenti

$$E(Y_t|X_t) = \beta_1 + \beta_2 X_t \quad \text{su visais} \quad t = 1, \ldots, T.$$

O tai atitinka sąlyginio vidurkio specifikavimą.

Prielaida apie triukšmo nulinį vidurkį nėra esminė. Jei tarsime, kad $E\varepsilon_t = a$ su visais t, tai

$$\begin{aligned} Y_t &= \beta_1 + \beta_2 X_t + \varepsilon_t + (a - a) = (a + \beta_1) + \beta_2 X_t + (\varepsilon_t - a) \\ &= \beta_1' + \beta_2 X_t + \varepsilon_t' \end{aligned}$$

ir $E\varepsilon_t' = E(\varepsilon_t - a) = 0$. Tačiau ekonominė prasmė atliekant tokią transformaciją aiškiai keičiasi. Mat, jei paklaidomis aprašome praleistą informaciją, tai paklaidų vidurkis apibūdina vidutinį praleistos informacijos kiekį. Todėl nenulinį vidurkį būtinai reikia įvertinti.

- *Homoskedastiškumo prielaida.* Atsitiktiniai dydžiai $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_T$ turi nulinį vidurkį $E\varepsilon_t = 0$ ir pastovią dispersiją

$$\sigma^2 = var(\varepsilon_t) = E\varepsilon_t^2 \text{ su visais } t = 1, 2, \dots, T. \tag{2.7}$$

Egzogeninio kintamojo terminais tai reiškia, kad

$$var(Y_t|X_t) = E(Y_t - E(Y_t|X_t))^2|X_t = \sigma^2 \quad \textit{su visais} \quad t = 1, \dots, T.$$

Dispersija, kaip žinia, charakterizuoja atsitiktinio dydžio reikšmių išsibarstymą apie vidurkį. Nagrinėjamame pavyzdyje, $var(Y|X = 480)$ rodo, kaip gali skirtis išlaidos maistui nuo vidutinių išlaidų maistui tų namų ūkių, kurių pajamos yra 480 Lt. Jei tas išsibarstymas visai nepriklauso nuo pajamų, tai ir yra homoskedastinis atvejis.

Jei dispersija $\sigma_t^2 = E\varepsilon_t^2$ priklauso nuo t ir $\sigma_t^2 \neq \sigma_s^2$ kuriems nors $t \neq s$, tai sakome, kad modelis yra *heteroskedastinis*. Juos smulkiau nagrinėsime penktame skyriuje.

- *Nekoreliuotų paklaidų prielaida.* Tariame, kad a.d. $\varepsilon_1, \dots, \varepsilon_T$ turi nulinius vidurkius ir yra nekoreliuoti, t.y.

$$E\varepsilon_t\varepsilon_s = 0, \quad \text{kai} \quad t \neq s. \tag{2.8}$$

Tai taip pat reiškia, kad duomenų surinkimo mechanizmas duoda nekoreliuotus duomenis:

$$cor(Y_t, Y_s) = 0, \quad \text{kai} \quad t \neq s.$$

Vienas iš svarbiausių argumentų, kodėl ekonometrinis modelis yra iš esmės stochastinis yra duomenų surinkimo stochastiškumas (iš baigtinės populiacijos atsitiktinai parenkami respondentai). Todėl darant koreliacines prielaidas paklaidoms būtina turėti omeny duomenų surinkimo mechanizmą.

Labai dažnai tariama, kad modelio paklaidos yra Gausinės.

- *Gausinio triukšmo prielaida.* Atsitiktiniai dydžiai $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_T$ yra nepriklausomi ir

$$\varepsilon_t \sim N(0, \sigma^2) \text{ su kiekvienu } \quad t = 1, \dots, T. \tag{2.9}$$

Šiuo atveju duomenys turi sąlyginį Gausinį skirstinį, t.y.

$$P(Y_t \leq x|X_t) = P(N(\beta_1 + \beta_2 X_t, \sigma^2) \leq x), \quad x \in R.$$

2.3 Parametrų įvertinimas

Sudarius ekonometrinį modelį reikia įvertinti jo parametrus. Tam naudojame turimus duomenis. Pajamų-išlaidų maistui modeliui turime surinkę šiuos duomenis.

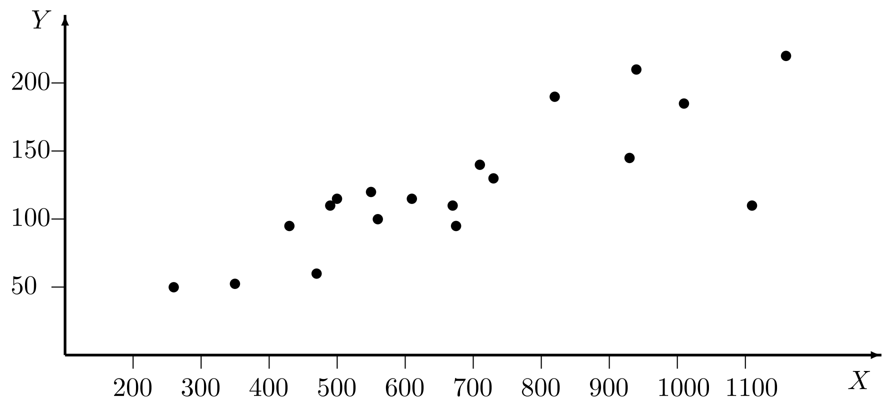
Stebinio *Nr.*	Išlaidos maistui	Pajamos per sav.	Stebinio *Nr.*	Išlaidos maistui	Pajamos per sav.
t	Y_t	X_t	t	Y_t	X_t
1	52.25	258.30	21	98.14	719.80
2	58.32	343.10	22	123.94	720.00
3	81.79	425.00	23	126.31	722.30
4	119.90	467.50	24	146.47	722.30
5	125.80	482.90	25	115.98	734.40
6	100.46	487.70	26	207.23	742.50
7	121.51	496.50	27	119.80	747.70
8	100/08	519.40	28	151.33	763.30
9	127.75	543.30	29	169.51	810.20
10	104.94	548.70	30	108.03	818.50
11	107.48	564.60	31	168.90	825.60
12	98.48	588.30	32	227.11	833.30
13	181.21	591.30	33	84.94	834.00
14	122.23	607.30	34	98.70	918.10
15	129.57	611.20	35	141.06	918.10
16	92.84	631.00	36	215.40	929.60
17	117.92	659.60	37	112.89	951.70
18	82.13	664.00	38	166.25	1014.00
19	182.28	704.20	39	115.43	1141.30
20	139.13	704.80	40	269.03	1154.60

Pajamų išlaidų maistui duomenys

Pirmiausia duomenis patartina atvaizduotus grafiškai (žr. 2.2 grafiką).

Iš grafiko matome, kad duomenys iš esmės neprieštarauja tiesiniam modeliui. Taigi, dar kartą konkretizuokime modelį. Tarkime, imties $(Y_t, X_t), t = 1, \ldots, T$ generavimo mechanizmas yra šis:

1. $Y_t = \beta_1 + \beta_2 X_t + \varepsilon_t$.



Grafikas 2.2: pajamu - išlaidų maistui duomenys

2. $E\varepsilon_t = 0$ (arba, ekvivalentiškai, $EY_t = \beta_1 + \beta_2 X_t$).

3. $var(\varepsilon_t) = \sigma^2 = var(Y_t|X_t)$.

4. $cov(\varepsilon_t, \varepsilon_s) = cov(Y_t, Y_s) = 0$, jei $s \neq t$.

5. $X_t \neq c$ su kiekvienu t.

Yra daugybė būdų, kaip įvertinti išreikštinius modelio parametrus (PIM modelyje koeficientus β_1 ir β_2).

2.3.1 Kovariacijų principas

Tiesiniame regresiniame modelyje abi puses padauginę iš X_t turime atsitiktinių dydžių lygybę

$$Y_t X_t = \beta_1 X_t + \beta_2 X_t^2 + \varepsilon_t X_t$$

su kiekvienu $t = 1, \ldots, T$. Suskaičiavę abiejų pusių vidurkius randame

$$EX_t Y_t = \beta_1 EX_t + \beta_2 EX_t^2 + EX_t \varepsilon_t.$$

Kadangi $EX_t\varepsilon_t = 0$ su kiekvienu t ir $EX_t = EX$, $EX_tY_t = EXY$ bei $EX_t^2 = EX^2$, sudarome sistemą

$$\begin{cases} EY = \beta_1 + \beta_2 EX \\ EYX = \beta_1 EX + \beta_2 EX^2. \end{cases}$$

Išsprendę šią sistemą randame

$$\beta_2 = \frac{EXY - EXEY}{EX^2 - (EX)^2}, \quad \beta_1 = EY - EX\frac{EXY - EXEY}{EX^2 - (EX)^2}.$$

Parametrų įverčius gauname į šias išraiškas įstatę atitinkamų dydžių statistikinius įverčius. Kaip žinome, atsitiktinio dyžio vidurkio $m(X) = EX$ statistikinis įvertinimas yra aritmetinis vidurkis

$$\widehat{m}(X) = \overline{X} = \frac{1}{T}\sum_{k=1}^{T} X_k.$$

Taigi, vidurkius EX ir EY įvertiname atitinkamai $\overline{X}$ ir $\overline{Y}$; vidurkį EXY – dydžiu

$$\overline{XY} = \frac{1}{T}\sum_{k=1}^{T} X_kY_k,$$

o vidurkį EX^2 – dydžiu

$$\overline{X^2} = \frac{1}{T}\sum_{k=1}^{T} X_k^2.$$

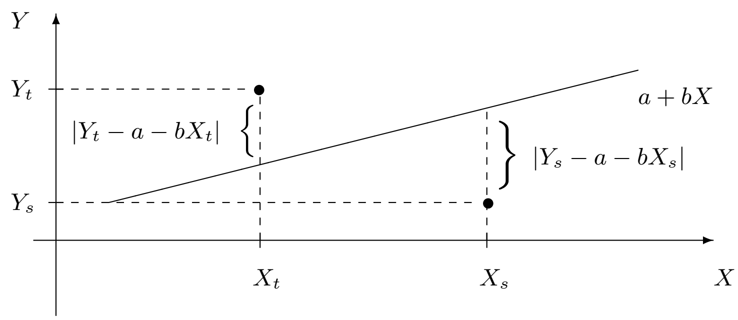
Taip gauname šiuos parametrų β_1 ir β_2 įvertinimus:

$$\widehat{\beta}_2 = \frac{\overline{XY} - \overline{X}\cdot\overline{Y}}{\overline{X^2} - (\overline{X})^2}, \tag{2.10}$$

$$\widehat{\beta}_1 = \overline{Y} - \overline{X}\frac{\overline{XY} - \overline{X}\cdot\overline{Y}}{\overline{X^2} - (\overline{X})^2}. \tag{2.11}$$

2.3.2 Mažiausių kvadratų principas

Nagrinėkime bet kurią tiesę $y = a + bx$. Taškuose X_t tos tiesės reikšmės yra $a + bX_t$, o jų nuokrypis nuo atitinkamų endogeninio kintamojo reikšmių yra $|Y_t - (a + bX_t)|$.



Grafikas 2.3: Mažiausių kvadratų principo paaiškinimas

Mažiausių kvadratų principo esmė – parinkti tiesę $a + bx$ taip, kad gaunamų nuokrypių kvadratų suma būtų mažiausia. Taigi, parametrų β_1, β_2 įvertinimas yra

$$(\widehat{\beta}_1, \widehat{\beta}_2) = \underset{a,b}{argmin} \sum_{t=1}^{T} (Y_t - (a + bX_t))^2.$$

Funkcijos

$$f(a, b) = \sum_{t=1}^{T} (Y_t - (a + bX_t))^2$$

minimumo tašką surasti nesunku. Taip gaunamai parametrų β_1 ir β_2 įvertinimai, kurie paprasčiausio tiesinio regresinio modelio atveju sutampa su gautais kovariacijų metodu:

$$\widehat{\beta}_2 = \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\overline{X^2} - (\overline{X})^2}, \tag{2.12}$$

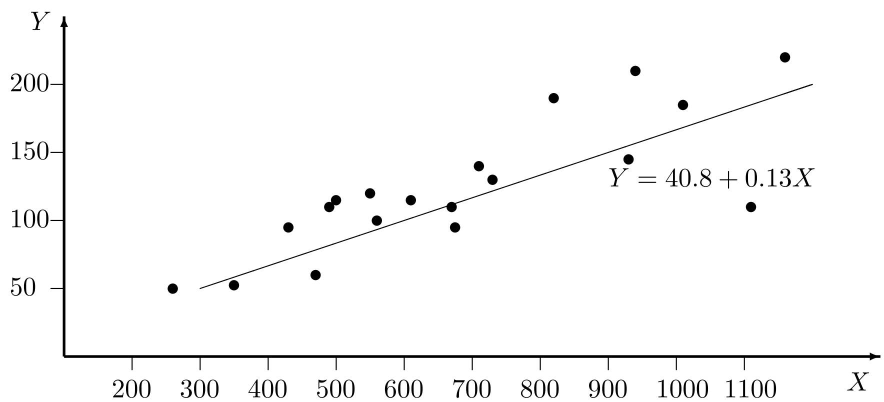
$$\widehat{\beta}_1 = \overline{Y} - \overline{X} \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\overline{X^2} - (\overline{X})^2}. \tag{2.13}$$

Iš formulių matyti, kam reikėjo prielaidos, jog regresoriaus reikšmės negali būti vienodos ($X_t \not\equiv c$ su visais $t = 1, \ldots, T$.). Jei $X_t = c$ su visais $t = 1, \ldots, T$, tai $\overline{X^2} - (\overline{X})^2$, irapskaičiuotieji parametrų įvertinimai netenka prasmės.

Į gautas (2.12) ir (2.13) išraiškas įstatę konkrečias imties reikšmes gausime parametrų β_2 ir β_1 įverčius. O formulės aprašo tų parametrų įvertinimus ir aišku, nepriklauso nuo konkrečios imties reikšmės. Tai labai svarbu. Taigi, tiek $\widehat{\beta}_1$, tiek $\widehat{\beta}_2$ aprašyti atitinkamai (2.12) ir (2.13) formulėmis yra atsitiktiniai dydžiai. Jų reikšmės nėra žinomos, kol nesurinkti duomenys arba kol neturime konkrečios imties realizacijos.

Naudodami lentelės duomenis, pajamų-išlaidų maistui uždavinyje surandame parametrų β_1 ir β_2 įverčius:

$$\widehat{\beta}_2 = \frac{40 \cdot 3834936.497 - 27920 \cdot 5212.520}{40 \cdot 21020623.02 - (27920)^2} = 0.1283,$$
$$\widehat{\beta}_1 = 40.7676.$$



Grafikas 2.4: Įvertintas išlaidų pajamų modelis

Dar kartą pabrėšime, kad įvertinimas yra formulė, kuria pasinaudoję galime suskaičiuoti konkrečius įverčius. Įvertis – konkreti parametro reikšmė, gauta į įvertinimą įstačius reikšmes, suskaičiuotas naudojant duotąją imtį.

2.3.3 Įverčių interpretacija

Kai parametrų įverčiai surasti, juos galime interpretuoti sprendžiamo ekonominio uždavinio kontekste. Reikšmė $\widehat{\beta}_2 = 0.1283$ yra parametro β_2

įvertis ir parodo, kiek padidėja savaitinės išlaidos maistui, kai pajamos padidėja vienu litu. Vadinasi, jei pajamos padidės 100 Lt per savaitę, tai vidutiniškai išlaidos maistui padidės 12.83 Lt. Supermarketų vadybininkas, gavęs informaciją apie galimą pajamų padidėjimą 100 Lt gali tikėtis parduoti už 12.83 Lt daugiau maisto prekių. Tai labai svarbi informacija ilgalaikiam planavimui.

Skaičius $\widehat{\beta}_1 = 40.7676$ yra parametro β_1 įvertis. Jis reiškia, griežtai kalbant, kiek vidutiniškai išleidžia maistui namų ūkis, kurio pajamos yra nulinės. Daugelyje ekonominių modelių reikia būti labai atsargiems interpretuojant laisvojo nario reikšmę. Svarbiausia problema yra ta, kad dažniausiai nėra pakankamai duomenų arti nulinės reikšmės. Būtent taip yra uždavinyje apie pajamas ir išlaidas maistui. Jei neturime pakankamai stebėjimų arti nulinės reikšmės, tai ir įvertinimas gali būti nepakankamai tikslus, tai yra, gali blogai atspindėti tiriamo regiono padėtį. Nors nagrinėjamame pajamų-išlaidų maistui pavyzdyje įvertintas modelis ir sako kad namų ūkiai su nulinėmis pajamomis vidutiniškai maistui išleidžia 40.7676 Lt per savaitę, būtų labai rizikinga tai suprasti paraidžiui.

Priminsime, kad ekonominio dydžio y elastingumas kito ekonominio dydžio x atžvilgiu yra

$$\eta = \eta(y|x) = \frac{y \text{ procentinis pasikeitimas}}{x \text{ procentinis pasikeitimas}} = \frac{dy}{dx} \cdot \frac{x}{y}.$$

Vidutinių išlaidų maistui elastingumas pajamų atžvilgiu yra

$$\eta = \frac{dE(Y)}{dX} \cdot \frac{X}{E(Y)}.$$

Jei tariame, kad išlaidų maistui priklausomybė nuo pajamų aprašoma tiesiniu modeliu, tai

$$\eta = \beta_2 \frac{X}{E(Y)}.$$

Norėdami įvertinti vidutinį elastingumą $E\eta$, parametrą β_2 galime pakeisti jo įverčiu $\widehat{\beta}_2$, o EX ir EY pakeisti atitinkamai $\overline{X}$ ir $\overline{Y}$. Taigi, vidutinių išlaidų maistui elastingumo pajamoms įvertinimas yra

$$\widehat{\eta} = \widehat{\beta}_2 \frac{\overline{X}}{\overline{Y}}.$$

Nagrinėjamame uždavinyje randame

$$\widehat{\eta} = 0.687.$$

Šis elastingumo įvertis rodo, kad savaitinėms namų ūkių pajamoms pakitus 1%, vidutinės išlaidos maistui padidės apytikriai 0.7%. Kai elastingumas yra mažesnis už vieną, išlaidos maistui klasifikuojamos kaip būtinos, o ne kaip prabangos dalykas.

Pagaliau, turėdami įvertintą modelį, galime daryti tam tikras prognozes. Tarkime, mus domina išlaidos maistui tų namų ūkių, kurių pajamos yra 750 Lt per savaitę. Tiesiniu modeliu tokią prognozę galime gauti įstatę $x = 750$ į įvertintą modelį. Gausime

$$\widehat{Y} = 40.7676 + 0.1283(750) = 136.98.$$

Taigi prognozė tokia: namų ūkiai, kurių pajamos yra 750 Lt per savaitę, maistui vidutiniškai išleidžia 136.98 Lt per savaitę.

2.3.4 Paklaidos dispersijos įvertinimas

Vienas svarbiausių paprasčiausios tiesinės regresijos parametrų yra paklaidų dispersija σ^2. Paprastai šis parametras yra nežinomas. Todėl ir jį reikia įvertinti. Jei $E\varepsilon_k = 0$, tai

$$\sigma^2 = E\varepsilon_k^2.$$

Prisiminę, kad vidurkio įvertinimui dažniausiai naudojamas imties empirinis vidurkis, galėtume manyti, kad σ^2 galima vertinti aritmetiniu vidurkiu $T^{-1}\sum \varepsilon_k^2$. Tačiau iš to įverčio nebūtų naudos, nes imties (ε_k) neturime (dydžiai ε_k nestebimi!). Bet turime tos imties įvertinimą – regresijos liekanas, kurios yra

$$\widehat{\varepsilon}_k = y_k - \widehat{y}_k = y_k - \widehat{\beta}_1 - \widehat{\beta}_2 X_k, \quad k = 1, \dots, T.$$

Todėl atrodo labai logiška nežinomą dispersiją įvertinti dydžiu $T^{-1}\sum \widehat{\varepsilon}_k^2$. Bet taip gautume paslinktą įvertinimą. Laimei, tą defektą nesunku ištaisyti, imant

$$\widehat{\sigma}^2 = \frac{1}{T-2}\sum_t \widehat{\varepsilon}_t^2.$$

Skaičius 2, kuris figūruoja vardiklyje, atspindi išreikštinių modelio parametrų skaičių, kuris paprasčiausios tiesinės regresijos atveju yra 2.

PIM modeliui suskaičiuojame

$$\widehat{\sigma}^2 = \frac{54311.3315}{38} = 1429.2456.$$

2.4 Mažiausių kvadratų įvertinimų savybės

Tarkime, turime imtį $(Y_t, X_t), t = 1, \ldots, T$, atitinkančią dydžius (Y, X). Nagrinėsime klasikinį tiesinį regresinį modelį (KTM), kuris aprašomas šiomis prielaidomis:

A1. $Y_t = b_1 + b_2 X_t + \varepsilon_t, \quad t = 1, \ldots, T;$

A2. $E\varepsilon_t = 0, \quad t = 1, \ldots, T;$

A3. $var(\varepsilon_t) = \sigma^2 > 0, \quad t = 1, \ldots, T;$

A4. $cov(\varepsilon_t, \varepsilon_s) = 0$, jei $t \neq s, t, s = 1, \ldots, T;$

A5. stebėjimai X_t yra deterministiniai, ir $X_t \neq const, \quad t = 1, \ldots, T.$

Jei, be išvardintų (A1) – (A5) sąlygų, dar teisinga

A6. $\varepsilon_t \sim N(0, \sigma^2), t = 1, \ldots, T,$

tai sakysime, kad imtis (Y_t, X_t) tenkina *klasikinį tiesinį gausinį (KTG) modelį*. Parametrų įvertinimai, gauti tiek kovariacijų metodu, tiek mažiausių kvadratų metodu yra:

$$\widehat{\beta}_2 = \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\overline{X^2} - (\overline{X})^2}, \tag{2.14}$$

$$\widehat{\beta}_1 = \overline{Y} - \overline{X} \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\overline{X^2} - (\overline{X})^2}. \tag{2.15}$$

Įvertinimai yra atsitiktiniai dydžiai. Konkrečios jų reikšmės priklauso nuo konkrečios imties realizacijos. Svarbu nustatyti bendras tų įvertinimų savybes, nepriklausančias nuo konkrečios imties. Taigi nagrinėdami gautus įvertinimus kaip atsitiktinius dydžius, turime atsakyti į tokius klausimus:

- Kokios yra įvertinimų paprasčiausios charakteristikos: vidurkiai, dispersijos, kovariacijos ir pan.?

- Kaip palyginti kovariacijų metodu gautus įvertinimus su įvertinimais gaunamais kitais metodais?

Šiame skyrelyje ir bandysime pirmiausia atsakyti į šiuos klausimus.

2.4.1 Vidurkinės savybės

Parametro β_2 įvertinimą patogu išreikšti tiesine forma

$$\widehat{\beta}_2 = \beta_2 + \sum_{t=1}^{T} w_t \varepsilon_t, \tag{2.16}$$

kai

$$w_t = \frac{X_t - \overline{X}}{\sum_i (X_i - \overline{X})^2}, \quad t = 1, \ldots, T.$$

Norėdami išvesti šią formulę, pirmiausia įsitikiname, kad

$$\widehat{\beta}_2 = \frac{\sum (X_t - \overline{X})(Y_t - \overline{Y})}{\sum (X_t - \overline{X})^2}.$$

Tai visai nesunku patikrinti. Pastebėję, kad, sumuodami dydžių nuokrypius nuo aritmetinio vidurkio, visados gauname nulį, t.y.

$$\sum (X_t - \overline{X}) = 0,$$

išvedame

$$\begin{aligned} &\widehat{\beta}_2 = \frac{\sum (X_t - \overline{X}) Y_t - \overline{Y} \sum (X_t - \overline{X})}{\sum (X_t - \overline{X})^2} = \\ &\frac{\sum (X_t - \overline{X}) Y_t}{\sum (X_t - \overline{X})^2} = \sum w_t Y_t. \end{aligned}$$

Galiausiai vietoje Y_t įstatome $\beta_1 + \beta_2 X_t + \varepsilon_t$. Taigi

$$\begin{aligned} &\widehat{\beta}_2 = \sum w_t Y_t = \sum w_t (\beta_1 + \beta_2 X_t + \varepsilon_t) = \\ &\beta_1 \sum w_t + \beta_2 \sum w_t X_t + \sum w_t \varepsilon_t = \\ &\beta_2 + \sum w_t \varepsilon_t, \end{aligned}$$

nes

$$\sum w_t = 0, \tag{2.17}$$

o

$$\sum w_t X_t = \frac{\sum (X_t - \overline{X}) X_t}{\sum (X_t - \overline{X})^2} = \frac{\sum X_t^2 - \overline{X} \sum X_t}{\sum X_t^2 - \overline{X} \sum X_t} = 1. \tag{2.18}$$

Gauta $\widehat{\beta}_2$ išraiška labai naudinga. Pirmiausia tuo įsitikiname skaičiuodami vidurkį:

$$E\widehat{\beta}_2 = \beta_2 + \sum w_t E\varepsilon_t = \beta_2,$$

nes $E\varepsilon_t = 0$ su kiekvienu t. Vadinasi, mažiausių kvadratų metodas duoda nepaslinktą parametro β_2 įvertinimą. Tuo pasinaudoję suskaičiuojame ir įvertinimo $\widehat{\beta}_1$ vidurkį:

$$E\widehat{\beta}_1 = E(\overline{Y} - \widehat{\beta}_2\overline{X}) = \beta_1 + \beta_2\overline{X} - E\beta_2\overline{X} = \beta_1.$$

Taigi, ir įvertinimas $\widehat{\beta}_1$ yra nepaslinktas. Čia reikia atkreipti dėmesį, kad skaičiuodami įvertinimų vidurkius iš esmės panaudojome klasikinio tiesinio regresinio modelio prielaidas. Jei jos neteisingos, įvertinimas gali ir nebūti nepaslinktas. Pavyzdžiui, taip bus, jei $E\varepsilon_t \neq 0$. Prisiminus, kad ε_t aprašo ir praleistus faktorius, jo vidurkis bus nenulinis, jei pražiopsojome kažką svarbaus. Taigi tuo atveju ir įvertinimas bus paslinktas.

Nepaslinktumas dar nereiškia, kad įvertinimas yra geras. Tai tik viena medalio pusė. Kita pusė – variacijos ir kovariacija. Atsitiktinio dydžio variacija aprašo vidutinį kvadratinį nuokrypį nuo vidurkio. Tai yra ir atitinkamo tikimybinio skirstinio išsibarstymo matas. Jei teisingos tiesinio regresinio modelio prielaidos, tai

$$var(\widehat{\beta}_1) = \sigma^2 \frac{\sum X_i^2}{T\sum(X_i - \overline{X})^2}, \tag{2.19}$$

$$var(\widehat{\beta}_2) = \frac{\sigma^2}{\sum(X_i - \overline{X})^2}, \tag{2.20}$$

$$cov(\widehat{\beta}_1, \widehat{\beta}_2) = \frac{-\sigma^2\overline{X}}{\sum(X_i - \overline{X})^2}. \tag{2.21}$$

Pirmiausia išveskime šias formules. Skaičiuodami $\widehat{\beta}_2$ dispersiją, pasinaudojame (2.16) išraiška. Taigi

$$\begin{aligned} var(\widehat{\beta}_2) &= var(\beta_2 + \textstyle\sum w_t\varepsilon_t) = var(\sum w_t\varepsilon_t) \\ &\qquad [\text{nes} \ \ \beta_2 \ \ \text{yra konstanta}] \\ &= \textstyle\sum w_t^2 var(\varepsilon_t) + 2\sum\sum_{t\neq s} w_t w_s cov(\varepsilon_t, \varepsilon_s) \\ &= \textstyle\sum w_t^2 var(\varepsilon_t) \\ &\qquad [\text{nes} \ \ cov(\varepsilon_t, \varepsilon_s) = 0] \\ &= \sigma^2 \textstyle\sum w_t^2 = \frac{\sigma^2}{\sum(X_t - \overline{X})^2}. \end{aligned}$$

Dabar aptarkime šias formules.

1. Kaip matyti iš (2.19) – (2.21) išraiškų, visur įeina parametras σ^2 – paklaidų dispersija. Tai visai natūralu. Kuo didesnė dispersija, tuo didesnės ir įverčių variacijos apie vidurkį. O tai reiškia didesnį neapibrėžtumą, slypintį duomenyse. Informacija, kurią slepia duomenys apie parametrus β_1 ir β_2, yra skurdesnė, jei dispersija σ^2 didesnė.

2. Kiekvienoje (2.19), (2.20), (2.21) išraiškoje taip pat yra dydis $\sum(X_t - \overline{X})^2$ – egzogeninio kintamojo reikšmių nuokrypių nuo vidurkinės reikšmės kvadratų suma. Kuo ji didesnė, tuo mažesni tie dydžiai. Vadinasi, kuo didesnis egzogeninio dydžio reikšmių išsibarstymas apie vidurkį, tuo tikslesnė informacija apie parametrus. Intuityviai tai yra suvokiama. Jei $X_t = c$ su visais t, tai dispersijos tampa begalinės. Tai paaiškina sąlygą.

3. Kuo didesnis imties tūris T, tuo mažesnės dispersijos. Reiškia, geriau turėti daugiau duomenų.

4. Kuo didesnė suma $\sum X_t^2$ tuo didesnė $\widehat{\beta}_1$ dispersija. Kodėl taip yra? Parametras β_1 aprašo endogeninio kintamojo reikšmes, kai egzogeninis kintamasis $X = 0$. Vadinasi, kuo toliau nuo tos reikšmės yra X_t tuo sunkiau interpretuoti parametrą β_1 ir dar sunkiau jį tiksliai įvertinti.

5. Kovariacijos išraiškoje yra dydis $\overline{X}$. Be to, kovariacijos ženklas yra priešingas vidurkio ženklui. Tai nesunku paaiškinti. Jei $\overline{X} \geq 0$, tai, didėjant tiesės $\widehat{\beta}_1 + \widehat{\beta}_2 x$ laisvajam nariui $\widehat{\beta}_1$, mažėja polinkio keficientas $\widehat{\beta}_2$.

Priminsime, kad dispersijos σ^2 įvertinimas yra

$$\widehat{\sigma}^2 = \frac{1}{T-2} \sum_{t=1}^{T} \widehat{\varepsilon}_t^2.$$

Šis dispersijos įvertinimas yra nepaslinktas:

$$E\widehat{\sigma}^2 = \sigma^2.$$

Norėdami tai įrodyti pasinaudosime regresijos liekanų apibrėžimu ir parametrų įvertinimais:

$$\begin{aligned}
&\sum E\widehat{\varepsilon}_t^2 = \sum E(Y_t - \widehat{Y}_t)^2 = \\
&\sum E[(\beta_1 - \widehat{\beta}_1) + (\beta_2 - \widehat{\beta}_2)X_t + \varepsilon_t]^2 = \\
&\sum E(\beta_1 - \widehat{\beta}_1)^2 + E(\beta_2 - \widehat{\beta}_2)^2 \sum X_t^2 + \sum E\varepsilon_t^2 + \\
&2\sum E(\beta_1 - \widehat{\beta}_1)\varepsilon_t + 2\sum E(\beta_2 - \widehat{\beta}_2)X_t\varepsilon_t + \\
&2\sum E(\beta_1 - \widehat{\beta}_1)(\beta_2 - \widehat{\beta}_2)X_t = \\
&Tvar(\beta_1) + var(\beta_2)\sum X_t^2 + T\sigma^2 + 2cov(\widehat{\beta}_1, \widehat{\beta}_2)\overline{X} + \\
&2\sum E(\beta_1 - \widehat{\beta}_1)\varepsilon_t + 2\sum E(\beta_2 - \widehat{\beta}_2)X_t\varepsilon_t.
\end{aligned} \tag{2.22}$$

Kadangi $\widehat{\beta}_2 = \beta_2 + \sum w_t\varepsilon_t$, tai atsižvelgę dar į (2.18) lygybę, gauname

$$\sum E(\beta_2 - \widehat{\beta}_2)X_t\varepsilon_t = -\sum_t X_t \sum_s w_s E\varepsilon_t\varepsilon_s = -\sigma^2 \sum_t X_t w_t = -\sigma^2.$$

Liko suskaičiuoti

$$\begin{aligned}
&\sum E(\beta_1 - \widehat{\beta}_1)\varepsilon_t = \sum E(\overline{Y} - \beta_2\overline{X} - \sum_s \varepsilon_s - \overline{Y} + \widehat{\beta}_2\overline{X})\varepsilon_t = \\
&\sum E(\widehat{\beta}_2 - \beta_2)\overline{X}\varepsilon_t - \sum E\varepsilon_t \sum_s \varepsilon_s = -\sigma^2 T.
\end{aligned}$$

Į (2.22) surinkę ką suskaičiavome, gauname

$$\begin{aligned}
&\sum E\widehat{\varepsilon}_t^2 = \frac{\sigma^2 \sum X_t^2}{\sum_t (X_t - \overline{X})^2} + \frac{\overline{X}\sigma^2}{\sum_t (X_t - \overline{X})^2} + T\sigma^2 - \\
&\frac{2\sigma^2(\overline{X})^2}{\sum_t (X_t - \overline{X})^2} - 2\sigma^2 - \sigma^2 = (T-2)\sigma^2.
\end{aligned}$$

PIM modelyje

$$\widehat{\sigma}^2 = \frac{54311.3315}{38} = 1429.2456$$

Turėdami dispersijos įvertinimą, galime gauti ir parametrų įvertinimų dispersijų ir kovariacijos įvertinimus:

$$\widehat{var}(\widehat{\beta}_1) = \widehat{\sigma}^2 \frac{\sum X_i^2}{T\sum (X_i - \overline{X})^2} \tag{2.23}$$

$$\widehat{var}(\widehat{\beta}_2) = \frac{\widehat{\sigma}^2}{\sum (X_i - \overline{X})^2} \tag{2.24}$$

$$\widehat{cov}(\widehat{\beta}_1, \widehat{\beta}_2) = \frac{-\widehat{\sigma}^2\overline{X}}{\sum (X_i - \overline{X})^2}. \tag{2.25}$$

Dažniau naudosime dydžius

$$se(\beta_1) = \sqrt{\widehat{var}(\beta_1)} \quad \text{ir} \quad se(\beta_2) = \sqrt{\widehat{var}(\beta_2)}. \tag{2.26}$$

Gautus rezultatus pritaikykime PIM modeliui. Tam tikslui reikia suskaičiuoti regresijos liekanas. Atlikę aritmetinius veiksmus gauname:

$$\begin{aligned}
\widehat{var}(\beta_1) &= 1429.2456\frac{2102063}{40 \cdot 1532463} = 490.1200;\\
se(\beta_1) &= \sqrt{490.1200} = 22.1387;\\
\widehat{var}(\beta_2) &= \frac{1429.2456}{1532463} = 0.0009326;\\
se(\beta_2) &= \sqrt{0.0009326} = 0.0305;\\
\widehat{cov}(\beta_1, \beta_2) &= 1429.2456\frac{-698}{1532463} = -0.6510.
\end{aligned}$$

2.4.2 Gauso-Markovo teorema

Kaip matėme, parametrų β_1 ir β_2 įvertinimai yra tiesinė $Y_1, \ldots, Y_T$ kombinacija: $\widehat{\beta}_2 = \sum_{t=1}^{T} w_t Y_t$. Tokie įvertinimai vadinami tiesiniais. Be to, tie įvertinimai yra nepaslinkti: $E\widehat{\beta}_1 = \beta_1, E\widehat{\beta}_2 = \beta_2$. Galime spręsti tokį optimizavimo uždavinį: tarp visų tiesinių nepaslinktų parametrų β_1 ir β_2 įvertinimų rasti tuos, kurių dispersijos būtų mažiausios.

Bet pasirodo, tokius įvertinimus jau radome. Tą sako Gauso-Markovo teorema.

1 teorema. (Gauso-Markovo.) *Klasikinio tiesinio regresinio modelio, aprašomo (A1) – (A5) prielaidomis, parametrų β_1 ir β_2 mažiausių kvadratų įvertinimai turi mažiausią dispersiją tarp visų tiesinių nepaslinktų įvertinimų.*

Teorema neteigia, kad mažiausių kvadratų įvertinimai yra geriausi iš visų galimų.

Lygindami įvertinimus, netgi nebūtinai tiesinius ar nepaslinktus, dažniausiai naudojame vidutinį kvadratinį nuokrypį. Jei θ – koks nors parametras, o $\widehat{\theta}$ – jo įvertinimas, tai labai svarbi charakteristika yra vidutinis kvadratinis nuokrypis:

$$MISE = E(\widehat{\theta} - \theta)^2.$$

Kad galiotų Gauso-Markovo teorema, turi būti išpildytos (A1)–(A5) sąlygos. Jei bent viena iš jų yra neteisinga, mažiausių kvadratų įvertinimai nebus geriausi tiesiniai nepaslinkti įvertinimai.

Gauso-Markovo teorema nepriklauso nuo Gausiškumo prielaidos.

Įrodymas. Teoremą įrodysime tik įvertinimui $\widehat{\beta}_2$. Tarkime

$$\beta_2^* = \sum_{t=1}^{T} a_t Y_t$$

yra koks nors tiesinis parametro β_2 įvertinimas. Čia a_t yra bet kokie skaičiai. Patogumo dėlei tarkime, kad $a_t = w_t + c_t$ su kiekvienu t. Taigi

$$\begin{aligned}\beta_2^* &= \sum (w_t + c_t) Y_t = \sum (w_t + c_t)(\beta_1 + \beta_2 X_t + \varepsilon_t) \\ &= \beta_1 \sum (w_t + c_t) + \beta_2 \sum (w_t + c_t) X_t + \sum (w_t + c_t)\varepsilon_t \\ &= \beta_1 \sum c_t + \beta_2 + \beta_2 \sum c_t X_t + \sum (w_t + c_t)\varepsilon_t,\end{aligned}$$

nes, $\sum w_t = 0$ o $\sum w_t X_t = 1$ (žr. (2.17) ir (2.18) formules). Suskaičiuojame vidurkį

$$\begin{aligned}E\beta_2^* &= \beta_1 \sum c_t + \beta_2 + \beta_2 \sum c_t X_t + \sum (w_t + c_t) E\varepsilon_t \\ &= \beta_1 \sum c_t + \beta_2 + \beta_2 \sum c_t X_t.\end{aligned}$$

Tam, kad įvertinimas β_2^* būtų nepaslinktas, būtinai turi būti teisingos šios sąlygos

$$\sum c_t = 0 \quad \text{ir} \quad \sum c_t X_t = 0. \tag{2.27}$$

Pritaikius šias sąlygas, supaprastėja ir įvertinimo β_2^* išraiška:

$$\beta_2^* = \beta_2 + \sum (w_t + c_t)\varepsilon_t.$$

Ji palengvina dispersijos skaičiavimą. Turime

$$\begin{aligned}var(\beta_2^*) &= \sum (w_t + c_t)^2 var(\varepsilon_t) = \sigma^2 \sum (w_t + c_t)^2 \\ &= \sigma^2 \sum w_t^2 + 2\sigma^2 \sum w_t c_t + \sigma^2 \sum c_t^2 = \sigma^2 \sum w_t^2 + \sigma^2 \sum c_t^2 \\ &= var(\widehat{\beta}_2) + \sigma^2 \sum c_t^2 \geq var(\widehat{\beta}_2).\end{aligned}$$

Čia pritaikėme lygybę $\sum w_t c_t = 0$, kuri, savo ruožtu, yra (2.27) išvada:

$$\begin{aligned}&\sum w_t c_t = \sum \left[\frac{c_t(X_t - \overline{X})}{\sum (X_t - \overline{X})^2}\right] = \\ &\frac{1}{\sum (X_t - \overline{X})^2} \sum c_t X_t - \frac{\overline{X}}{\sum (X_t - \overline{X})^2} \sum c_t = 0.\end{aligned}$$

Iš įrodymo taip pat matyti, kad $var(\beta_2^*) = var(\widehat{\beta}_2)$ vieninteliu atveju – kai $c_t = 0$ su visais t. Bet tuo atveju $\beta_2^* = \widehat{\beta}_2$.

Panašiai teoremą įrodome ir parametro β_1 mažiausių kvadratų įvertinimui.

■

2.4.3 Įvertinimų tikimybiniai skirstiniai

Iki šiol, tirdami tiesinio regresinio modelio parametrų įvertinimus, nesirėmėme Gausiškumo hipoteze. Dabar tarkime, kad modelio paklaidos turi normalinį skirstinį su nuliniu vidurkiu ir dispersija σ^2. Tai yra, nagrinėkime tiesinį regresinį Gausinį modelį. Tuomet teisingas toks teiginys.

2.4.1 teiginys. *Klasikinio tiesinio gausinio modelio atveju parametrų β_1 ir β_2 mažiausių kvadratų įvertinimų skirstiniai yra:*

$$\widehat{\beta}_2 \overset{\mathcal{D}}{\sim} N\Big(\beta_2, \frac{\sigma^2}{\sum(X_k - \overline{X})^2}\Big), \tag{2.28}$$

$$\widehat{\beta}_1 \overset{\mathcal{D}}{\sim} N\Big(\beta_1, \frac{\sigma^2 \sum X_k^2}{T\sum(X_k - \overline{X})^2}\Big). \tag{2.29}$$

Įrodymui reikia pastebėti, kad nepriklausomų (arba nekoreliuotų) normalinių atsitiktinių dydžių tiesinė kombinacija yra normalinis dydis, t.y., jei atsitiktiniai dydžiai $\xi_1, \ldots, \xi_n$ yra nepriklausomi ir ξ_i turi normalinį skirstinį su vidurkiu a_i ir dispersija $\sigma_i^2, i = 1, \ldots, n$, tai atsitiktinis dydis $\sum_{k=1}^n \xi_k$ turi taip pat normalinį skirstinį su vidurkiu $a = \sum a_i$ ir dispersija $\sigma^2 = \sum \sigma_i^2$. Kadangi $\widehat{\beta}_2 = \sum_{i=1}^T w_i Y_i$ ir

$$w_i Y_i \overset{\mathcal{D}}{\sim} N(w_i(\beta_1 + \beta_2 X_i), w_i^2\sigma^2), \quad \text{kai} \quad i = 1, \ldots, T,$$

tai

$$\widehat{\beta}_2 = \sum_{i=1}^T w_i Y_i \overset{\mathcal{D}}{\sim} N\Big(\sum w_i(\beta_1 + \beta_2 X_i), \sigma^2 \sum w_i^2\Big).$$

Kadangi $\sum w_i = 0$ o $\sum_i w_i^2 = 1/\sum(X_i - \overline{X})^2$, tai (2.28) yra teisinga. Analogiškai įrodome ir (2.29).

Pasinaudoję tuom, kad $\sigma^{-1}(\xi - a) \overset{\mathcal{D}}{\sim} N(0,1)$, kai $\xi \overset{\mathcal{D}}{\sim} N(a,\sigma^2)$, iš (2.28) ir (2.29) gauname, kad

$$\Big(\sum(X_i - \overline{X})^2\Big)^{1/2}\frac{\widehat{\beta}_2 - \beta_2}{\sigma} \quad \overset{\mathcal{D}}{\sim} \quad N(0,1); \tag{2.30}$$

$$\Big(T\sum(X_i - \overline{X})^2\Big)^{1/2}\Big(\sum X_t^2\Big)^{-1/2}\frac{\widehat{\beta}_1 - \beta_1}{\sigma} \quad \overset{\mathcal{D}}{\sim} \quad N(0,1). \tag{2.31}$$

Šios formulės mažai naudingos jei dispersija σ^2 nežinoma. Tada reikia naudoti jos įverinimą. Tai keičia ir gaunamos statistikos skirstinį.

Dispersijos įvertinimo skirstinį aprašo šis teiginys.

2.4.2 teiginys. *KTG modelio dispersijos σ^2 mažiausių kvadratų įvertinimui teisinga:*

$$\frac{(T-2)\widehat{\sigma}^2}{\sigma^2} \quad \overset{\mathcal{D}}{\sim} \quad \chi^2_{T-2}.$$

Šį teiginį įrodysime vėliau.

2.5 Statistikinis tyrimas

Ekonometrinio modelio statistikinį tyrimą sudaro

- pasikliautinių intervalų parametrams nustatymas;
- hipotezių apie parametrų reikšmes tyrimas.

Su šiais uždaviniais susipažinsime nagrinėdami paprasčiausią regresinį modelį ir pajamų-išlaidų maistui pavyzdį.

2.5.1 Pasikliautiniai intervalai

2.5.1 teiginys. *Klasikinio tiesinio gausinio modelio atveju*

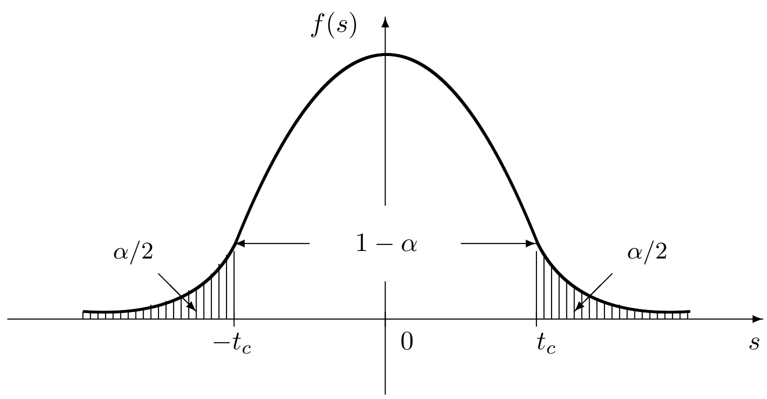
$$\tau_i = \frac{\widehat{\beta}_i - \beta_i}{se(\widehat{\beta}_i)} \quad \overset{\mathcal{D}}{\sim} \quad t_{(T-2)}, \quad i = 1,2.$$

Dydis τ_k labai reikšmingas kalbant tiek apie intervalines prognozes, tiek ir tikrinant hipotezes. Stjudento atsitiktinio dydžio $t_{(T-2)}$ kritinė reikšmė t_c, atitinkanti reikšmingumo lygmenį α (kuris dažniausiai parenkamas 0.01 arba 0.05, ar koks kitas mažas skaičius) randama iš lygties

$$P(t_{(T-2)} > t_c) = P(t_{(T-2)} \leq -t_c) = \alpha/2.$$

Kitaip tariant,

$$P(-t_c \leq t_{(T-2)} \leq t_c) = 1 - \alpha.$$



Grafikas 2.5: Stjudento skirstinio kvantiliai

Žinodami, kad atsitiktinis dydis τ_2 turi Stjudento skirstinį su $T-2$ laisvės laipsniais, matome, kad

$$P(-t_c \leq \frac{\widehat{\beta}_2 - \beta_2}{se(\widehat{\beta}_2)} \leq t_c) = 1 - \alpha$$

arba

$$P(\widehat{\beta}_2 - t_c se(\widehat{\beta}_2) \leq \beta_2 \leq \widehat{\beta}_2 + t_c se(\widehat{\beta}_2)) = 1 - \alpha. \tag{2.32}$$

Atsitiktinis intervalas $[\widehat{\beta}_2 - t_c se(\widehat{\beta}_2),\ \ \widehat{\beta}_2 + t_c se(\widehat{\beta}_2)]$ vadinamas *intervaliniu parametro β_2 įvertinimu*. Gautas (2.32) sąryšis reiškia, kad su tikimybe $1-\alpha$ intervalas $[\widehat{\beta}_2 - t_c se(\widehat{\beta}_2),\ \ \widehat{\beta}_2 + t_c se(\widehat{\beta}_2)]$ turi tikrąją parametro β_2 reikšmę. Kai dydžiai $\widehat{\beta}_2$ ir $se(\widehat{\beta}_2)$ yra konkrečios reikšmės, tai gautas skaitinis intervalas vadinamas $(1-\alpha)\times 100\%$ *intervaliniu β_2 įverčiu* arba $(1-\alpha)\times 100\%$ *simetriniu pasikliautinumo intervalu*.

Su intervaliniu įvertinimu ir intervaliniu įverčiu reikia būti atsargiems. Bet kuris intervalinis įvertis gali turėti, bet gali ir neturėti tikrosios parametro reikšmės, nes ta reikšmė apskritai yra nežinoma.

Išlaidų maistui pavyzdyje $T = 40$. Parinkę $\alpha = 0.05$, iš lentelių randame kritinę reikšmę $t_c = 2.02$. Taigi

$$P(\widehat{\beta}_2 - 2.024se(\widehat{\beta}_2) \leq \beta_2 \leq \widehat{\beta}_2 + 2.024se(\widehat{\beta}_2)) = 0.95.$$

Norėdami gauti intervalinį įvertį, suskaičiuojame $\widehat{\beta}_2 = 0.1283$ ir $se(\widehat{\beta}_2) = 0.0305$. Įstatę šias reikšmes, gauname 95-nuošimčių pasikliautinį intervalą $[0.0666, 0.1900]$. Ar β_2 priklauso šiam intervalui? Šito nežinome ir niekad nežinosime. Tad kokia gi nauda iš intervalinio įverčio? Galima tikrai pasakyti, kad naudodami tą pačią procedūra daug kartų, vidutiniškai 95 nuošimčius kartų, tikroji parametro reikšmė priklausys intervalui. Jokiu būdu negalima tvirtinti, kad su tikimybe 0.95 tikroji β_2 reikšmė priklausys intervalui $[0.0666, 0.1900]$. Mat intervalas yra deterministinis, o tvirtinimas būtų stochastinis.

2.5.2 Hipotezių apie parametrų reikšmes tikrinimas

Hipotezės patikrinimas palygina prielaidą apie tiriamą reiškinį su informacija, kuri yra imtyje. Tiksliau pasakius, prielaidos, kurios tikrinamos, dažniausiai susijusios su ekonominio modelio parametrais. Turint ekonominį ar statistikinį modelį, hipotezės formuojamos apie ekonominį elgesį. Tos hipotezės susiveda dažniausiai į prielaidas apie modelio parametrus. Kiekvieną hipotezės patiktrinimą sudaro keturi ingredientai:

1. Nulinė hipotezė H_0;
2. Alternatyvi hipotezė H_1;
3. Testinė statistika;
4. Atmetimo arba kritinė sritis.

Nulinė hipotezė. Žymima H_0. Paprastai aprašo kurią nors parametro reikšmę. Tarkime, $H_0 : \beta_2 = 0$. Jei ši hipotezė teisinga, tai, pavyzdžiui, PIM modelyje pajamos neturi jokios įtakos išlaidoms maistui.

Alternatyvi hipotezė. Paprastai kartu su nuline hipoteze yra formuluojama ir alternatyva jai, tai yra kita hipotezė. Žymima H_1. Pavyzdžiui, nulinei hipotezei $H_0 : \beta_2 = 0$ galimos trys alternatyvos:

- $H_1 : \beta_2 \neq 0$;
- $H_1 : \beta_2 > 0$. Atmetę nulinę hipotezę alternatyvos naudai, darome išvadą, kad parametras β_2 yra teigiamas. Tai eliminuoja bet kokią neigiamų reikšmių galymybę. Tokios alternatyvos labai dažnai pasitaiko ekonometrijoje, nes labai dažnai ekonomikos teorija leidžia nustatyti parametro ženklą. Pavyzdžiui, PIM modelyje alternatyvi hipotezė kaip tik ir būtų $H_1 : \beta_2 > 0$, nes ekonomikos teorija aiškiai teigia, jog išlaidos maistui didėja, didėjant pajamoms;
- $H_1 : \beta_2 < 0$.

Testinė statistika. Informacija, kurią imtis suteikia apie tiriamą nulinę hypotezę, turi atsispindėti testinėje statistikoje (imties funkcijoje). Remdamiesi testinės statistikos reikšmėmis, nusprendžiame, ar atmesti nulinę hipotezę, ar ne. Vienas svarbiausių reikalavimų testinei statistikai yra tas, kad jos skirstinys turi būti žinomas (bent apytikriai), kai teisinga nulinė hipotezė, ir tas skirstinys turi būti kitoks, kai teisinga alternatyva.

Nagrinėkime PIM modeliui nulinę hipotezę $H_0 : \beta_2 = 0$ ir alternatyvą $H_1 : \beta_2 \neq 0$. Pastebėsime, kad ekonomiškai labiau pagrįsta yra alternatyva $\beta_2 > 0$. Klasikinio tiesinio gausinio modelio atveju, statistika

$$\tau = \frac{\widehat{\beta}_2 - \beta_2}{se(\widehat{\beta}_2)} \sim t_{T-2}.$$

Jei teisinga nulinė hipotezė, t.y., $\beta_2 = 0$, tai

$$\tau = \frac{\widehat{\beta}_2}{se(\widehat{\beta}_2)} \sim t_{T-2}.$$

Tai ir būtų viena iš galimų testo statistikų, mat, jei nulinė hypotezė nėra teisinga, tai jos skirstinys jau nebūtų stjudento su $T-2$ laisvės laipsniais.

Kritinė arba *atmetimo sritis.* Tai tokia testinės statistikos reikšmių sritis, kai atmetama nulinė hipotezė alternatyvos naudai. Tokią sritį galime sukonstruoti tik tada, kai žinome testinės statistikos skirstinį su sąlyga, kad teisinga nulinė hipotezė. Praktiškai kritinę sritį sudaro tokios statistikos reikšmės, kurios įgyjamos su maža tikimybe, kai teisinga nulinė hipotezė. Jei, panaudojus turimą imtį, testinės statistikos suskaičiuota reikšmė pateko į mažos tikimybės sritį, tai mažai tikėtina, kad statistikos skirstinys yra būtent tas.

Kad būtų aiškiau, vėl panagrinėkime PIM modelį ir testo statistiką

$$t = \frac{\widehat{\beta}_2}{se(\widehat{\beta}_2)} \sim t_{T-2}.$$

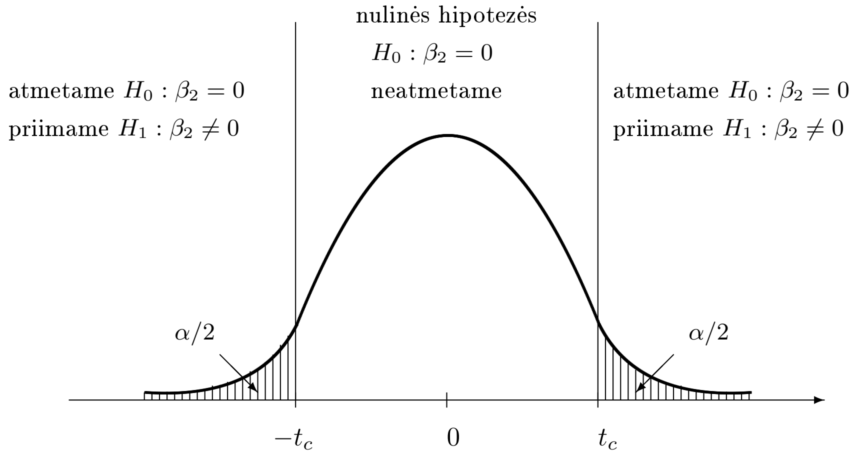
Kritinę sritį randame parinkę reikšmingumo lygmenį α, kuris paprastai imamas lygus $0.01, 0.05$ arba 0.10. Iš lygčių

$$P(t \geq t_c) = P(t \leq -t_c) = \alpha/2$$

randame kritinę reikšmę t_c. Kitaip tariant, atmetimo sritis yra stjudento skirstinio uodegos. Jei konkreti statistikos reikšmė, suskaičiuota duotai imčiai, patenka į kritinę sritį, tai rodo, kad ta reikšmė nėra būdinga statistikai, kai teisinga nulinė hipotezė.

Jei testinės statistikos reikšmė patenka į kritinę sritį, nulinę hipotezę atmetame alternatyvos naudai.

Jei testinės statistikos reikšmė nepatenka į kritinę sritį, tai sakome, kad duomenys neprieštarauja nulinei hipotezei.



Grafikas 2.6: Kritinė sritis nulinei hipotezei $\beta_2 = 0$ su alternatyva $\beta_2 \neq 0$.

Taigi tikrindami statistikinę hipotezę atliekame šiuos žingsnius:

1. Apibrėžiame nulinę hipotezę ir alternatyvas.

2. Nustatome testinę statistiką ir jos skirstinį, kai teisinga nulinė hipotezė.

3. Parenkame reikšmingumo lygmenį α ir surandame kritinę sritį.

4. Suskaičiuojame testinės statistikos reikšmę turimai imčiai.

5. Padarome išvadas.

Pritaikykime šiuos žingsnius išlaidų maistui modelyje.

1. Nulinė hipotezė yra $H_0 : \beta_2 = 0.$, alternatyva – $H_1 : \beta_2 \neq 0$. Jei teisinga nulinė hipotezė, klasikinis tiesinis gausinis modelis nerodo jokio ekonominio sąryšio tarp pajamų ir išlaidų maistui per savaitę. Jei teisinga alternatyva, tai toks sąryšis įmanomas.

2. Testinė statistika
$$t = \frac{\widehat{\beta}_2}{se(\widehat{\beta}_2)} \sim t_{(T-2)},$$
kai teisinga nulinė hipotezė.

3. Parinkime $\alpha = 0.05$. Kritinis taškas tada yra $t_c = 2.024$ ir jis atitinka stjudento skirstinio su 38 laisvės laipsniais 0.025-kvantilį.

4. Iš turimų duomenų surandame $t = 4.21$.

5. Išvados. Kadangi $t = 4.20 > t_c = 2.024$, tai nulinę hypotezę atmetame alternatyvos naudai.

Toks testas dar vadinamas parametro β_2 reikšmingumo testu (nustatome ar parametras yra reikšmingas, ar ne).

I ir II rūšių klaidos. Ar hipotezė atmetama, ar ne, visada yra galimybė suklysti. Tai neišvengiama. Bet kurioje hipotezių tikrinimo situacijoje yra du būdai padaryti teisingą sprendimą ir du būdai suklysti. Teisingi sprendimai bus šiais atvejais:

- Nulinė hipotezė klaidinga, ir sprendimas yra ją atmesti.
- Nulinė hipotezė teisinga, ir sprendimas yra jos neatmesti.

Priimtas sprendimas bus klaidingas, jei

- Nulinė hipotezė teisinga, bet sprendimas priimtas ją atmesti (I rūšies klaida).
- Nulinė hipotezė klaidinga, bet nuspręsta jos neatmesti (II rūšies klaida).

Atmesdami nulinę hipotezę rizikuojame padaryti pirmosios rūšies klaidą. Pirmos rūšies klaidos tikimybė yra α – reikšmingumo lygmuo. Rizikuojame padaryti antros rūšies klaidą neatmesdami nulinės hypotezės kai ji klaidinga. Antriosios rūšies klaidos tikimybė yra nekontroliuojama ir jos negalima suskaičiuoti. Žinomi tokie faktai:

- Antrosios rūšies klaidos tikimybė priklauso nuo pasirinkto reikšmingumo lygmens α, t.y. nuo pirmos rūšies klaidos tikimybės. Kai ji didėja, mažėja antrosios rūšies klaidos tikimybė.

- Kuo hipotetinė parametro reikšmė yra arčiau tikrosios reikšmės, tuo didesnė antrosios rūšies klaidos tikimybė. Jei nulinė hipotezė yra $\beta_2 = 0$ ir tikroji to parametro reikšmė labai artima nuliui, tai antros rūšies klaidos tikimybė bus labai didelė. Galima sakyti, kad testas neturi pakankamai galios atskirti tikrosios parametro reikšmės nuo (klaidingos) hipotetinės, jei tos reikšmės artimos.

- Kuo didesnis imties tūris T, tuo mažesnė antros rūšies klaidos tikimybė, jei fiksuotas reikšmingumo lygmuo.

- Beveik jokiai hipotezei, kurią tenka tikrinti ekonomistams, neegzistuoja paties geriausio testo, kurį būtų galima taikyti visiems atvejams. Geriausias čia reiškia, kad to testo pirmosios rūšies klaidos tikimybė yra fiksuota, o antrosios – mažiausia.

Testo p-reikšmės. Aprašant statistikinio testo rezultatus, dažnai pateikiamos taip vadinamos testo p-reikšmės. Jei testo statistika yra τ, o $\widehat{\tau}$, tos statistikos reikšmė, atitinkanti imtį, tai dvipusė p-reikšmė yra $P(|t| \geq |\widehat{t}|)$.

Lygindami reikšmingumo lygmenį α su p-reikšme, galime nustatyti ar atmesti nulinę hipotezę. Taisyklė yra paprasta.

> *Jei p-reikšmė yra mažesnė už pasirinktą reikšmingumo lygmenį α, tai nulinę hipotezę atmetame alternatyvos naudai.*

PIM modeliui $p = 0.000155$. Tai yra mažiausia tikimybė, su kuria vis dar nepriimtume nulinės hipotezės.

Bendresnės hipotezės.

Bendresnė nulinė hipotezė gali būti tokia:

$$H_0 : \quad \beta_2 = c,$$

o alternatyva

$$H_1 : \quad \beta_2 \neq c.$$

Čia c iš anksto duotas skaičius. Tokios hipotezės patikrinimui KTG modelio atveju galime naudoti statistiką

$$\tau = \frac{\widehat{\beta}_2 - c}{se(\widehat{\beta}_2)}.$$

Jei teisinga nulinė hipotezė, tuomet ta statistika turi stjudento skirstinį su $T-2$ laisvės laipsniais. Nulinę hipotezę atmetame, jei $|t| \geq t_c$ arba jei tos statistikos p-reikšmė yra mažesnė nei pasirinktas reikšmingumo lygmuo.

PIM modeliui patikrinkime hipotezę, kad $\beta_2 = 0.1$.

1. Nulinė hipotezė $H_0 : \beta_2 = 0.1$, alternaryva – $H_1 : \beta_2 \neq 0.1$.

2. Testinė statistika
$$\tau = \frac{\widehat{\beta}_2 - 0.1}{se(\widehat{\beta}_2)} \overset{\mathcal{D}}{\sim} t_{(38)},$$
jei teisinga nulinė hipotezė.

3. Parinkime $\alpha = 0.05$. Atitinkama stjudento skirstinio kritinė reikšmė yra $t_c = 2.024$.

4. Naudodami turimus duomenis suskaičiuojame $\widehat{\beta}_2 = 0.1283, \quad se(\widehat{\beta}_2) = 0.0305$. Statistikos reikšmė yra $t = 0.9263$.

5. Kadangi $t = 0.9263 < t_c = 2.024$, tai turimi duomenys nulinei hipotezei neprieštarauja. Be to, p-reikšmė $p = 0.3601 > \alpha = 0.05$. Tai reiškia, kad reikšmingumo lygmenį galima dar mažinti.

Tarp hipotezių tikrinimo ir pasikliautinųjų intervalų yra labai natūralus sąryšis.

2.6 Prognozavimas regresiniu modeliu

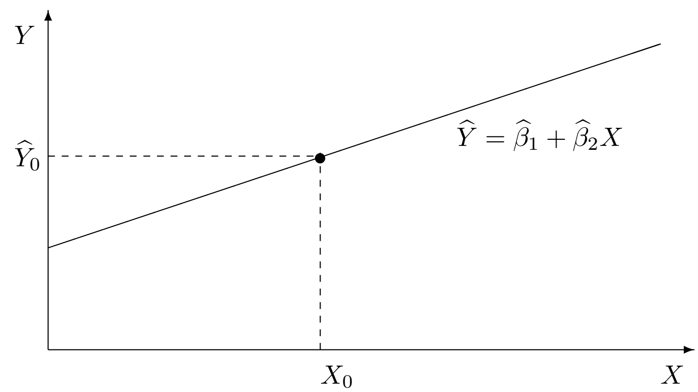
2.6.1 Taškinė prognozė

Vienas iš svarbiausių motyvų nagrinėti ekonometrinį modelį yra galimybė prognozuoti priklausomo kintamojo Y reikšmes. Turėdami tiesinį regresinį modelį, aprašomą (A1)–(A5) prielaidomis, kintamojo Y reikšmė, atitinkanti paaiškinamojo kintamojo reikšmę $X = X_0$, yra

$$Y_0 = \beta_1 + \beta_2 X_0 + \varepsilon_0;$$

čia ε_0 yra atsitiktinis dydis, kurio vidurkis yra nulis, dispersija lygi σ^2 ir kuris nekoreliuoja su $\varepsilon_1, \ldots, \varepsilon_T$. Šioje lygybėje pakeitę nežinomus parametrus jų įvertinimais, o nežinomą paklaidą – vidutine reikšme (nuliu), gauname Y_0 prognozę

$$\widehat{Y}_0 = \widehat{\beta}_1 + \widehat{\beta}_2 X_0.$$



Grafikas 2.7: Taškinė prognozė

Gauta reikšmė dar vadinama mažiausių kvadratų taškine prognoze, siekiant pabrėžti parametrų įvertinimo pasirinktą metodą. Norint ištirti prognozės empirines savybes, paprastai nagrinėjama prognozės paklaida

$$f = \widehat{Y}_0 - Y_0 = (\widehat{\beta}_1 - \beta_1) + (\widehat{\beta}_2 - \beta_2)X_0 - \varepsilon_0.$$

Taikydami mažiausių kvadratų įvertinimų savybes, nesunkiai gauname, kad prognozės paklaidos vidurkis yra nulis:

$$Ef = E(\widehat{Y}_0 - Y_0) = 0.$$

Tai reiškia, kad prognozavimo vidutinė paklaida yra nulinė arba kad $\widehat{Y}_0$ yra nepaslinktas Y_0 įvertinimas. Be to, galime suskaičiuoti ir prognozės dispersiją:

$$var(f) = \sigma^2 \Big[1 + \frac{1}{T} + \frac{(X_0 - \overline{X})^2}{\sum (X_t - \overline{X})^2} \Big].$$

Iš variacijos išraiškos matome, kad, kuo labiau X_0 nutolęs nuo vidurkio $\overline{X}$, tuo didesnė paklaidos variacija. Jei modelis yra Gausinis, tai ir prognozės paklaida turi normalinį skirstinį.

Kadangi modelio paklaidų dispersija yra nežinoma, tai nežinoma ir prognozės variacija. Jos įvertinimą gauname, pakeitę nežinomą dispersiją σ^2 įvertinimu $\widehat{\sigma}^2$:

$$\widehat{var}(f) = \widehat{\sigma}^2 \Big[1 + \frac{1}{T} + \frac{(X_0 - \overline{X})^2}{\sum (X_t - \overline{X})^2} \Big].$$

Kvadratinė šaknis iš gautos reikšmės vadinama standartine prognozės paklaida:

$$se(f) = \sqrt{\widehat{var}(f)}.$$

Išlaidų maistui modelyje prognozuodami vidutines išlaidas maistui, atitinkančias 750 Lt mėnesines pajamas, gausime

$$\widehat{Y}_0 = 40.7676 + 0.1283(750) = 136.98.$$

Tos prognozės variacijos įvertis yra

$$\widehat{var}(f) = 1467.4986,$$

o standartinė prognozės paklaida

$$se(f) = \sqrt{1467.4986} = 38.3079.$$

Šitas dydis bus panaudotas vėliau, kai kalbėsime apie intervalinę prognozę.

2.6.2 Intervalinis prognozavimas

Taškinės prognozės $\widehat{Y}_0 = \widehat{\beta}_1 + \widehat{\beta}_2 X_0$ paklaida yra atsitiktinis dydis

$$f = \widehat{Y}_0 - Y_0 = (\widehat{\beta}_1 - \beta_1) + (\widehat{\beta}_2 - \beta_2) X_0 - \varepsilon_0.$$

Jei modelis buvo KTG, tai f turi normalinį skirstinį su nuliniu vidurkiu ir dispersija

$$var(f) = \sigma^2 \Big(1 + \frac{1}{T} + \frac{(X_0 - \overline{X})^2}{\sum (X_t - \overline{X})^2} \Big).$$

Vadinasi,

$$\frac{f}{\sqrt{var(f)}} \quad \sim \quad N(0,1).$$

Kadangi paklaidos f variacijos išraiškoje yra σ^2, kuris dažniausiai nėra žinomas, jį reikia pakeisti įverčiu. Taip gauname ir dispersijos įvertį

$$\widehat{var}(f) = \widehat{\sigma}^2\Big(1 + \frac{1}{T} + \frac{(X_0 - \overline{X})^2}{\sum(X_t - \overline{X})^2}\Big).$$

Tada

$$\frac{f}{\sqrt{\widehat{var}(f)}} = \frac{f}{se(f)} \quad \sim \quad t_{(T-2)}. \tag{2.33}$$

Remdamiesi šiuo rezultatu, galime sukonstruoti dydžio Y_0 intervalinę prognozę. Jei t_c yra stjudento skirstinio $t_{(T-2)}$ kritinė reikšmė, atitinkanti lygmenį α, t.y.

$$P(t_{(T-2)} > t_c) = \alpha/2,$$

tai

$$P(-t_c \leq t_{(T-2)} \leq t_c) = 1 - \alpha.$$

Įstatę vietoje $t_{(T-2)}$ atsitiktinį dydį iš (2.33) formulės, turime

$$P(-t_c \leq \frac{\widehat{Y}_0 - Y_0}{se(f)} \leq t_c) = 1 - \alpha.$$

Iš šios formulės gauname

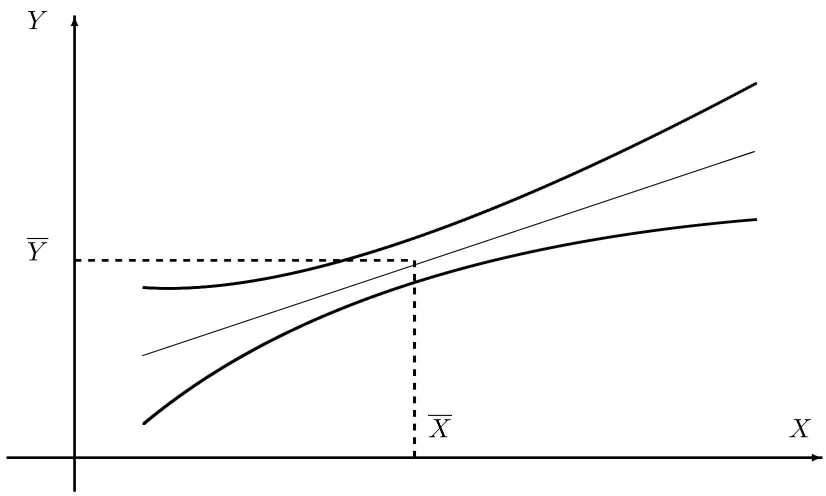
$$P(\widehat{Y}_0 - t_c se(f) \leq Y_0 \leq \widehat{Y}_0 + t_c se(f)) = 1 - \alpha.$$

Intervalas $[\widehat{Y}_0 - t_c se(f),\ \widehat{Y}_0 + t_c se(f)]$ vadinamas Y_0 $(1-\alpha)100\%$-ine intervaline prognoze. Iš to intervalo pavidalo aiškiai matyti, kad, kuo Y_0 reikšmė labiau nutolusi nuo vidurkio $\overline{Y}$, tuo didesnis prognozės intervalas (žr. 2.8 grafiką).

PIM modelyje 95%-inė intervalinė prognozė dydžiui Y_0, kai $X_0 = 750$ yra

$$[59.44,\ 214.52].$$

Tai reiškia, kad namų ūkis, kurio savaitės pajamos yra 750 Lt su 95 nuošimčių patikimumu maistui išleis tarp 59.44 Lt ir 214.52 Lt. Toks didelis intervalas reiškia, kad taškinė prognozė 136.98 Lt maisto išlaidoms yra nereali. Tai galėtų reikšti, kad yra praleisti kiti faktoriai, įtakojantys išlaidas maistui, ir tie faktoriai sudaro didelę paklaidą. Apie praplėstą modelį bus kalbama vėliau.



Grafikas 2.8: intervalinis prognozavimas

2.7 Ekonometrinis tyrimas

Kol kas susipažinsime tik su determinacijos koeficientu ir jo panaudojimu modelio atitikimo tyrimui. Taip pat aptarsime funkcionalinės modelio formos parinkimą.

2.7.1 Determinacijos koeficientas

Vienas iš svarbiausių ekonometrinės analizės tikslų – paaiškinti endogeninio dydžio variaciją. Sudarydami modelį siekiame ir šio tikslo. PIM pavyzdyje namų ūkio išlaidos maistui kinta pagal modelį

$$Y_t = \beta_1 + \beta_2 X_t + \varepsilon_t.$$

Čia atskirta pagrindinė priežastis – pajamos, o visi kiti galimi faktoriai atspindimi dydžiu ε_t. Tardami, kad $E\varepsilon_t = 0$, kartu tariame, kad vidutinės išlaidos maistui, kai pajamos fiksuotos, yra

$$E(Y_t|X_t) = \beta_1 + \beta_2 X_t.$$

Šis sąlyginis vidurkis aprašo sisteminę kintamojo dydžio Y_t komponentę $\beta_1+\beta_2 X_t$, o ε_t yra atsitiktinė to dydžio komponentė. Nė viena iš sudedamųjų dalių nėra stebima, tačiau, pasitelkę turimus duomenis, abi jas galime įvertinti. Būtent, įvertinę parametrus β_1 ir β_2 gauname

$$\widehat{Y}_t = \widehat{\beta}_1 + \widehat{\beta}_2 X_t,$$

ir

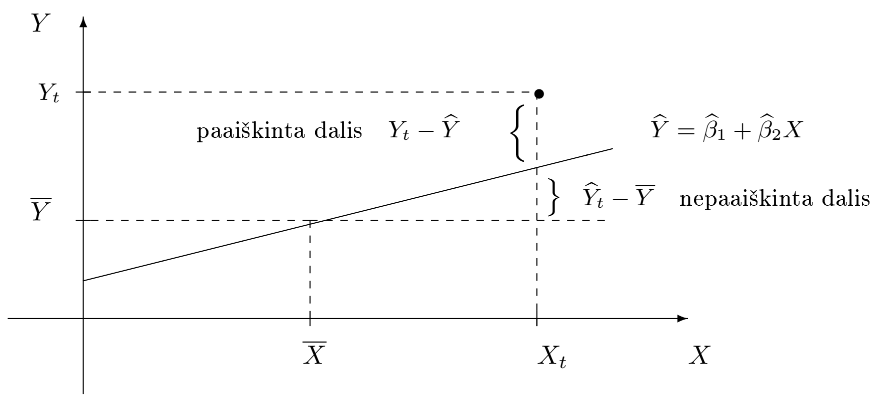
$$\widehat{\varepsilon}_t = Y_t - \widehat{Y}_t.$$

Dydžiai $\widehat{\varepsilon}_1, \ldots, \widehat{\varepsilon}_T$ vadinami modelio liekanomis. Taigi

$$Y_t = \widehat{Y}_t + \widehat{\varepsilon}_t.$$

Iš abiejų pusių atėmę empirinį vidurkį $\overline{Y}$, gauname

$$Y_t - \overline{Y} = (\widehat{Y}_t - \overline{Y}) + \varepsilon_t.$$



Grafikas 2.9: Paaiškinta ir nepaaiškinta Y_t komponentės

Nesunku matyti, kad skirtumas tarp Y_t ir vidurkio $\overline{Y}$ susideda iš tos dalies, kurią paaiškina regresinis modelis, t.y. $\widehat{Y}_t - \overline{Y}$ ir dalies, kuri lieka nepaaiškinta (žr. 2.9 grafiką). Pilnoji variacija yra dydis

$$TSS := \sum_{t=1}^{T} (Y_t - \overline{Y})^2,$$

čia $\overline{Y} = T^{-1}\sum_{t=1}^{T} Y_t$. Pilnoji variacija aprašo dydžio Y stebimų reikšmių nuokrypių nuo vidutinės reikšmės didumą. $\widehat{Y}$ pilnoji variacija yra

$$RSS := \sum_{t=1}^{T}(\widehat{Y}_t - \overline{Y})^2.$$

Jeigu modelis buvo parinktas teisingai ir įvertinimo metodas geras, tai turėtų būti $TSS \approx RSS$ arba $RSS/TSS \approx 1$. Santykis RSS/TSS aprašo tą egzogeninio kintamojo reikšmių pilnosios variacijos dalį, kurią paaiškina ekonometrinis modelis. Šis santykis vadinamas determinacijos koeficientu arba charakteristika R^2. Taigi

$$R^2 = \frac{RSS}{TSS} = \frac{\sum_{t=1}^{T}(\widehat{Y}_t - \overline{Y})^2}{\sum_{t=1}^{T}(Y_t - \overline{Y})^2}. \tag{2.34}$$

Visi ekonometriniai paketai būtinai pateikia R^2 reikšmę. Artimas vienetui R^2 yra modelio privalumas. Bet spekuliuoti tokiu metodu modelio kokybei nustatyti nederėtų.

Nagrinėkime parametrų mažiausių kvadratų įvertinimus. Tada

$$\begin{aligned}
&\sum(Y_t - \overline{Y})^2 = \sum[(\widehat{Y}_t - \overline{Y}) + \widehat{\varepsilon}_t]^2 = \\
&\sum(\widehat{Y}_t - \overline{Y})^2 + \sum\widehat{\varepsilon}_t^2 + 2\sum(\widehat{Y}_t - \overline{Y})\widehat{\varepsilon}_t = \\
&\sum(\widehat{Y}_t - \overline{Y})^2 + \sum\widehat{\varepsilon}_t^2.
\end{aligned} \tag{2.35}$$

(2.35) formulė aprašo pilnosios variacijos skaidinį į paaiškintąją ir nepaaiškintąją komponentes. Regresijos liekanų kvadratų sumą pažymėję ESS, gauname

$$TSS = RSS + ESS.$$

Šis išdėstymas lydi kiekvieną regresinę analizę. Jis dažniausiai aprašomas vadinamojoje ANOVA lentelėje (analysis of variance):

Variacijos šaltinis	DF	Kvadratų suma	Vidurkis
Paaiškintoji variacija	1	RSS	$RSS/1$
Nepaaiškintoji variacija	$T-2$	ESS	$ESS/(T-2)$
Pilnoji variacija	$T-1$	TSS	

Šioje lentelėje laisvės laipsniai (DF) yra:

- RSS turi vieną laisvės laipsnį (paaiškinamųjų kintamųjų skaičius);
- ESS turi $T-2$ laisvės laipsnis (stebėjimų skaičius minus išreikštinių parametrų skaičius);
- TSS turi $T-1$ laisvės laipsnį (stebėjimų skaičius minus parametrų skaičiui modelyje (šiuo atveju parametras yra tik vienas β_1)).

Determinacijos koeficientas R^2 yra aprašomoji charakteristika. Jis neaprašo regresinio modelio kokybės. Reikia įsidėmėti, kad regresinė analizė nekelia tikslo surasti modelio su didžiausiu galimu R^2. Susikoncentruoti vien šio parametro maksimizavimui – bloga strategija.

Kadangi mažiausių kvadratų metodas reiškia, kad minimizuojamas dydis $\sum_{t=1}^{t}\widehat{\varepsilon}_t^2$, tai gauname tokią išvadą: *(??) modeliui determinacijos koeficientas yra maksimalus mažiausių kvadratų metodui.*

Priminsime, kad dviejų dydžių X ir Y koreliacijos koeficientas yra

$$\rho = \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}}.$$

Turėdami atsitiktinio vektoriaus (X,Y) stebėjimus (X_i,Y_i), $i=1,\dots,T$, koreliacijos koeficientą įvertinime taip:

$$\widehat{\rho} = \frac{\widehat{cov}(X,Y)}{\sqrt{\widehat{var}(X)\widehat{var}(Y)}},$$

kai

$$\widehat{cov}(X,Y) = \frac{1}{T-1}\sum(X_i-\overline{X})(Y_i-\overline{Y}), \widehat{var}(X) = \frac{1}{T-1}\sum(X_i-\overline{X})^2.$$

Tiesinio regresinio modelio atveju matome, kad $\rho^2 = R^2$. Tai yra, determinacijos koeficientas yra lygus koreliacijos koeficiento įvertinimo kvadratui. Koreliacijos koeficientas parodo tiesinės priklausomybės tarp dydžių X ir Y stiprumą. Ta interpretacija nedaug skiriasi nuo determinacijos koeficiento interpretacijos.

Paprasčiausio tiesinio regresinio modelio atveju R^2 gali būti suskaičiuotas kaip regresijos tarp Y_i ir $\widehat{\beta}_1+\widehat{\beta}_2X_i$ koreliacijos koeficiento kvadratas. Dažnai dėl šio atitikimo determinacijos koeficientas naudojamas kaip modelio atitikimo matas ("goodness of fit test").

2.7.2 Rezultatų pateikimas

Praktikoje visi regresinės analizės skaičiavimai atliekami kompiuteriu. Tam naudojama įvairiausia programinė įranga. Kiekviena iš jų pateikia rezultatus savaip. Viena tai daro aiškiau, kita – ne taip aiškiai. Aptarsime tipinį rezultatų pateikimą.

Dauguma ekonometrijai skirtų paketų pateikia koeficientų įverčius, jų standartines paklaidas ir atitinkamas t -reikšmes bei p-reikšmes. Tai dažniausiai užrašoma į lentelę. Pavyzdžiui,

(1) Kintamasis	(2) Koeficientas	(3) Standartinė pakl.	(4) t−reikšmė	(5) p−reikšmė
Laisvas narys	40.7676	22.1387	1.841	0.0734
X	0.1283	0.0305	4.201	0.0002

Reikia suprasti, kad kompiuteris modelio parametro vardų nežino. Laisvąjį narį daugelis paketų vadina INTERCEPT, o regresantą – X. Standartiniai nuokrypiai – tai $se(\beta_1)$ ir $se(\beta_2)$.

t-statistikos reikšmė pateikiama abiems parametrams: kai teisinga nulinė hipotezė $H_0: \quad \beta_1 = 0$ prieš alternatyvą $H_1 : \beta_1 \neq 0$ ir atitinkamai, kai $H_0: \quad \beta_2 = 0$, prieš alternatyvą $H_1 : \beta_2 \neq 0$. Tos reikšmės yra

$$t_1 = \frac{\widehat{\beta}_1}{se(\beta_1)} = \frac{40.7676}{22.1387} = 1.84$$

ir

$$t_2 = \frac{\widehat{\beta}_2}{se(\beta_2)} = \frac{0.1283}{0.0305} = 4.20$$

Paskutiniame lentelės stulpelyje surašytos atitinkamos p-reikšmės, atitinkančios nulinę hipotezę $H_0 : \beta_k = 0$ prieš dvipusę alternatyvą $H_1 : \beta_k \neq 0, \quad k = 1, 2$. Taigi ir p-reikšmės yra dvipusės, jos randamos iš lentelių:

$$P(|t_{(38)}| > 1.841) = 0.0734, \quad P(|t_{(38)}| > 4.201) = 0.0002.$$

Vienpuses p-reikšmes suskaičiuojame gautą atitinkamą dvipusę reikšmę padaliję iš dviejų.

Kiekviena programa paprastai pateikia ir R^2 reikšmę. PIM modeliui $R^2 = 0.317$.

Dažniausiai lentelės su kompiuteriu gautais rezultatais pateikiamos darbo gale. Tekste, aprašant rezultatus, naudojamas šis užrašas:

$$\begin{array}{rlll} \widehat{Y}_t = & 40.7676+ & 0.1283x_t, & R^2 = 0.317 \\ & (22.1387) & (0.0305)(s.e.) & \end{array} \tag{2.36}$$

Dažnai pasitaiko ir kitas būdas: vietoj standartinių nuokrypių pateikti atitinkamas t-reikšmes:

$$\begin{array}{rllll} \widehat{Y}_t = & 40.7676+ & 0.1283x_t, & & R^2 = 0.317 \\ & (1.48) & (4.20) & (t) & \end{array} \tag{2.37}$$

Komentaras dėl R^2 reikšmės: mažos reikšmės paprastai gaunamos naudojant vietos aplinkybių duomenis. Laiko eilutėms tos reikšmės būna didesnės. Tačiau neverta stengtis modelį tempti prie didelio R^2. Jei R^2 mažas, reikia ieškoti to priežasčių ir bandyti jas paaiškinti.

2.7.3 Kitos funkcionalinės formos

Funkcionalinės formos parinkimas reiškia kintamųjų transformacijos parinkimą. Bendra dviejų kintamųjų tiesinio modelio forma yra

$$f(Y) = \beta_1 + \beta_2 g(X) + \varepsilon,$$

kai f ir g – duotos transformacijos. Tiesinio modelio esmė yra ta, kad modelis yra tiesinis išreikštinių parametrų atžvilgiu.

Taip gaunamos įvairios tiesinio modelio funkcionalinės formos. Dažniausiai pasitaikančios transformacijos ir atitinkami parametrai surašyti šioje lentelėje.

Transformacijos	Modelis	Nuolydis	Elastingumas
Tiesinė	$Y = \beta_1 + \beta_2 X + \varepsilon$	β_2	$\beta_2 \frac{X}{Y}$
Tiesinė-atvirkštinė	$Y = \beta_1 + \beta_2 \frac{1}{X} + \varepsilon$	$-\beta_2 \frac{1}{X}$	$-\beta_2 \frac{1}{XY}$
Log-Log	$\log(Y) = \beta_1 + \beta_2 \log(X) + \varepsilon$	$\beta_2 \frac{Y}{X}$	β_2
Log-tiesinė	$\log(Y) = \beta_1 + \beta_2 X + \varepsilon$	$\beta_2 Y$	$\beta_2 X$
Tiesinė-log	$Y = \beta_1 + \beta_2 \log(X) + \varepsilon$	$\beta_2 \frac{1}{X}$	$\beta_2 \frac{1}{Y}$
Log-atvirkštinė	$\log(Y) = \beta_1 - \beta_2 \frac{1}{X}$	$\beta_2 \frac{Y}{X^2}$	$\beta_2 \frac{1}{X}$

Labai dažnai tiesinio modelio funkcionalinė forma parenkama pagal turimus duomenis.

Paklausos ir pasiūlos modeliams dažniausiai naudojama tiesinė forma. Alternatyva – log-log forma.

Būna ir kitų transformacijų. Pavyzdžiui, tipinis produkto X_t gamybos kainos Y_t modelis yra

$$Y_t = \beta_1 + \beta_2 X_t^2 + \varepsilon_t.$$

2.8 Kiti ekonominiai uždaviniai

Regresinį modelį naudojome tirdami išlaidų maistui priklausomybę nuo pajamų. Paprasčiausias modelis gali būti pritaikytas ir kitiems, analogiškiems ekonominiams uždaviniams. Keletą tokių uždavinių paminėsime.

- *Paklausos sąryšiai*: kurios nors prekės (prekių grupės) paklausos priklausomybė nuo kainos (kainų indekso);
- *Gamybos funkcijos tyrimas*: siekiama nustatyti pagamintos produkcijos kiekio ir panaudotos darbo jėgos sąryšį;
- *Agreguoto suvartojimo tyrimas*: pajamų ir agreguoto (sumarinio) suvartojimo sąryšiai;

- *Pasiūlos tyrimas*: pasiūlos priklausomybė nuo kainos;
- *Agreguotų investicijų tyrimas*: investicijų ir palūkanų normos sąryšiai;
- *Aplinkosauginiai tyrimai*: užterštumo ir parduodamo benzino sąryšiai; oro taršos priklausomybė nuo geografinės padėties; oro taršos priklausomybė nuo laiko;
- *Filipso kreivės tyrimas*: sąryšis tarp infliacijos ir užimtumo; A.W.Philips 1958 metais pasiūlė modelį: sąryšį tarp darbo užmokesčio paikeitimo ir nedarbo lygio. Jei w_t – darbo užmokestis laiko momentu t, tai procentinis jo pasikeitimas yra
$$\%\Delta w_t = \frac{w_t - w_{t-1}}{w_{t-1}}.$$
Jei tarsime, kad $\%\Delta w_t$ yra proporcingas darbo jėgos paklausai d_t, tai turėsime
$$\%\Delta w_t = \gamma d_t.$$
Čia γ yra ekonominis parametras. Kadangi nedarbo lygis u_t yra atvirkščiai susijęs su darbo paklausa, tai
$$d_t = \alpha + \eta \frac{1}{u_t}.$$
Sujungę gautas dvi lygtis, išvedame modelį
$$\%\Delta w_t \gamma\alpha + \gamma\eta \frac{1}{u_t}.$$
Modelis yra netiesinis tiek parametrų, tiek kintamųjų atžvilgiu. Tačiau, pažymėję $Y_t = \%\Delta w_t$, $X_t = 1/u_t$, $\beta_1 = \eta\alpha$, $\beta_2 = \gamma\eta$, gauname paprasčiausią tiesinį regresinį modelį.

3 skyrius. Bendrasis tiesinis regresinis modelis

3.1 Ekonominis uždavinys ir ekonometrinis modelis

Kaskart greito maisto restorano mėsainių skyriui reikia nuspręsti kiek kitą savaitę investuoti reklamai ir kokią nustatyti mėsainių kainą. Vadybininkus ypač domina pelnas (bendrosios pajamos), gaunamas iš investicijų į reklamą.

- Ar, padidinus investicijas į reklamą, pelnas (bendrosios pajamos) padidėja?
- Jei taip, ar pelnas yra pakankamas, kad pasiteisintų išaugusios išlaidos reklamai?

Be to, vadybininkams ne mažiau svarbi kainų strategija.

- Ar kainų sumažinimas padidintų pajamas?
- Ar tas padidėjimas esminis?

Jei kainos sumažinimas reikštų tik nedidelį pardavimo padidėjimą, tai pelnas gali ir sumažėti (paklausa neelastiška kainoms). O toks kainos sumažinimas, kuris ryškiai padidintų pardavimą, aiškiai padidintų ir pajamas (paklausa elastiška kainai). Visi tie klausimai svarbūs efektyviai vadybai. Uždavinį sutrumpintai vadinsime PRK – pelnas, reklama, kaina.

Nustatykime kintamuosius ir ekonominį modelį. Paaiškinamasis kintamasis – pažymėkime jį tr – pajamos per savaitę. Tirsime jo priklausomybę nuo kainos p ir reklamos išlaidų a tai savaitei. Kintamieji tr ir a matuojami tūkstančiais litų, o p – litais. Ekonomikos teorija pirmiausia rekomenduotų tiesinę priklausomybę:

$$tr = \beta_1 + \beta_2 p + \beta_3 a. \tag{3.1}$$

Pakomentuokime (3.1) modelį. Pirmiausia išsiaiškinkime galimas parametrų β_1, β_2 ir β_3 interpretacijas.

Parametras β_2 rodo pajamų tr pasikeitimą tūkstančiais litų, kai kaina pakinta vienu litu, o išlaidos reklamai nekinta:

$$\beta_2 = \frac{\partial tr}{\partial p}.$$

Parametras β_2 gali būti tiek teigiamas, tiek neigiamas. Jei jis yra teigiamas, tai kainos padidėjimas iššaukia pajamų padidėjimą, o tai reiškia, kad paklausa yra neelastiška kainai. Paklausos elastiškumas kainai reiškia, kad didėjant kainai, pajamos mažėja ($\beta_2 < 0$.)

Parametras β_3 aprašo pajamų reakciją į išlaidas reklamai. Kitaip tariant, β_3 lygus pajamų tr pasikeitimui (1000 Lt), kai išlaidos reklamai a padidėja vienu vienetu (1000 Lt), o p lieka pastovus:

$$\beta_3 = \frac{\partial\, tr}{\partial a}.$$

Tikėtina, kad $\beta_3 > 0$. Tai yra, tikėtina, kad padidinus išlaidas reklamai, bendrosios pajamos padidės. Jei $\beta_3 < 1$, tai išlaidas reklamai padidinus 1000 Lt, bendrosios pajamos padidės, bet mažiau nei 1000 Lt. Jei $\beta_3 > 1$ – tai daugiau nei 1000 Lt. Tokiu būdu parametras β_3 parodo ir išlaidų reklamai atsiperkamumą. Taigi siekiant nustatyti reklamos politiką, labai svarbu teisingai įvertinti parametrą β_3.

Laisvasis narys β_1 vaidina pagalbinį vaidmenį ir yra skirtas modelio subalansavimui, siekiant geresnės prognozės.

Į ekonominį sąryšį, aprašomą (3.1) formule, žiūrime kaip į vidutinių pajamų priklausomybę nuo kainos ir išlaidų reklamai, t.y. (3.1) lygtis išties yra

$$E(tr|p,a) = \beta_1 + \beta_2 p + \beta_2 a.$$

Bet kaina ir išlaidos reklamai PRK pavyzdyje yra determinuoti, t.y. iš anksto nustatomi dydžiai. Todėl $E(tr|p,a) = E(tr)$, ir lygtimi

$$E(tr) = \beta_1 + \beta_2 p + \beta_2 a$$

aprašome vidutinių pajamų priklausomybę nuo kainos ir išlaidų reklamai. Atitinkamas ekonometrinis modelis yra

$$tr = \beta_1 + \beta_2 p + \beta_3 a + \varepsilon. \tag{3.2}$$

Tiesinė dalis $\beta_1+\beta_2 p+\beta_3 a$ aprašo sisteminę pajamų tr priklausomybę nuo kainos p ir išlaidų reklamai a. Tuo tarpu atsitiktinė paklaida ε aprašo tiek praleistus faktorius, darančius įtaką pajamoms, tiek, kaip vėliau matysime, galimas kainų agregavimo paklaidas, praleistą informaciją ir pan. Surinktus duomenis $(tr_t, p_t, a_t, t = 1, \ldots, T)$ aprašome pagal pasirinktą modelį taip:

$$tr_t = \beta_1 + \beta_2 p_t + \beta_3 a_t + \varepsilon_t, \quad t = 1, \ldots, T. \tag{3.3}$$

Vadinasi, kiekvieną konkretų duomenį traktuojame kaip atsitiktinio dydžio, aprašomo (3.2) lygtimi, tam tikrą kopiją. Surinktuose duomenyse paslėptai informacijai paprastai būdingas neapibrėžtumas. Juos aprašome atsitiktiniais dydžiais $\varepsilon_t, t = 1, \ldots, T$, kurie dar vadinami *modelio inovacijomis* arba *modelio paklaidomis*.

Gautas (3.3) modelis yra atskiras bendrojo tiesinio regresinio modelio atvejis, atitinkantis du regresorius.

3.2 KTR MODELIS

Bendrasis tiesinis vienos lygties regresinis modelis, atitinkantis d regresorių, aprašomas lygtimi

$$Y = \beta_1 + \beta_2 X_2 + \cdots + \beta_d X_d + \varepsilon; \tag{3.4}$$

čia Y – endogeninis kintamasis (dar vadinamas regresantu arba paaiškinamuoju kintamuoju), $X_2, \ldots, X_d$ – egzogeniniai kintamieji (dažnai dar vadinami regresoriais arba paaiškinančiaisiais kintamaisiais); ε – atsitiktinis faktorius; $\beta_1, \beta_2, \ldots, \beta_d$ – išreikštiniai modelio parametrai. Atitinkamas duomenų

$$Y_t, X_{2t}, X_{3t}, \ldots, X_{dt}, \quad t = 1, \ldots, T$$

generavimo mechanizmas aprašomas modeliu

$$Y_t = \beta_1 + \beta_2 X_{2t} + \cdots + \beta_d X_{dt} + \varepsilon_t, \qquad t = 1, 2, \ldots, T. \tag{3.5}$$

Čia $\varepsilon_t, t = 1, 2, \ldots, T$ – atsitiktiniai faktoriai, dar vadinami modelio inovacijomis, triukšmu arba paklaidomis.

(3.2) modelis gaunamas iš bendrojo tiesinio regresinio (3.5) modelio, kai $d = 3$, $Y_t = tr_t$, $X_{2t} = p_t$, $X_{3t} = a_t$. Taigi (3.2) modelį galime perrašyti taip

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \varepsilon_t, \quad t = 1, \ldots, T. \tag{3.6}$$

Toliau (3.5) modelį patogiau užrašyti vektoriniu būdu

$$Y_t = \boldsymbol{X}_t^{\tau} \boldsymbol{\beta} + \varepsilon_t, \quad t = 1, \ldots, T, \tag{3.7}$$

kai $\boldsymbol{X}_t = (1, X_{2t}, \ldots, X_{dt})^{\tau}$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^{\tau}$ arba vektoriniu-matriciniu:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{3.8}$$

kai

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_T \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_T \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_d \end{pmatrix},$$

$$\boldsymbol{X} = \begin{pmatrix} 1 & X_{21} & \ldots & X_{d1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{2T} & \ldots & X_{dT} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_1^{\tau} \\ \vdots \\ \boldsymbol{X}_d^{\tau} \end{pmatrix}.$$

Matrica $\boldsymbol{X}$ vadinama *projektine* arba *plano*[1] matrica. Atskiru atveju, kai $d = 3$, ji yra

$$\boldsymbol{X} = \begin{pmatrix} 1 & X_{21} & X_{31} \\ \vdots & \vdots & \vdots \\ 1 & X_{2T} & X_{3T} \end{pmatrix}.$$

[1] angl. design matrix

Jei nepasakyta kitaip, laikysime, kad $T > d$. Be to, laikysime, kad atžvilgiu egzogeninių kintamųjų teisingos šios dvi prielaidos.

$T1$. *Egzogeninių kintamųjų pastovumo*: kintamieji $X_2, X_3, \ldots, X_d$ yra nestochastiniai dydžiai.

$T2$. *Kolinearumo*: tarp vektorių eilučių

$$(1, \ldots, 1), \quad (X_{21}, \ldots, X_{2T}), \ldots, (X_{d1}, \ldots, X_{dT})$$

nėra tiesinės priklausomybės. Kitaip sakant, matrica $\boldsymbol{X}$ yra pilno stulpelių rango.

Paklaidoms $(\varepsilon_t, t = 1, \ldots, T)$ dažniausiai taikysime šias sąlygas:

$E1$. *Nulinių vidurkių*: $E\varepsilon_1 = \cdots = E\varepsilon_T = 0$, t.y. $E\boldsymbol{\varepsilon} = 0$. Ši sąlyga reiškia, kad visų modeliu padarytų netikslumų, praleistų kintamųjų ir pan. vidutinė įtaka kintamajam Y yra nulinė.

$E2$. *Homoskedastiškumo*: $var(\varepsilon_t) = \sigma^2$ su visais $t = 1, \ldots, T$. Tai yra atsitiktinių dydžių $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_T$ skirstiniai turi vienodas dispersijas σ^2. Paprastai dispersija σ^2 yra nežinomas neišreikštinis modelio parametras ir aprašo neapibrėžtumų, slypinčių duomenyse, apimtį. Homoskedastinis modelis reiškia, kad kiekviename stebėjime slypinčiai informacijai būdingas vienodas neapibrėžtumas.

$E3$. *Nekoreliuotų paklaidų*:

$$cov(\varepsilon_t, \varepsilon_s) = E(\varepsilon_t - E\varepsilon_t)(\varepsilon_s - E\varepsilon_s) = 0, \quad \text{kai} \quad t \neq s.$$

Labai dažnai sutinkama *gausinių paklaidų sąlyga:*

E4. atsitiktiniai dydžiai $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ yra nepriklausomi ir turi vienodą normalinį pasiskirstimą su nuliniu vidurkiu ir dispersija σ^2.

Priminsime, kad atsitiktinio vektoriaus $\xi = (\xi_1, \ldots, \xi_d)^\tau$ kovariacinė matrica yra

$$\begin{gathered} cov(\xi) = E(\xi - E\xi)(\xi - E\xi)' = \\ E \begin{pmatrix} \xi_1 - E\xi_1 \\ \vdots \\ \xi_d - E\xi_d \end{pmatrix} (\xi_1 - E\xi_1, \ldots, \xi_d - E\xi_d) = \\ \begin{pmatrix} var(\xi_1) & cov(\xi_1, \xi_2) & \ldots & cov(\xi_1, \xi_d) \\ \vdots & \vdots & \ddots & \vdots \\ cov(\xi_d, \xi_1) & cov(\xi_d, \xi_2) & \ldots & var(\xi_d) \end{pmatrix}. \end{gathered}$$

Jei paklaidos $\varepsilon_1, \ldots, \varepsilon_T$ yra homoskedastinės ir nekoreliuotos, tai

$$cov(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}.$$

Čia ir toliau $\boldsymbol{I}$ žymi vienetinę matricą, kurios dimensija aiški iš konteksto. Tais atvejais, kai dimensiją bus būtina pabrėžti, prirašysime indeksą: $d \times d$ vienetinę matricą žymėsime $\boldsymbol{I}_d$.

Paminėtos sąlygos charakterizuoja klasikinį tiesinį modelį, t.y.

> (3.5) *modelis, kurio atžvilgiu teisingos* $T1, T2, E1, E2$ *ir* $E3$ *sąlygos, vadinamas klasikiniu tiesiniu regresiniu modeliu* (sutrumpintai KTR modeliu).
>
> *Klasikinį tiesinį regresinį modelį, kurio paklaidoms teisinga* $(E4)$ *sąlyga vadiname gausiniu klasikiniu tiesiniu regresiniu modeliu* (sutrumpintai GKTR modelis).

Beje, jei teisinga $(E4)$ sąlyga, tai teisingos ir $(E2, E3)$ sąlygos.

3.3 Parametrų įvertinimas

Kaip ir paprasčiausio regresinio modelio, taip ir bendrojo modelio parametrus galime įvertinti įvairiais metodais. Populiariausi iš jų: mažiausių kvadratų, koreliacijų, maksimalaus tikėtinumo, mažiausių absoliutinių nuokrypių, kvantilinis. Čia aptarsime tik mažiausių kvadratų metodą. Kiti bus aptariami vėlesniuose kursuose.

3.3.1 Mažiausių kvadratų metodas

Tiesinio (3.4) modelio parametrų $\boldsymbol{\beta}$ įvertinimas mažiausių kvadratų metodu remiasi paprastu pastebėjimu, kad

$$\underset{b \in R}{argmin}\, E(Y-b)^2 = EY.$$

Priminsime, kad *argmin* reiškia tą argumento reikšmę, kuriame funkcija įgyja savo mažiausią reikšmę. Tai yra, kvadratinio vidurkio prasme atsitiktinio dydžio vidurkis yra geriausia jo skaitmeninė aproksimacija . Nustatę tiesinę atsitiktinio dydžio Y vidurkio priklausomybę nuo kintamųjų $X_1, \dots, X_d$, $EY = \beta_1 + \beta_2 X_1 + \dots + \beta_d X_d$, parametrus $\beta_1, \dots, \beta_d$ galime įvertinti suskaičiavę $argmin_{(b_1, b_2, \dots, b_d) \in R^d} E(Y - b_1 - b_2 X_2 - \dots - b_d X_d)^2$. Kvadratinį nuokrypį įvertinę empiriniu, apibrėžiame

$$(\widehat{\beta}_1, \widehat{\beta}_2, \dots, \widehat{\beta}_d) = \underset{(b_1, b_2, \dots, b_d) \in R^d}{argmin} \sum_{t=1}^{T} (Y_t - b_1 - b_2 X_{2t} - \dots - b_d X_{dt})^2.$$

Tai ir yra parametrų $(\beta_1, \beta_2, \ldots, \beta_d)$ mažiausių kvadratų įvertinimas. Juos surasti daug paprasčiau panaudojus matricinę modelio išraišką. Aibrėžkime funkciją

$$f(\boldsymbol{b}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b})^\tau (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}) = \sum_{t=1}^{T} (Y_t - \boldsymbol{X}_t \boldsymbol{b})^2, \quad \boldsymbol{b} = (b_1, \ldots, b_d)^\tau \in R^d.$$

3.3.1 apibrėžimas. *Modelio, aprašomo* (3.8) *lygtimi, parametro* $\boldsymbol{\beta}$ *mažiausių kvadratų įvertinimas yra*

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{b} \in R^d}{argmin}\, f(\boldsymbol{b}).$$

Kitaip tariant $f(\widehat{\boldsymbol{\beta}}) = \min_{\boldsymbol{b} \in R^d} f(\boldsymbol{b})$. Funkcijos $f(\boldsymbol{b})$ minimumo tašką rasti nesunku. Kadangi

$$f(\boldsymbol{b}) = \boldsymbol{Y}^\tau \boldsymbol{Y} - 2\boldsymbol{b}^\tau \boldsymbol{X}^\tau \boldsymbol{Y} + \boldsymbol{b}^\tau \boldsymbol{X}^\tau \boldsymbol{X} \boldsymbol{b},$$

suskaičiavę šios funkcijos išvestinę taške $\widehat{\boldsymbol{\beta}}$ ir, prilyginę ją nuliui, gauname lygtį

$$-2\boldsymbol{X}^\tau \boldsymbol{Y} + 2\boldsymbol{X}^\tau \boldsymbol{X} \widehat{\boldsymbol{\beta}} = 0. \tag{3.9}$$

Pagal kolinearumo prielaidą, matricos $\boldsymbol{X}^\tau \boldsymbol{X}$ rangas lygus d. Vadinasi egzistuoja atvirkštinė matrica $(\boldsymbol{X}^\tau \boldsymbol{X})^{-1}$ ir (3.9) lygtį galime išspręsti. Jos sprendinys yra

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\tau \boldsymbol{X})^{-1} \boldsymbol{X}^\tau \boldsymbol{Y}. \tag{3.10}$$

Pažymėję

$$\boldsymbol{A} = (\boldsymbol{X}^\tau \boldsymbol{X})^{-1} \boldsymbol{X}^\tau, \tag{3.11}$$

$\boldsymbol{\beta}$ įvertinimą užrašome $\widehat{\boldsymbol{\beta}} = \boldsymbol{A}\boldsymbol{Y}$. Pritaikę (3.8) formulę, matome, kad

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\tau \boldsymbol{X})^{-1} \boldsymbol{X}^\tau \boldsymbol{Y} = (\boldsymbol{X}^\tau \boldsymbol{X})^{-1} \boldsymbol{X}^\tau (\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) =$$
$$(\boldsymbol{X}^\tau \boldsymbol{X})^{-1} \boldsymbol{X}^\tau \boldsymbol{X} \boldsymbol{\beta} + (\boldsymbol{X}^\tau \boldsymbol{X})^{-1} \boldsymbol{X}^\tau \boldsymbol{\varepsilon} = \boldsymbol{\beta} + \boldsymbol{A}\boldsymbol{\varepsilon}.$$

Taigi

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \boldsymbol{A}\boldsymbol{\varepsilon}. \tag{3.12}$$

Gautąja formule ne kartą remsimės.

Regresijos liekanos yra

$$\widehat{\varepsilon}_t = Y_t - \widehat{Y}_t = Y_t - (\widehat{\beta}_1 + \widehat{\beta}_2 X_{2t} + \cdots + \widehat{\beta}_d X_{dt},$$

$t = 1, \ldots T$. Liekanų vektorių $\widehat{\boldsymbol{\varepsilon}} = (\widehat{\varepsilon}_1, \ldots, \widehat{\varepsilon}_d)^\tau$ galime išreikšti regresijos paklaidų vektoriumi $\boldsymbol{\varepsilon}$:

$$\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{M}\boldsymbol{\varepsilon}, \quad \boldsymbol{M} = I - \boldsymbol{X}(\boldsymbol{X}^\tau \boldsymbol{X})^{-1} \boldsymbol{X}^\tau. \tag{3.13}$$

Klasikinio tiesinio regresinio modelio dispersiją įvertiname pasinaudoję regresijos liekanomis. Paprastai naudojamas šis dispersijos įvertinimas:

$$\widehat{\sigma}^2 = \frac{1}{T-d} \sum_{t=1}^{T} \widehat{\varepsilon}_t^2 = \frac{1}{T-d} \widehat{\boldsymbol{\varepsilon}}^\tau \widehat{\boldsymbol{\varepsilon}}. \tag{3.14}$$

3.3.2 PRK modelio įvertinimas

Dviejų regresorių (3.6) modeliui, pažymėję

$$y_t = Y_t - \overline{Y}, \quad \overline{Y} = \frac{1}{T}\sum_{t=1}^{T} Y_t,$$

$$x_{2t} = X_{2t} - \overline{X_2}, \quad x_{3t} = X_{3t} - \overline{X_3}, \quad \overline{X_i} = \frac{1}{T}\sum_{t=1}^{T} X_{it}, \ i = 2, 3,$$

turime šiuos parametrų $\beta_1, \beta_2, \beta_3$ įvertinimus:

$$\begin{aligned} \widehat{\beta}_2 &= \frac{\sum y_t x_{2t} \sum x_{3t}^2 - \sum y_t x_{3t} \sum x_{2t} x_{3t}}{\sum x_{2t}^2 \sum x_{3t}^2 - \left(\sum x_{2t} x_{3t}\right)^2} \\ \widehat{\beta}_3 &= \frac{\sum y_t x_{3t} \sum x_{2t}^2 - \sum y_t x_{2t} \sum x_{2t} x_{3t}}{\sum x_{2t}^2 \sum x_{3t}^2 - \left(\sum x_{2t} x_{3t}\right)^2} \\ \widehat{\beta}_1 &= \overline{Y} - \widehat{\beta}_2 \overline{X_2} + \widehat{\beta}_3 \overline{X_3}. \end{aligned} \tag{3.15}$$

Pasinaudoję gautomis formulėmis ir duomenimis (žr. žemiau esančią lentelę), randame parametrų įverčius nagrinėjamam PRK modeliui:

$$\begin{aligned} \widehat{\beta}_1 &= 104.79, \\ \widehat{\beta}_2 &= -6.642, \\ \widehat{\beta}_3 &= 2.984. \end{aligned} \tag{3.16}$$

Taigi įverintas PRK modelis yra

$$\widehat{tr}_t = 104.79 - 6.642 p_t + 2.984 a_t. \tag{3.17}$$

Sav.	Pajamos 1000 Lt	Kaina Lt	Reklama 1000Lt	Sav.	Pajamos 1000 Lt	Kaina Lt	Reklama 1000Lt
1	123.1	1.92	12.4	2	124.3	2.15	9.9
3	89.3	1.67	2.4	4	141.3	1.68	13.8
5	112.8	1.75	3.5	6	108.1	1.55	1.8
7	143.9	1.54	17.8	8	124.2	2.10	9.8
9	110.1	2.44	8.3	10	111.7	2.47	9.8
11	123.8	1.86	12.6	12	123.5	1.93	11.5
13	110.2	2.47	7.4	14	100.9	2.11	6.1
15	123.3	2.10	9.5	16	115.7	1.73	8.8
17	116.6	1.86	4.9	18	153.5	2.19	18.8
19	149.2	1.90	18.9	20	89.0	1.67	2.3

21	132.6	2.43	14.1	22	97.5	2.13	2.9
23	106.1	2.33	5.9	24	115.3	1.75	7.6
25	98.5	2.05	5.3	26	135.1	2.35	16.8
27	124.2	2.12	8.8	28	98.4	2.13	3.2
29	114.8	1.89	5.4	30	142.5	1.50	17.3
31	122.6	1.93	11.2	32	127.7	2.27	11.2
33	113.0	1.66	7.9	34	144.2	1.73	17.0
35	109.2	1.59	3.3	36	106.8	2.29	7.3
37	145.0	1.86	15.3	38	124.0	1.91	12.7
39	106.7	2.34	6.1	40	153.2	2.13	19.6
41	120.1	2.05	6.3	42	119.3	1.89	9.0
43	150.6	2.12	18.7	44	92.2	1.87	2.2
45	130.5	2.09	16.0	46	112.5	1.76	4.5
47	111.8	1.77	4.3	48	120.1	1.94	9.3
49	107.4	2.37	8.3	50	128.6	2.10	15.4
51	124.6	2.29	9.2	52	127.2	2.36	10.2

Vieno greito maisto restorano duomenys

Kokias galime padaryti pirmines išvadas?

1. Neigiamas koeficientas prie kainos p_t rodo, kad paklausa elastinga kainai ir kainos padidėjimas vienu litu bendras pajamas vidutiniškai sumažintų 6.64 Lt. Arba, sumažinę kainą vienu litu, padidintume pajamas vidutiniškai 6.64 Lt. Taigi kainos mažinimas, siūlant įvairias nuolaidas ar naujus produktus, padidintų pajamas.
2. Reklamos koeficientas yra teigiamas. Vadinasi, padidinus reklamos išlaidas 1000 Lt, pajamos padidėtų vidutiniškai 2984 Lt. Šia informacija jau galime pasinaudoti. Jei žinosime, kiek kainuoja, tarkime, naujos rūšies mėsainių pagaminimas, galėsime išsiaiškinti ar jų reklamai išleisti pinigai padidins pelną.
3. Nenulinis laisvasis narys rodo, kad, jei išlaidos reklamai ir kainos yra nulinės, tai gaunamos pajamos vidutiniškai sudarys 104.79 Lt. Akivaizdu, kad tai nėra korektiška. Šiame modelyje laisvasis narys vaidina modelio stabilizatoriaus vaidmenį ir pagerina prognozavimo tikslumą.

Įvertintu modeliu vadybininkai jau gali pasinaudoti. Tarkime, jei ateinančią savaitę mėsainių kaina (agreguota) nustatyta 2 Lt, o reklamai skiriama 10000 Lt, tai prognozuojamos pajamos būtų 121.34 Lt:

$$\widehat{tr} = 104.785 - 6.6419 \cdot 2 + 2.9843 \cdot 10 = 121.34.$$

3.3.1 pastaba. *Neigiamas koeficientas prie kainos reiškia, kad, mažinant kainą, pajamos didėja. Išeitų, jog naudingiausia būtų kainą padaryti nuline. Bet akivaizdu, kad taip nėra. Mat regresiniu modeliu aprašome sąryšį tarp ekonominių* *dydžių, kurių reikšmės yra panašios į imties reikšmes. Neprotinga ekstrapoliuoti modelį ekstreminėms reikšmėms.*

PRK modeliui dispersiją įvertiname pagal (3.14) formulę:

$$\widehat{\sigma}^2 = \frac{1805.168}{52-3} = 36.84.$$

3.4 Statistikinės įvertinimų savybės

3.4.1 Paprasčiausios savybės

3.4.1 teiginys. *Jei* $E\boldsymbol{\varepsilon} = 0$, *tai įvertinimas* $\widehat{\boldsymbol{\beta}}$ *yra nepaslinktas, t.y.* $E\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}$.

Įrodymas. Atsižvelgę į (3.12) lygtį, turime, kad

$$E\widehat{\boldsymbol{\beta}} = E(\boldsymbol{\beta} + \boldsymbol{A}\boldsymbol{\varepsilon}) = \boldsymbol{\beta} + \boldsymbol{A}E\boldsymbol{\varepsilon} = \boldsymbol{\beta}.$$

■

3.4.2 teiginys. *Jei modelis yra homoskedastinis, tai*

$$cov(\widehat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^\tau\boldsymbol{X})^{-1}. \tag{3.18}$$

Įrodymas. Kadangi pagal homoskedastiškumo prielaidą $E\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\tau = \sigma^2\boldsymbol{I}$, tai

$$cov(\widehat{\boldsymbol{\beta}}) = E(\widehat{\boldsymbol{\beta}} - E\widehat{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}} - E\widehat{\boldsymbol{\beta}})^\tau = E(\boldsymbol{A}\boldsymbol{\varepsilon})(\boldsymbol{A}\boldsymbol{\varepsilon})^\tau = E(\boldsymbol{A}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\tau\boldsymbol{A}^\tau) =$$
$$\boldsymbol{A}E\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\tau\boldsymbol{A}^\tau = \boldsymbol{A}\sigma^2\boldsymbol{I}\boldsymbol{A}^\tau = \sigma^2\boldsymbol{A}\boldsymbol{A}^\tau.$$

Be to,

$$\boldsymbol{A}\boldsymbol{A}^\tau = ((\boldsymbol{X}^\tau\boldsymbol{X})^{-1}\boldsymbol{X}^\tau)((\boldsymbol{X}^\tau\boldsymbol{X})^{-1}\boldsymbol{X}^\tau)^\tau = ((\boldsymbol{X}^\tau\boldsymbol{X})^{-1})^\tau = (\boldsymbol{X}^\tau\boldsymbol{X})^{-1}.$$

Pastaroji lygybė teisinga todėl, kad matrica $\boldsymbol{X}^\tau\boldsymbol{X}$ yra simetrinė. Taigi (3.18) formulė įrodyta. ■

Matricos $(\boldsymbol{X}^\tau\boldsymbol{X})^{-1}$ elementus pažymėję m_{ij} turime, kad

$$\sigma^2 m_{ij} = cov(\widehat{\beta}_i, \widehat{\beta}_j), \quad i \neq j$$

ir

$$\sigma^2 m_{ii} = var(\widehat{\beta}_i).$$

Kadangi σ^2, kaip jau minėta, dažniausiai nežinome, tai nežinome ir $cov(\widehat{\beta})$. Jos įvertinimą gauname vietoje σ^2 įstatę dispersijos įvertinimą $\widehat{\sigma}^2$. Tokiu būdu

$$\widehat{cov}(\widehat{\beta}) = \widehat{\sigma}^2(\boldsymbol{X}^\tau \boldsymbol{X})^{-1}.$$

Iš čia randame kiekvieno parametro β_i įvertinimo $\widehat{\beta}_i$ dispersijos įvertinimą

$$\widehat{var}(\widehat{\beta}_i) = \widehat{\sigma}^2 m_{ii}$$

bei kovariacijų $cov(\widehat{\beta}_i, \widehat{\beta}_j)$ įvertinimus

$$\widehat{cov}(\widehat{\beta}_i, \widehat{\beta}_j) = \widehat{\sigma}^2 m_{ij}, \quad i \neq j.$$

Dviejų regresorių atveju turime

$$var(\widehat{\beta}_2) = \frac{\sigma^2}{(1 - r_{23}^2)\sum x_{2t}^2}, \tag{3.19}$$

$$cov(\widehat{\beta}_2, \widehat{\beta}_3) = \frac{-r_{23}\sigma^2}{(1 - r_{23}^2)\big(\sum x_{2t}^2\big)^{1/2}\big(\sum x_{3t}^2\big)^{1/2}}. \tag{3.20}$$

Čia r_{23} yra koreliacijos koeficientas tarp stebėjimų (X_{2t}) ir (X_{3t}):

$$r_{23} = \frac{\sum x_{2t}x_{3t}}{\big(\sum x_{2t}^2\big)^{1/2}\big(\sum x_{3t}^2\big)^{1/2}}.$$

Kitų parametrų įvertinimų variacijos bei kitos kovariacijos turi panašias išraiškas. Labai svarbu suprasti parametrų įvertinimų variacijos ir kovariacijų struktūrą. Parametro β_2 atveju variacijos struktūra yra tokia:

1. Dispersija σ^2 įtakoja mažiausių kvadratų įvertinimų dispersiją. To ir reikėjo tikėtis, nes σ^2 atpindi neapibrėžtumą, slypintį modelio specifikavime. Jei tas neapibrėžtumas yra didelis, vadinasi duomenys gali būti labiau išsibarstę apie vidutinę reikšmę $EY_t = \beta_1 + \beta_2 X_2 + \beta_3 X_3$. Tokiu atveju informacijos apie tikrąsias parametrų $\beta_1, \beta_2, \beta_2$ reikšmes bus mažiau ir duomenyse. Ir atvirkščiai, jeigu dispersija σ^2 maža, tai duomenys bus labiau koncentruoti apie vidurkį ir juose bus daugiau informacijos apie parametrus.

2. Imties tūris T daro įtaką per vardiklį $\sum_{t=1}^T x_{2t}^2$. Kadangi tai yra neneigiamų narių suma, tai, kuo didesnis T, tuo didesnė suma ir mažesnė įvertinimo variacija. Taigi didesnė imtis duoda tikslesnį įvertinimą.

3. Kvadratų suma $\sum_{t=1}^T x_{2t}^2$ aprašo paaiškinančiojo kintamojo (regresoriaus) variaciją apie vidurkį $\overline{X_2}$. Taigi norint tiksliau įvertinti parametrą β_2, reikia, kad X_2 stebėjimų variacija būtų kuo didesnė. Intuityviai tai yra aišku, nes β_2 aprašo regresoriaus X_2 įtaką kintamajam Y.

4. $var(\widehat{\beta}_2)$ vardiklyje yra dydis $1-r_{23}^2$, o r_{23} matuoja koreliaciją tarp dydžių X_2 ir X_3. Prisiminkime, kad koreliacija matuoja tiesinį sąryšį tarp kintamųjų. Jei dydžiai X_2 ir X_3 koreliuoja, tai $1-r_{23}^2$ yra mažesnė už vienetą trupmena. Kuo didesnė koreliacija tarp tų dydžių, tuo arčiau nulio bus koeficientas $1-r_{23}^2$, tuo didesnė bus variacija $var(\widehat{\beta}_2)$.

Mėsainių pavyzdyje suskaičiuojame

$$\widehat{cov}(\widehat{\boldsymbol{\beta}}) = \begin{pmatrix} 42.026 & -19.863 & -0.16111 \\ -19.863 & 10.184 & -0.05402 \\ -0.16111 & -0.05402 & 0.02787 \end{pmatrix}.$$

3.4.3 teiginys. *Dispersijos įverinimas yra nepaslinktas, t.y.* $E\widehat{\sigma}^2 = \sigma^2$.

Įrodymas. Pasinaudoję (3.13) sąryšiu, gauname:

$$E\widehat{\sigma}^2 = E(T-d)^{-1}\widehat{\varepsilon}^{\tau}\widehat{\varepsilon} =$$
$$(T-d)^{-1}E(\boldsymbol{M\varepsilon})^{\tau}\boldsymbol{M\varepsilon} = (T-d)^{-1}E\boldsymbol{\varepsilon}^{\tau}\boldsymbol{M}^{\tau}\boldsymbol{M\varepsilon} =$$

3.4.2 Gauso-Markovo teorema

Priminsime, kad $T \times T$ matrica $\boldsymbol{V}$ vadinama teigiamai apibrėžta (žymėsime $\boldsymbol{V} \geq 0$), jei $\boldsymbol{x}^{\tau}\boldsymbol{V}\boldsymbol{x} \geq 0$ su bet kuriuo $\boldsymbol{x} \in R^T$. Jei $\boldsymbol{V}_1$ ir $\boldsymbol{V}_2$ yra dvi teigiamai apibrėžtos matricos, tai sakysime, kad $\boldsymbol{V}_1 \geq \boldsymbol{V}_2$, jei $\boldsymbol{V}_1 - \boldsymbol{V}_2 \geq 0$.

2 teorema. (Gauso–Markovo) *Klasikinio tiesinio modelio parametro $\boldsymbol{\beta}$ mažiausių kvadratų įvertinimas $\widehat{\boldsymbol{\beta}}$, tarp visų tiesinių nepaslinktų parametro $\boldsymbol{\beta}$ įvertinimų turi mažiausią kovariacinę matricą.*

Įrodymas. Jau įrodėme, kad mažiausių kvadratų įvertinimas $\widehat{\boldsymbol{\beta}}$ yra nepaslinktas ir tiesinis. Tiesinių parametro $\boldsymbol{\beta}$ įvertinimų klasę galime apibrėžti taip: $\widetilde{\boldsymbol{\beta}} = \boldsymbol{CY}$, kai $\boldsymbol{C}$ yra bet kuri $T \times d$ eilės matrica. Įvertinimas $\widetilde{\boldsymbol{\beta}}$ yra nepaslinktas, kai

$$E\widetilde{\boldsymbol{\beta}} = E\boldsymbol{CY} = E\boldsymbol{C}(\boldsymbol{X\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{CX\beta} = \boldsymbol{\beta}.$$

Vadinasi, būtinai

$$\boldsymbol{CX} = \mathcal{I}. \tag{3.21}$$

Kadangi $\widetilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}} + [\boldsymbol{C} - \boldsymbol{A}]\boldsymbol{Y}$, pritaikę (3.12) formulę, išvedame

$$\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta} + \boldsymbol{A\varepsilon} + (\boldsymbol{C} - \boldsymbol{A})\boldsymbol{Y} =$$
$$= \boldsymbol{\beta} + \boldsymbol{A\varepsilon} + (\boldsymbol{C} - \boldsymbol{A})\boldsymbol{\varepsilon} = \boldsymbol{\beta} + \boldsymbol{C\varepsilon}.$$

Dabar galime suskaičiuoti įvertinimo $\widetilde{\boldsymbol{\beta}}$ kovariacinę matricą

$$cov(\widetilde{\boldsymbol{\beta}}) = \sigma^2 \boldsymbol{C}\boldsymbol{C}^\tau = \sigma^2 \boldsymbol{A}\boldsymbol{A}^\tau + \sigma^2(\boldsymbol{C}\boldsymbol{C}^\tau - \boldsymbol{A}\boldsymbol{A}^\tau).$$

Bet $\boldsymbol{C}\boldsymbol{A}^\tau = \boldsymbol{A}\boldsymbol{C}^\tau = \boldsymbol{A}\boldsymbol{A}^\tau = (\boldsymbol{X}^\tau\boldsymbol{X})^{-1}$, todėl $(\boldsymbol{C}\boldsymbol{C}^\tau - \boldsymbol{A}\boldsymbol{A}^\tau) = (\boldsymbol{C}-\boldsymbol{A})(\boldsymbol{C}-\boldsymbol{A})^\tau$. Kadangi pastaroji matrica yra neneigiamai apibrėžta, tai

$$cov(\widetilde{\boldsymbol{\beta}}) \geq \sigma^2 \boldsymbol{A}\boldsymbol{A}^\tau = cov(\widehat{\boldsymbol{\beta}}).$$

■

3.4.3 GKTR MODELIO SAVYBĖS

Dar daugiau savybių galime surasti klasikiniam tiesiniam Gausiniam modeliui. Šiame skyrelyje tirsime klasikinį tiesinį Gausinį modelį:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{3.22}$$

Paklaidų vektorius $\boldsymbol{\varepsilon}$ turi normalinį skirstinį su nuliniu vidurkiu ir kovariacijų matrica $cov(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}$. Taigi ir vektorius $\boldsymbol{Y} \overset{\mathcal{D}}{\sim} N_T(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$.

Nagrinėsime parametrų įvertinius, surastus mažiausių kvadratų metodu. Priminsime, kad parametro $\boldsymbol{\beta}$ mažiausių kvadratų įvertinimas yra

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\tau\boldsymbol{X})^{-1}\boldsymbol{X}^\tau\boldsymbol{Y} = \boldsymbol{\beta} + \boldsymbol{A}\boldsymbol{\varepsilon},$$

o dispersijos σ^2 įvertinimas $\widehat{\sigma}^2 = (T-d)^{-1}\widehat{\varepsilon}^\tau\widehat{\varepsilon}$; čia $\widehat{\varepsilon} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ – regresijos liekanų vektorius.

3.4.4 teiginys. *Klasikinio tiesinio Gausinio (3.22) modelio atveju d-matė statistika*

$$\boldsymbol{\theta}_T^{\boldsymbol{\beta}} = \sigma^{-1}(\boldsymbol{X}^\tau\boldsymbol{X})^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$$

turi standartinį d-matį normalinį skirstinį.

Įrodymas. Pasinaudosime tokia savybe: jei $\xi \overset{\mathcal{D}}{\sim} N_d(\boldsymbol{a}, \Sigma)$ ir $\boldsymbol{V}$ yra $d \times d$ matrica, tai

$$\boldsymbol{V}\xi \quad \overset{\mathcal{D}}{\sim} \quad N_d(\boldsymbol{V}\boldsymbol{a}, \boldsymbol{V}\Sigma\boldsymbol{V}^\tau).$$

Tai išplaukia iš kintamųjų keitimo formulės. Kadangi $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \boldsymbol{A}\boldsymbol{\varepsilon}$ ir $\boldsymbol{\varepsilon} \overset{\mathcal{D}}{\sim} N_T(0, \sigma^2\boldsymbol{I})$, tai

$$\begin{aligned}
&\boldsymbol{A}\boldsymbol{\varepsilon} \overset{\mathcal{D}}{\sim} N_T(0, \sigma^2\boldsymbol{A}\boldsymbol{A}^\tau),\\
&\sigma^{-1}(\boldsymbol{X}\boldsymbol{X}^\tau)^{1/2}\boldsymbol{A}\boldsymbol{\varepsilon} \overset{\mathcal{D}}{\sim} N_T(0, (\boldsymbol{X}\boldsymbol{X}^\tau)^{1/2}\boldsymbol{A}\boldsymbol{A}^\tau((\boldsymbol{X}\boldsymbol{X}^\tau)^{1/2})^\tau).
\end{aligned}$$

Lieka pastebėti, kad

$$(\boldsymbol{X}\boldsymbol{X}^\tau)^{1/2}\boldsymbol{A}\boldsymbol{A}^\tau((\boldsymbol{X}\boldsymbol{X}^\tau)^{1/2})^\tau = (\boldsymbol{X}\boldsymbol{X}^\tau)^{1/2}\boldsymbol{A}\boldsymbol{A}^\tau(\boldsymbol{X}\boldsymbol{X}^\tau)^{1/2} =$$
$$(\boldsymbol{X}\boldsymbol{X}^\tau)^{1/2}(\boldsymbol{X}\boldsymbol{X}^\tau)^{-1}(\boldsymbol{X}\boldsymbol{X}^\tau)^{1/2} = \boldsymbol{I}.$$

■

3.4.5 teiginys. *Klasikinio tiesinio Gausinio (3.22) modelio atveju statistika*

$$\frac{T-d}{\sigma^2}\widehat{\sigma}^2 = \frac{1}{\sigma^2}\widehat{\boldsymbol{\varepsilon}}^\tau\widehat{\boldsymbol{\varepsilon}}$$

turi χ^2 *skirstinį su* $T-d$ *laisvės laipsniais.*

Įrodymas. Kadangi $\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{M}\boldsymbol{\varepsilon}$, tai $\widehat{\boldsymbol{\varepsilon}}^\tau\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{\varepsilon}^\tau\boldsymbol{M}^\tau\boldsymbol{M}\boldsymbol{\varepsilon}$. Be to, $\boldsymbol{M}^\tau = \boldsymbol{M}$ ir $\boldsymbol{M}^2 = \boldsymbol{M}$, o $rank\,\boldsymbol{M} = T-d$. Todėl egzistuoja tokia ortogonalioji matrica $\boldsymbol{O}$, su kuria $\boldsymbol{O}\boldsymbol{M}\boldsymbol{O}^\tau$ yra diagonalinė matrica, kurios diagonalėje yra $T-d$ vienetukų ir d nulių. Kadangi normalinis skirstinys yra invariantinis ortogonaliųjų transformacijų atžvilgiu, tai $\boldsymbol{\varepsilon} \overset{\mathcal{D}}{\sim} \boldsymbol{O}\boldsymbol{\varepsilon}$. Taigi

$$\widehat{\boldsymbol{\varepsilon}}^\tau\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{\varepsilon}^\tau\boldsymbol{M}^\tau\boldsymbol{M}\boldsymbol{\varepsilon} \overset{\mathcal{D}}{\sim} \boldsymbol{\varepsilon}^\tau\boldsymbol{O}^\tau\boldsymbol{M}^\tau\boldsymbol{M}\boldsymbol{O}\boldsymbol{\varepsilon} \overset{\mathcal{D}}{\sim} \varepsilon_1^2 + \cdots + \varepsilon_{T-d}^2.$$

Kadangi $\sigma^{-2}\varepsilon_t^2$ turi standartinį normalinį skirstinį, tai

$$\frac{1}{\sigma^2}\widehat{\boldsymbol{\varepsilon}}^\tau\widehat{\boldsymbol{\varepsilon}} \overset{\mathcal{D}}{\sim} \frac{1}{\sigma^2}(\varepsilon_1^2 + \cdots + \varepsilon_{T-d}^2)$$

turi χ^2_{T-d} skirstinį. ■

3.4.6 teiginys. *Klasikinio tiesinio Gausinio (3.22) modelio atveju su kiekvienu* $\boldsymbol{a} \in R^d$ *statistika*

$$\widehat{\theta}^{\boldsymbol{\beta}}_{T,\boldsymbol{a}} = \frac{1}{\widehat{\sigma}\sqrt{\boldsymbol{a}^\tau\boldsymbol{a}}}\boldsymbol{a}^\tau(\boldsymbol{X}^\tau\boldsymbol{X})^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$$

turi Stjudento skirstinį su $T-d$ *laisvės laipsniais.*

Įrodymas. Pažymėję

$$\xi = \frac{1}{\widehat{\sigma}\sqrt{\boldsymbol{a}^\tau\boldsymbol{a}}}\boldsymbol{a}^\tau(\boldsymbol{X}^\tau\boldsymbol{X})^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}); \quad \eta^2 = \frac{T-d}{\sigma^2}\widehat{\sigma}^2,$$

gauname, kad

$$\widehat{\theta}^{\boldsymbol{\beta}}_{T,\boldsymbol{a}} = \sqrt{T-d}\,\frac{\xi}{\eta}.$$

Atsitiktinis dydis ξ turi standartinį normalinį skirstinį. Tikrai, jei $\nu \sim_P N_d(0,\Sigma)$, tai su kiekvienu $\boldsymbol{a} \in R^d$ atsitiktinis dydis $\boldsymbol{a}^\tau \nu \sim_P N(0, \boldsymbol{a}^\tau \Sigma \boldsymbol{a})$. Vadinasi,

$$\sigma^{-1}\boldsymbol{a}^\tau(\boldsymbol{X}^\tau\boldsymbol{X})^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \overset{\mathcal{D}}{\sim} N(0,\boldsymbol{a}^\tau\boldsymbol{a})$$

ir

$$\xi = \frac{1}{\sigma\sqrt{\boldsymbol{a}^\tau\boldsymbol{a}}}\boldsymbol{a}^\tau(\boldsymbol{X}^\tau\boldsymbol{X})^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})\boldsymbol{a} \quad \overset{\mathcal{D}}{\sim} \quad N(0,1).$$

Remiantis (3.4.5) teiginiu η^2 turi χ^2 skirstinį su $T-d$ laisvės laipsniais. Lieka įrodyti, kad atsitiktiniai dydžiai ξ ir η yra nepriklausomi. Tam pakanka įrodyti, kad atsitiktinis vektorius $\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta} = \boldsymbol{A\varepsilon}$ ir atsitiktinis dydis $\widehat{\boldsymbol{\varepsilon}}^\tau\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{\varepsilon}^\tau\boldsymbol{M}^\tau\boldsymbol{M\varepsilon}$ yra nepriklausomi. Tai savo ruožtu, bus teisinga, jei įrodysime atsitiktinių vektorių $\boldsymbol{A\varepsilon}$ ir $\boldsymbol{M\varepsilon}$ nepriklausomumą. Kadangi abu vektoriai turi Gauso skirstinius, tai jie yra nepriklausomi, jeigu jie yra nekoreliuoti. Suskaičiuokime

$$cor(\boldsymbol{A\varepsilon},\boldsymbol{M\varepsilon}) = E\boldsymbol{A\varepsilon\varepsilon}^\tau\boldsymbol{M}^\tau = \sigma^2\boldsymbol{AM}^\tau =$$
$$\sigma^2((\boldsymbol{X}^\tau\boldsymbol{X})^{-1}\boldsymbol{X}^\tau)(\boldsymbol{I}-\boldsymbol{X}(\boldsymbol{X}^\tau\boldsymbol{X})^{-1}\boldsymbol{X}^\tau) = 0.$$

∎

3.5 Intervalinis parametrų įvertinimas

Intervalinis parametrų įvertinimas remiasi lygtimi

$$P\Big(-t_c \le \frac{\widehat{\beta}_k-\beta_k}{se(\widehat{\beta}_k)} \le t_c\Big) = 1-\alpha. \tag{3.23}$$

Čia α – pasirenkamas lygmuo. Dydis t_c nežinomas. Jį randame, kai žinome statistikos

$$\frac{\widehat{\beta}_k-\beta_k}{se(\widehat{\beta}_k)}$$

skirstinį. GKTR modelio atveju statistikos skirstinys yra t skirstinys su $T-d$ laisvės laipsniais. Iš t skirstinio lentelių randame tokį t_c, kad $P(t \ge t_c) = \alpha/2$. Intervalas

$$[\widehat{\beta}_k - t_c se(\widehat{\beta}_k), \widehat{\beta}_k + t_c se(\widehat{\beta}_k)]$$

yra parametro β_k $100(1-\alpha)$-procentinis paskiliautinis intervalas. Grįždami prie mėsainių pavyzdžio, turime

$$\begin{aligned} T &= 52, & d &= 3; \\ \widehat{\beta}_1 &= 104.79, & se(\widehat{\beta}_1) &= \sqrt{\widehat{var}(\widehat{\beta}_1)} = 6.483; \\ \widehat{\beta}_2 &= -6.642, & se(\widehat{\beta}_2) &= \sqrt{\widehat{var}(\widehat{\beta}_2)} = 3.191; \\ \widehat{\beta}_3 &= 2.984, & se(\widehat{\beta}_3) &= \sqrt{\widehat{var}(\widehat{\beta}_3)} = 0.1669. \end{aligned}$$

Šią informaciją panaudosime 95 procentų paskliautinių intervalų radimui. Laisvės laipsnių skaičius yra $T-d=52-3=49$. Kritinė reikšmė yra $t_c=2.01$. Parametrui β_2 pasikliautinis intervalas yra $[-13.06,-0.23]$. Intervalas akivaizdžiai per didelis ir nėra informatyvus. Kitaip tariant, parametro β_2 įvertis $\widehat{\beta}_2=-6.642$ yra nerealus, nes jo standartinė paklaida didelė.

Parametrui β_3 95-procentinis paskliautinis intervalas yra $(2.65, 3.32)$, Tai pakankamai siauras intervalas. Jis reiškia, kad, padidinus reklamos išlaidas 1000 Lt, pelnas padidės tarp 2650 ir 2320 Lt.

Parinkę Borelio aibę $A\subset R^d$ iš lygties

$$(2\pi)^{-d/2}\int_A \exp\{-\boldsymbol{x}^\tau \boldsymbol{x}/2\}d\boldsymbol{x}=\theta \tag{3.24}$$

randame, kad

$$P(\boldsymbol{\beta}\in\widehat{\boldsymbol{\beta}}+\sigma(\boldsymbol{X}\boldsymbol{X}^\tau)^{1/2}A)=\theta. \tag{3.25}$$

Priminsime, kad $F(A)=\{F(\boldsymbol{x}):\boldsymbol{x}\in A\}$, o $A+\boldsymbol{x}=\{\boldsymbol{y}+\boldsymbol{x}:\boldsymbol{y}\in A\}$. Vadinasi, $\widehat{\boldsymbol{\beta}}+\sigma(\boldsymbol{X}\boldsymbol{X}^\tau)^{1/2}A$ yra vektoriaus $\boldsymbol{\beta}$ θ-pasikliautinumo aibė. Akivaizdu, kad (3.24) lygties sprendinys yra daugiareikšmis. Todėl tikslinga nagrinėti tokias aibes, kurias aprašo vienas parametras ir kurį galime vienareikšmiškai rasti, išsprendę (3.24) lygtį. Pavyzdžiui, nagrinėkime aibę $A=A_r=\{\boldsymbol{x}\in R^d:\boldsymbol{x}^\tau\boldsymbol{x}\le r\}$. Tai – Euklidinis rutulys, kurio centras yra koordinačių pradžioje, o spindulys r. Šiuo atveju (3.24) lygtis virsta lygtimi

$$(2\pi)^{-d/2}\int_{A_r}\exp\{-\boldsymbol{x}^\tau\boldsymbol{x}/2\}d\boldsymbol{x}=\int_0^r e^{-u^2/2}du=\theta.$$

Ją išsprendę, randame $r=r_\theta$ ir atitinkamą pasikliautinę sritį

$$\{\boldsymbol{u}\in R^d:\quad \sigma(\boldsymbol{u}+\widehat{\boldsymbol{\beta}})^\tau\boldsymbol{X}^\tau\boldsymbol{X}(\boldsymbol{u}+\widehat{\boldsymbol{\beta}})=r_\theta\}.$$

Tai yra elipsoidas.

3.4.5 teiginį galime pritaikyti dispersijos pasikliautinėms sritims aprašyti. Iš lygties

$$\theta=P(t_1\le\chi^2_{T-d}\le t_2)$$

parinkę t_1,t_2, gauname, kad

$$\Big(\frac{\widehat{\sigma}^2}{(T-d)t_2},\quad\frac{\widehat{\sigma}^2}{(T-d)t_1}\Big)$$

yra parametro σ^2 θ-procentinis pasikliautinis intervalas.

3.6 KTR MODELIS: HIPOTEZIŲ TIKRINIMAS

3.6.1 Vienpusės hipotezės

Pradėsime mėsainių pavyzdžiu ir, sekdami jau anksčiau turėta vienpusės hipotezės patikrinimo schema, patikrinsime paklausos elastingumą kainoms. Paklausos elastingumo atžvilgiu svarbu išsiaiškinti, ar kainos sumažėjimas reiškia pelno sumažėjimą (paklausa neelastinga kainai), ar atvirkščiai – kainos sumažėjimas padidina pelną (paklausa elastiška kainai). Pasinaudodami anksčiau aptarta hipotezių patikrinimo schema, pirmiausia nustatiome nulinę hipotezę.

1. $H_0: \quad \beta_2 \geq 0$ (reikštų, kad paklausa nėra elastiška kainai)
2. $H_1: \quad \beta_2 < 0$ (paklausa yra elastiška kainai)
3. Sudarydami statistiką, elgiamės taip, tarsi nagrinėtume hipotezę $\beta_2 = 0$. Jei ji teisinga, tai testinė statistika būtų
$$t = \frac{\widehat{\beta}_2}{se(\widehat{\beta}_2)} \sim t_{(T-d)},$$
čia
$$se(\widehat{\beta}_2) = \left(\widehat{var}(\widehat{\beta}_2)\right)^{1/2}.$$
4. Kritinę sritį sudarys tokios reikšmės, kurios neturėtų pasirodyti, jei hipotezė teisinga. 5% pasikliovimo kritinę reikšmę t_c randame iš lygties
$$P(t_{(T-d)} \leq t_c) = 0.05.$$
Nulinę hipotezę atmetame, jei $t \leq t_c$. Turėdami imtį, kurios tūris $T = 52$, ir modelį su dviem regresoriais, iš lentelių randame $t_c = -1.68$.
5. Suskaičiuojame statistikos reikšmę
$$t = \frac{\widehat{\beta}_2}{se(\widehat{\beta}_2)} = \frac{-6.642}{3.191} = -2.08.$$
Kadangi $t = -2.08 < t_c = -1.68$, nulinę hipotezę atmetame alternatyvos $H_1: \quad \beta_2 > 0$ naudai. Šiuo atveju alternatyva labiau atspindinti duomenis.
6. Suskaičiuojame duomenis atitinkančią p-reikšmę: $p = 0.021 = P(t_{(38)} < -2.08)$.

Kita įdomi hipotezė: ar išlaidų reklamai padidinimas pakankamai padidins pajamas (tiek kad atsipirktų pačios išlaidos reklamai). Taip bus, jei $\beta_3 > 1$. Patikrinkime šią hipotezę.

1. $H_0: \quad \beta_3 \leq 1.$
2. $H_1: \quad \beta_3 > 1.$
3. Testinė statistika yra
$$t = \frac{\widehat{\beta}_3 - 1}{se(\widehat{\beta}_3)} \sim t_{(T-d)}.$$
4. Pasirinkę reikšmingumo lygmenį $\alpha = 0.05$, nulinę hipotezę atmesime, jei $t \geq t_c = 1.68$.
5. Testinės statistikos reikšmė yra
$$t = \frac{2.984 - 1}{0.1669} = 11.89.$$
Kadangi $11.89 > 1.68$, nulinę hipotezę atmetame alternatyvos naudai. Šiuo atveju ir p reikšmė yra labai maža (apie 10^{-12}). Taigi, turime statistikinį patvirtinimą, kad reklamos išlaidų padidinimas pasiteisins išaugusiomis pajamomis.

Čia panaudoti vienpusiai testai. Todėl ir atitinkama kritinė sritis imta vienpusė, tai yra 1.68.

3.6.2 Koeficientų reikšmingumo testai

Nagrinėdami daugelio kintamųjų regresinį modelį, paprastai esame įsitikinę, kad kiekvienas parinktas regresorius įtakoja endogeninį kintamąjį. Jei kintamasis X_k nedaro įtakos dydžiui Y, tai turėtų būti $\beta_k = 0$. Šį teiginį reikia patikrinti. Atitinkamas testas vadinamas reikšmingumo testu.

Norėdami patikrinti ar duomenyse yra informacijos apie endogeninio kintamojo Y sąryšį su kintamuoju X_k, tikriname nulinę hipotezę
$$H_0: \quad \beta_k = 0$$
atžvilgiu alternatyvos
$$H_1: \quad \beta_k \neq 0.$$
Jei nulinė hipotezė teisinga, tai Gausinio modelio atveju
$$t = \frac{\widehat{\beta}_k}{se(\beta_k)} \quad \overset{\mathcal{D}}{\sim} \quad t_{T-d}.$$

Kritinę sritį nustatome atsižvelgdami į alternatyvą. Dvipusės alternatyvos atveju (nelygu nuliui) kritinė sritis yra $\{|t| > t_c\}$, kai t_c randama iš lygties

$$P(|t_{T-d}| \geq t_c) = \alpha/2.$$

Patikrinkime kiekvieno koeficiento reikšmingumą nagrinėjamame PRK pavyzdyje.

Pirmiausia patikrinkime kainos įtaką pelnui.

1. $H_0: \quad \beta_2 = 0$
2. $H_1: \quad \beta_2 \neq 0$
3. Testinė statistika yra $t = \widehat{\beta}_2/se(\beta_2)$.
4. Parinkę reikšmingumo lygmenį $\alpha = 0.05$, randame kritinę reikšmę $t_c = 2.01$. Taigi kritinė sritis yra $\{|t| > 2.01\}$.
5. Suskaičiuojame statistikos reikšmę turimiems duomenims:

   $$t = \frac{-6.642}{3.191} = -2.08.$$

   Kadangi $|-2.08| = 2.08 > 2.01$, nulinę hipotezę atmetame alternatyvos naudai.
6. p-reikšmė yra $P(|t_49| > 2.08) = 0.042 < 0.05$.

Patikrinkime reklamos reikšmingumą.

1. $H_0: \quad \beta_3 = 0$
2. $H_1: \quad \beta_3 \neq 0$
3. Testinė statistika yra $t = \widehat{\beta}_3 se(\beta_3)$. Kai teisinga nulinė hipotezė ir modelis yra Gausinis, tai $t \overset{\mathcal{D}}{\sim} t_{(49)}$.
4. Parinkę reikšmingumo lygmenį $\alpha = 0.05$, randame kritinę reikšmę $t_c = 2.01$. Taigi kritinė sritis yra $\{|t| > 2.01\}$.
5. Suskaičiuojame statistikos reikšmę turimiems duomenims:

   $$t = \frac{2.984}{0.1669} = 17.88$$

   Kadangi $17.88 > 2.01$, nulinę hipotezę atmetame alternatyvos naudai.

3.6.3 Tiesinių hipotezių testai

Nagrinėkime modelį

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

ir jo atžvilgiu hipotezę:

$$H_0: \quad \boldsymbol{R\beta} = \boldsymbol{u}.$$

Čia $\boldsymbol{R}$ yra duota $q \times d$ matrica, o $\boldsymbol{u} \in R^q$ – duotas vektorius. Kitaip sakant, reikia patikrinti, ar parametrai $\boldsymbol{\beta}$ tenkina hipoteze nusakytą apribojimą. Šios hipotezės tikrinimui galime pritaikyti keletą metodų, kurie tiesinio modelio ir tiesinių apribojimų atveju yra ekvivalentūs. Pirmojo metodo esmė tokia. Nagrinėjame parametro $\boldsymbol{\beta}$ mažiausių kvadratų įvertinimą $\widehat{\boldsymbol{\beta}}$. Suskaičiuojame, "kiek skiriasi" $\boldsymbol{R}\widehat{\boldsymbol{\beta}}$ nuo $\boldsymbol{u}$. Jei tas skirtumas didelis, hipotezę atmetame kaip nepagrįstą, o jei nėra didelis – tariame, kad duomenys hipotezei neprieštarauja. Norint šį metodą realizuoti, reikia apibrėžti statistiką, kuri aprašytų dydžių $\boldsymbol{R}\widehat{\boldsymbol{\beta}}$ ir $\boldsymbol{u}$ atstumą.

Pirmiausia tarkime, kad $q = 1$. Tai yra, $\boldsymbol{u} = u \in R$ yra skaliaras, o $\boldsymbol{R}$ – $1 \times d$ vektorius eilutė. Tuomet tiesiog galime nagrinėti skirtumą $\boldsymbol{R}\widehat{\boldsymbol{\beta}} - u$. Apibrėžkime statistiką

$$\sigma^{-1}(\boldsymbol{R}(\boldsymbol{X}^\tau \boldsymbol{X})^{-1}\boldsymbol{R}^\tau)^{-1/2}(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - u).$$

Jeigu teisinga hipoteze H_0, t.y. $u = \boldsymbol{R\beta}$, tai

$$\begin{aligned}&\tau_T = \sigma^{-1}(\boldsymbol{R}(\boldsymbol{X}^\tau \boldsymbol{X})^{-1}\boldsymbol{R}^\tau)^{-1/2}(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{u}) = \\ &\sigma^{-1}(\boldsymbol{R}(\boldsymbol{X}^\tau \boldsymbol{X})^{-1}\boldsymbol{R}^\tau)^{-1/2}\boldsymbol{R}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim_P N(0,1).\end{aligned}$$

Tačiau dažniausiai dispersija σ^2 nėra žinoma. Todėl ją reikia pakeisti įvertinimu $\widehat{\sigma}^2$.

3 teorema. *KTGM atveju, jei teisinga hipotezė H_0 ir $q = 1$, tai statistika*

$$\widehat{\tau}_T = \widehat{\sigma}^{-1}(\boldsymbol{R}(\boldsymbol{X}^\tau \boldsymbol{X})^{-1}\boldsymbol{R}^\tau)^{-1/2}(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{u}) \sim_P t_{T-d}.$$

Įrodymas. Statistiką $\widehat{\tau}_T$ galime užrašyti

$$\widehat{\tau}_T = \frac{\sqrt{T-d}\,\tau_T}{\sqrt{(T-d)\widehat{\sigma}^2/\sigma^2}}.$$

Kadangi $\tau_T \sim_P N(0,1)$, o $(T-d)\widehat{\sigma}^2/\sigma^2 \sim_P \chi^2_{T-d}$, lieka pastebėti, kad tie du dydžiai yra nepriklausomi. Tam pakanka įsitikinti, kad atsitiktiniai dydžiai $\mathcal{R}\widehat{\boldsymbol{\beta}}$ ir $\widehat{\boldsymbol{\varepsilon}}^\tau\widehat{\boldsymbol{\varepsilon}}$ yra nepriklausomi. ■

Jeigu alternatyva yra $\boldsymbol{R\beta} \neq u$, tai testas yra dvipusis, o jei alternatyva yra $\boldsymbol{R\beta} > u$ (arba $\boldsymbol{R\beta} < u$), – vienpusis. Bet kurio testo atveju pirmiausia pasirenkame patikimumo lygmenį – mažą skaičių α ir randame kritinę sritį G_α. Statistikos $\widehat{\tau}_T$ skirstinys yra stjudento su $T-d$ laisvės laipsniais. Dvipusio testo atveju iš stjudento kvantilių lentelės randame $\pm t_{T-d}(\alpha/2)$ tokius, kad

$$\begin{aligned}&\alpha = P(\widehat{\tau}_T < -t_{T-d}(\alpha/2) \quad \text{arba} \;\; \widehat{\tau}_T > t_{T-d}(\alpha/2)) = \\ &1 - P(-t_{T-d}(\alpha/2) \leq \widehat{\tau}_T \leq t_{T-d}(\alpha/2)).\end{aligned}$$

Tada kritinė sritis yra

$$G_\alpha = (-\infty, -t_{T-d}(\alpha/2)) \cup (t_{T-d}(\alpha/2), \infty).$$

Nulinę hipotezę atmetame su reikšmingumo lygmeniu α, jei statistikos $\widehat{\tau}_T$ reikšmė patenka į kritinę sritį G_α.

Panagrinėkime hipotezę

$$H_0: \quad \beta_i = 0$$

prieš alternatyvą

$$H_1: \quad \beta_i \neq 0.$$

Tai yra jau nagrinėtas koeficientų reikšmingumas. Nesunku matyti, kad tai yra atskiras bendros tiesinės hipotezės atvejis, kai

$$\mathcal{R} = (0, \ldots, 0, 1, 0, \ldots, 0).$$

Tarkime, m_{ii} yra matricos $(\boldsymbol{X}^\tau \boldsymbol{X})^{-1}$ ii-tasis elementas. Tuomet $\mathcal{R}(\boldsymbol{X}^\tau \boldsymbol{X})^{-1}\mathcal{R}^\tau = m_{ii}$ ir atitinkama t-statistika yra

$$\tau = \frac{\widehat{\beta}_i}{\widehat{\sigma}_T \sqrt{m_{ii}}},$$

kuri turi stjudento skirstinį su $T - d$ laisvės laipsniais. Be to, nesunku matyti, kad $t = \widehat{\beta}_i / se(\beta_i)$. Tai yra ta pati testinė statistika, kurią naudojome testuodami parametrų reikšmingumą.

3.6.1 pavyzdys. Nagrinėkime hipotezę $\beta_i + \beta_j = 0$, kuri yra tiesinė ir aprašoma matrica

$$\boldsymbol{R} = (0, \ldots, 0, 1, 0, \ldots, 0, 1, 0, \ldots, 0).$$

Šiuo atveju $\boldsymbol{R}(\boldsymbol{X}^\tau \boldsymbol{X})^{-1}\boldsymbol{R}^\tau = m_{ii} + 2m_{ij} + m_{jj}$ ir atitinkama statistika

$$\tau = \frac{\widehat{\beta}_i + \widehat{\beta}_j}{\widehat{\sigma}(m_{ii} + 2m_{ij} + m_{jj})^{1/2}}$$

turi stjudento skirstinį su $T - d$ laisvės laipsniais.

Tarkime, kad modelis yra klasikinis gausinis, t. y. $\varepsilon \sim_P N_T(0, \sigma^2 \boldsymbol{I})$. Jei H_0 teisinga, tai

$$\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{u} = \boldsymbol{R}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim_P N_q(0, \sigma^2 \boldsymbol{R}(\boldsymbol{X}^\tau \boldsymbol{X})^{-1}\boldsymbol{R}^\tau).$$

Jei matrica $\boldsymbol{R}(\boldsymbol{X}^\tau \boldsymbol{X})^{-1}\boldsymbol{R}^\tau$ yra neišsigimusi, t.y. egzistuoja

$$(\boldsymbol{R}(\boldsymbol{X}^\tau \boldsymbol{X})^{-1}\boldsymbol{R}^\tau)^{-1/2} = \boldsymbol{U},$$

tai

$$\boldsymbol{U}(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{u}) \sim_P N_q(0, \sigma^2 \boldsymbol{I}).$$

Taigi statistika

$$\sigma^{-2}(\boldsymbol{R}\widehat{\boldsymbol{\beta}}-\boldsymbol{u})^{\tau}(\boldsymbol{R}(\boldsymbol{X}^{\tau}\boldsymbol{X})^{-1}\boldsymbol{R})^{-1}(\boldsymbol{R}\widehat{\boldsymbol{\beta}}-\boldsymbol{u}) \sim_P \chi^2_{(q)}.$$

Ji aprašo atstumą tarp $\mathcal{R}\widehat{\boldsymbol{\beta}}$ ir $\boldsymbol{u}$. Tačiau, kaip matėme anksčiau, dispersija σ^2 dažniausiai yra nežinoma. Todėl ją reikia pakeisti įvertinimu. Galiausiai apibrėžkime statistiką taip:

$$\mathcal{T}_{T,q} = q^{-1}\widehat{\sigma}^{-2}(\boldsymbol{R}\widehat{\boldsymbol{\beta}}-\boldsymbol{u})^{\tau}(\boldsymbol{R}(\boldsymbol{X}^{\tau}\boldsymbol{X})^{-1}\boldsymbol{R}^{\tau})^{-1}(\boldsymbol{R}\widehat{\boldsymbol{\beta}}-\boldsymbol{u}).$$

3.6.1 teiginys. *Klasikinio tiesinio gausinio modelio atveju, jei teisinga hipotezė H_0, tai statistika $\mathcal{T}_{T,q}$ turi Fišerio skirstinį su $(q, T-d)$ laisvės laipsniais.*

Statistika $\mathcal{T}_{T,q}$ dažnai vadinama tiesiog F statistika.

Kol kas esame patikrinę hipotezes apie kiekvieno atskirai paimto regresinio modelio parametro reikšmingumą. Tam tikslui naudojome t-testą. Šiame skyrelyje patikrinsime kiek sudėtingesnę hipotezę. Nagrinėkime klasikinį tiesinį modelį

$$Y_y = \beta_1 + \beta_2 X_{2t} + \cdots + \beta_d X_{dt} + \varepsilon_t.$$

Mus domina nulinė hipotezė

$$H_0: \quad \beta_2 = \beta_3 = \cdots = \beta_d = 0$$

atžvilgiu alternatyvos

$$H_1: \quad \text{bent vienas iš koeficientų} \quad \beta_k \neq 0.$$

Jei nulinė hipotezė teisinga, tai nė vienas iš paaiškinančiųjų kintamųjų nėra reikšmingas. Taigi ir modelis tuo atveju yra beprasmis. Alternatyva nenurodo, kuris iš paaiškinančiųjų kintamųjų yra reikšmingas, tačiau pasako, kad modelis nėra bevertis. Tokios nulinės hipotezės patikrinimas atžvilgiu duotos alternatyvos dažnai vadinamas *regresinio modelio bendrojo reikšmingumo* testavimu. Labai plačiai naudojamas testas, paremtas dviejų regresijos liekanų palyginimu. Vienos regresijos liekanos gaunamos nagrinėjant modelį be apribojimų, o kitos – tariant kad nulinė hipotezė teisinga. Testas remiasi idėja, kad, jei skirtumas tarp tų dviejų regresijos liekanų sumų yra reikšmingas, tai prielaida kad teisinga nulinė hipotezė yra teisinga, aiškiai sumažina galimybę, jog modelis teisingai aprašo duomenis. Todėl duomenys nepalaiko nulinės hipotezės. Jei nulinė hipotezė teisinga, tai tikėtina, kad duomenys palyginami su sąlygomis parametrams. Taigi liekanų kvadratų suma neturėtų smarkiai pasikeisti, jei teisinga nulinė hipotezė. Atitinkama statistika yra

$$F = \frac{(ESS_r - ESS_u)/q}{ESS_u/(T-d)}.$$

Čia ESS_u yra regresijos be apribojimų liekanų kvadratų suma, o ESS_r – regresijos su apribojimais liekanų kvadratų suma. Statistika F turi Fišerio skirstinį su $(q, T-d)$ laisvės laipsniais.

Iliustracijai panagrinėkime PRK modelį. Norime patikrinti hipotezę

$$H_0: \quad \beta_2 = 0, \quad \beta_3 = 0$$

prieš alternatyvą

$$H_1: \quad \text{arba } \beta_2 \neq 0, \quad \text{arba } \beta_3 \neq 0.$$

Iš lentelių randame

$$F = \frac{5888.09}{36.84} = 159.83.$$

Taip pat randame 5 procentų kritinę reikšmę, atitinkančią F skirstinį su $(2, 49)$ laisvės laipsniais. Ta reikšmė yra $F_c = 3.187$. Kadangi $159.83 > 3.187$, nulinę hipotezę atmetame alternatyvos naudai.

3.7 SUDERINTUMAS

Anksčiau apibrėžtas dydis R^2 tinka ir daugelio kintamųjų regresiniam modeliui. Teisingos ir tos pačios formulės. Priminsime, kad determinacijos koeficientas yra

$$R^2 = \frac{RSS}{TSS} = \frac{\sum(\widehat{Y}_t - \overline{Y})^2}{\sum(Y_t - \overline{Y})^2} =$$
$$1 - \frac{ESS}{TSS} = 1 - \frac{\sum \widehat{\varepsilon}_t^2}{\sum(Y_t - \overline{Y})^2}.$$

Nagrinėjamame pavyzdyje dalinė ANOVA lentelė yra tokia.

Variacija	DF	Kvadratų suma
Paaiškinta	2	11776.18
Nepaaiškinta	49	1805.168
Pilnoji	51	13581.35

Apskaičiuotas determinacijos koeficientas yra

$$R^2 = 1 - \frac{1805.168}{13581.35} = 0.867.$$

Vadinasi, net 86.7 procentus pilnosios variacijos pelno duomenyse paaiškinama modeliu, t. y. kainų ir išlaidų reklamai variacijomis. Tai labai aukštas procentas. Jis reiškia, kad duomenyse tik 13.3 procentai pelno variacijos lieka nepaaiškinta, ir ta dalis variacijos priskirtina paklaidoms.

Kaip minėta, determinacijos koeficientas yra vienas iš rodiklių, aprašančių modelio suderintumą su duomenimis. Mat R^2 yra ir koreliacijos koeficientas tarp $(\widehat{Y_t})$ ir (Y_t). Koreliacijos koeficientas, savo ruožtu, matuoja tiesinį sąryšį tarp kintamųjų.

Viena problema, kuri atsiranda, naudojant R^2, yra ta, kad jo reikšmę galima dirbtinai padidinti, pridėjus papildomus paaiškinamuosius kintamuosius. Todėl daugelio kintamųjų regresiniams modeliams naudojamas pataisytas determinacijos koeficientas

$$\overline{R}^2 = 1 - \frac{ESS/(T-d)}{TSS/(T-1)}.$$

Nagrinėjamame pavyzdyje $\overline{R}^2 = 0.8617$. Taip apibrėžtas koeficientas vadinamas apibendrintu determinacijos koeficientu. Jis gerokai silpniau reaguoja į pridedamus papildomus kintamuosius dėl laisvės laipsnių skaičiaus $T-d$ vardiklyje.

3.8 Kai kurių ekonominių hipotezių patikrinimas

Nagrinėkime šiek tiek bendresnį ekonometrinį modelį mėsainių pavyzdžiui. Tikslinga kelti tokį klausimą: ar pajamų tiesinė priklausomybė nuo kainos ir reklamos išlaidų tinkamai atspindi realią situaciją? Ar modelis nėra labai jau idealistinis? Abejonių kelia tai, kad pajamos nepriklauso nuo išlaidų reklamai didėjimo greičio. Tai yra, kaip greitai bedidintume išlaidas reklamai (pvz. kiekvieną savaitę reklamai išleidžiama vis daugiau), pajamos kinta tuo pačiu dydžiu. Tai lengva matyti, nes pajamų išvestinė pagal išlaidas reklamai nepriklauso nuo išlaidų reklamai. Vienas iš būdų modeliui pagerinti – paaiškinančiuoju kintamuoju paimti dar ir a^2, ir nagrinėti modelį

$$tr_t = \beta_1 + \beta_2 p_t + \beta_3 a_t + \beta_4 a_t^2 + \varepsilon_t. \tag{3.26}$$

Taip gauname modelį, kuriame tikėtinų pajamų atsakas į reklamos išlaidas priklauso nuo reklamos išlaidų. Vidutinių pajamų atsakas į a yra

$$\frac{\partial E(tr_t)}{\partial a_t} = \beta_3 + 2\beta_4 a_t.$$

Kai a_t padidėja vienu vienetu (tūkstančiu litų), o p_t nekinta, $E(tr)$ padidėja $(\beta_3 + 2\beta_4 a_t) \cdot 1000$ Lt. Norėdami nustatyti tikėtinus koeficientų β_3 ir β_4 ženklus, turime pastebėti, kad pajamų atsakas į reklamą turėtų būti teigiamas, kai $a_t = 0$. Todėl tikėtina, kad $\beta_3 > 0$. Norint pasiekti pajamų sumažėjimą, atsakas turėtų mažėti, jei išlaidos reklamai didėja. Todėl visai tikėtina, kad $\beta_4 < 0$.

Gautą modelį reikia įvertinti. Aišku, laikydami reklamos išlaidų kvadratą tiesiog papildomu paaiškinančiuoju kintamuoju, gautą modelį galime užrašyti

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_2 X_{3t} + \beta_4 X_{4t} + \varepsilon_t,$$

kai

$$Y_t = tr_t,\ X_{2t} = p_t,\ X_{3t} = a_t,\ X_{4t} = a_t^2.$$

Naudodami turimus duomenis įvertiname parametrus ir gauname įvertintą modelį

$$\widehat{tr}_t = 104.81 - 6.582 p_t + 2.948 a_t + 0.0017 a_t^2$$

Ką galime pasakyti apie gautą rezultatą?

Jei modelis įvertintas netiksliai, vienas iš būdų jį pataisyti yra surinkti daugiau ir geresnių duomenų. Taip galėsime patikrinti modelio stabilumą. Turime dar 28 savaičių duomenis.

Sav.	Pajamos 1000 Lt	Kaina Lt	Reklama 1000 Lt	Sav.	Pajamos 1000 Lt	Kaina Lt	Reklama 1000 Lt
53	129.9	2.87	16.0	66	1o8.6	1.61	4.8
54	101.5	2.05	4.0	67	158.8	2.62	27.7
55	136.3	2.55	19.6	68	147.2	1.74	20.6
56	97.6	3.49	10.2	69	146.3	3.21	25.4
57	118.9	3.45	17.5	70	121.2	1.50	10.2
58	130.5	3.45	18.3	71	107.0	1.78	4.9
59	128.5	2.58	18.2	72	121.2	2.43	12.1
60	138.3	2.87	22.1	73	125.4	2.04	12.3
61	103.6	1.76	4.1	74	141.9	2.99	19.7

62	151.8	2.97	24.9	75	120.0	2.83	14.3
63	128.5	2.77	14.7	76	101.9	2.47	4.8
64	128.5	2.64	18.6	77	130.4	2.04	11.6
65	143.7	1.50	20.9	78	139.9	1.87	19.8

Vieno greito maisto restorano papildomi duomenys

Panaudoję šiuos duomenis gauname tokį įvertintą modelį:

$$\widehat{tr}_t = 110.46 - 10.198 p_t + 3.361 a_t - 0.0268 a_t^2.$$

Šie duomenys suderinami su ekonominiais argumentais dėl parametrų ženklų.

Dabar panagrinėkime, kaip patikrinti, ar reklama turi įtakos pelnui. Nagrinėkime praplėstą modelį

$$tr_t = \beta_1 + \beta_2 p_t + \beta_3 a_t + \beta_4 a_t^2 + \varepsilon_t.$$

Jei $\beta_3 = 0$ arba $\beta_4 = 0$, tai reklama pelno neįtakoja. Atlikime testą naudodami praplėstus duomenis.

1. Nulinė hipotezė $H_0: \quad \beta_3 = 0, \beta_4 = 0$. Ji turi pavidalą $\boldsymbol{R\beta} = u$, kai

$$\boldsymbol{R} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{u} = (0,0)^\tau.$$

2. Alternatyva H_1 : arba $\beta_3 \neq 0$ arba $\beta_4 \neq 0$.

3. Testinė statistika

$$F = \frac{(ESS_r - ESS_u)/2}{ESS_u/(78-4)} = 261.41.$$

4. Kritinė sritis atitinkanti paskiliautinumo lygmenį $\alpha = 0.05$ yra $F_c = 3.120$,t.y. $P(F_{(2.74)} > F_c) = 0.05$.

5. Kadangi $F = 261.41 > F_c = 3.120$ nulinę hipotezę atmetame alternatyvos naudai.

Išspręskime tokį optimizavimo uždavinį. PRK pavyzdyje reikia parinkti optimalų išlaidų reklamai planą.

Marginalinis pelnas atitinkantis vienetinį reklamos padidinimą yra

$$\frac{\Delta E(tr_t)}{\Delta a_t} = \beta_3 + 2\beta_4 a_t.$$

Marginalioji kaina reklamos vienetui yra pačios reklamos kaina plius papildomos produkcijos, reikalingos reklamos dėl reklamos išaugusiai paklausai patenkinti, gamybos kaina. Ignoruodami tą plius kainą, reklamos kainą reikia padidinti ten kur marginalinis pelnas lygus vienam litui atitinka vieno lito išlaidas reklamai, tai yra, kur

$$\beta_3 + 2\beta_4 a_t = 1.$$

Įvertinę parametrus gauname ir reklamos išlaidų įvertinimą:

$$3.361 + 2(-0.0268)\widehat{a}_t = 1.$$

Taigi, $\widehat{a}_t = 44.0485$. vadinasi, optimali reklamos išlaidų suma t-tai savaitei yra $44048.50 Lt$. Tarkime, kad yra manoma, jog tokia suma per didelė ir pakaktų $40000 Lt$. Patikrinkime hipotezę

1. $H_0 : \beta_3 + 2\beta_4(40) = 1.$

2. Alternatyva $H_1 : \beta_3 + 2\beta_4(40) \neq 1.$

3. Testinė statistika
$$t = \frac{(\widehat{\beta}_3 + 80\widehat{\beta}_4) - 1}{se(\beta_3 + 80\beta_4)}$$
esant teisingai nulinei hipotezei turi stjudento skirstinį su 74 laisvės laipsniais. Vienintelė problema skaičiuojant tos statistikos reikšmę yra suskaičiuoti vardiklį. Kadangi
$$\widehat{var}(\widehat{\beta}_3 + 80\widehat{\beta}_4) = \widehat{var}(\widehat{\beta}_3) + 80^2\widehat{var}(\widehat{\beta}_4) + 2 \cdot 80\widehat{cov}(\widehat{\beta}_3, \widehat{\beta}_4) = 0.76366.$$
Taigi $t = 0.252$.

4. Kritinė sritis atitinkanti reikšmingumo lygmenį 0.05 yra $t_c = 1.993$. Taigi, matome, kad nulinės hipotezės atmesti negalime.

Tą patį atsakymą gautume ir pritaikę F-testą.

Kitas klausimas kurį aptarkime yra toks. Jei optimali suma reklamai yra $40000 Lt$ o kaina $2 Lt$ ar tikėtina, kad pelnas bus $175000 Lt$? Nagrinėjamo modelio kontekste

$$\begin{aligned} E(tr_t) &= \beta_1 + \beta_2 p_t + \beta_3 a_t + \beta_4 a_t^2 = \\ &\beta_1 + \beta_2(2) + \beta_3(40) + \beta_4(40)^2 = 175. \end{aligned}$$

Ar suderintos tos dvi hipotezės. Patiktrinkime.

1. Nulinė hipotezė

$$H_0: \quad \beta_3 + 2\beta_4(40) = 1, \quad \beta_1 + 2\beta_2 + 40\beta_3 + 1600\beta_4 = 175$$

2. Alternatyva kad bent viena iš šių sąlygų nepatenkinta.

3. Testinė statistika $F = 1.75$.

4. Kritinė sritis $F_c = 3.120$ Kadangi $F < F_c$ turimi duomenys neprieštarauja nulinei hipotezei.

3.9 Prognozavimas

Prognozavimo bendruoju tiesiniu regresiniu modeliu ideologija nesiskiria nuo prognozavimo paprasčiausiu tiesiniu regresiniu modeliu. Tarkime, turime tiesinį regresinį modelį

$$Y_t = \boldsymbol{X}_t\boldsymbol{\beta} + \varepsilon_t, \quad t = 1, \ldots, T.$$

Turėdami duotuas paaiškinančiųjų dydžių reikšmes, sakykime, $\boldsymbol{X}_0$, tariame kad endogeninio kintamojo reikšmė Y_0 susijusi tuo pačiu modeliu, t.y.

$$Y_0 = \boldsymbol{X}_0\boldsymbol{\beta} + \varepsilon_0.$$

Be to, tariame, kad paklaidos $\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_T$ yra nekoreliuotos ir homoskedastinės. Tuomet įvertinę parametrus, randame taškinę prognozę

$$\widehat{Y}_0 = \boldsymbol{X}_0\widehat{\boldsymbol{\beta}}.$$

Prognozės paklaida yra $f = Y_0 - \widehat{Y}_0$. Galima suskaičiuoti paklaidos dispersiją $var(f)$ ir jos įvertinimą $\widehat{var}(f)$. Dydis $se(f)\sqrt{\widehat{var}(f)}$ vadinamas standartine prognozės paklaida.

4 Heteroskedastiškumas ir autokoreliacija

4.1 Heteroskedastiniai modeliai

Ekonometriniais metodais tirdami ekonominius reiškinius, retai kada išsiversime homoskedastiniais modeliais. Pavyzdžiui, nagrinėkime namų ūkių išlaidų maistui modelį

$$Y_t = \beta_2 X_t + \beta_1 + \varepsilon_t;$$

čia X – namų ūkio pajamos, Y – išlaidos maistui, β_2 – koeficientas aprašantis polinkį išleisti maistui, β_0 – būtinosios išlaidos maistui. Beveik akivaizdu, kad mažesnes pajamas turinčių namų ūkių išlaidos maistui bus panašesnės, nei dideles pajamas turinčiųjų. Tai galime paaiškinti didesnėmis galimybėmis tenkinti skonį, kuris šiaip jau yra visų žmonių skirtingas. Todėl išlaidų maistui modelyje anksčiau daryta prielaida $E\varepsilon_t^2 = \sigma^2$ su visais t nėra logiška. Tiksliau būtų tarti, kad dispersijos nėra vienodos, t.y. $E\varepsilon_t^2 = \sigma_t^2$ ir dispersijos σ_t^2 yra skirtingos skirtingiems t. Išlaikydami nekoreliuoto triukšmo prielaidą, turėsime, kad

$$cov(\boldsymbol{\varepsilon}) = diag(\sigma_1^2, \ldots, \sigma_n^2).$$

Jei regresinio modelio paklaidų dispersijos skiriasi, t.y. $var(\varepsilon_t) = \sigma_t^2$, tai sakoma, kad modelis yra heteroskedastinis.

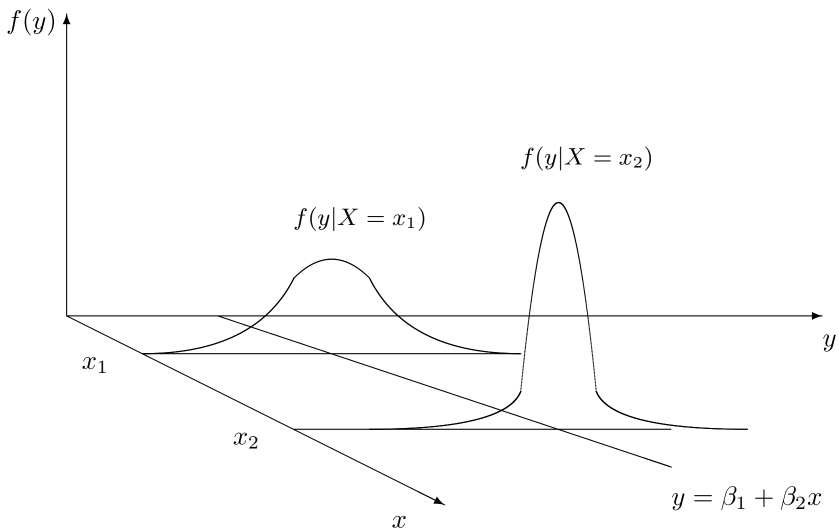
Tokie modeliai labai dažnai naudojami tiriant vietos aplinkybių duomenis. Pavyzdžiui, nagrinėjant kainų duomenis, įvairių firmų išlaidų bei pajamų duomenis, pirkimo kiekius įvairių tipų parduotuvėse.

Kokios galimos hetroskedastiškumo pasekmės taikant įprastinį mažiausių kvadratų metodą? Mažiausių kvadratų įvertinimai išlieka tiesiniai ir nepaslinkti, bet nebėra geriausias, t.y. jų variacija nebėra mažiausia tarp visų galimų tiesinių nepaslinktų įvertinimų.

Tikrai, pavyzdžiui,

$$var(\beta_2) = var(\sum w_t \varepsilon_t) = \sum_t w_t^2 \sigma_t^2.$$

Mažiausia dispersija gaunama, kai $\sigma_t^2 = \sigma^2$ su visais t.



Grafikas 4.1: heteroskedastiškumas

Standartinės paklaidos mažiausių kvadratų įvertinimams nebėra tikslūs. Tai atsiliepia pasikliautinumo intervalų tikslumui, ir atitinkamai, hipotezių patikrinimui.

Dažniausiai heteroskedastiškumas arba žinomas iš anksto, arba modeliuojamas. Pastaruoju atveju ieškoma geriausio būdo modeliui įvertinti.

Paprasčiausia heteroskedastiškumo forma yra taip vadinamas *grupinis heteroskedastiškumas*. Tarkime, duomenys gali būti sugrupuoti į dvigrupes: pirmoje T_1 duomenų ir jų dispersija yra vienoda – σ_1^2, o antroje - T_2 duomenų kurių dispersija lygi σ_2^2. Tokiam modeliui homoskedastiškumo testas suvedamas į hipotezę:

$$H_0 : \quad \sigma_1^2 = \sigma_2^2 = \sigma^2$$

su alternatyva

$$H_1 : \quad \sigma_1^2 > \sigma_2^2.$$

Nagrinėkime regresinius modelius atskirai su kiekviena duomanų grupe. Tarkime, gavome dispersijų įvertinimus, atitinkamai $\widehat{\sigma}_1^2$ ir $\widehat{\sigma}_2^2$. Intuityviai aišku, kad artimi įvertinimai liudija apie hipotezės napgrįstumą. Formaliai, prisiminkime, kad

$$(T_1 - d)\widehat{\sigma}_1^2/\sigma_1^2 \sim \chi^2_{T_1-d};$$
$$(T_2 - d)\widehat{\sigma}_2^2/\sigma_2^2 \sim \chi^2_{T_2-d}.$$

Kadangi statistikos suskaičiuotos iš skirtingų duomenų grupių, tai jos yra nepriklausomos. Vadinasi santykis

$$F = \frac{\widehat{\sigma}_1^2}{\widehat{\sigma}_2^2} \sim F(T_1 - d, T_2 - d).$$

Tai ir yra testinė statistika.

Grįžkime prie išlaidų maistui modelio:

$$Y_t = \beta_1 + \beta_2 X_t + \varepsilon_t.$$

Tikėtina, kad didesnes pajamas gaunančių namų ūkių išlaidos maistui labiau kintamos nei mažesnes pajamas gaunančių. Mat tuomet pradeda veikti skonio faktorius ir pan. Taigi, tikėtina, kad dispersija $var(\varepsilon_t) = \sigma_t^2$ priklauso nuo t. Tačiau, jei visos dispersijos bus skirtingos, tai turėsime T skirtingų parametrų ir jų negalėsime įvertinti. Tokiu atveju reikia dispersijas modeliuoti. Vienas iš labiausiai paplitusių modelių – proporcingumo. Tariame, kad

$$var(\varepsilon_t) = \sigma^2 X_t, \quad t = 1, \dots, T.$$

Išlaidų maistui modelyje tai reikštų, kad didesnės pajamos duoda didesnį išlaidų maistui išsibarstymą apie vidurkį. Bendresnis proporcingumo modelis būtų

$$var(\varepsilon_t) = f(X_t)\sigma^2, \quad t = 1, \dots, T,$$

kai f – kuri nors neneigiama funkcija. Tokie modeliai paprastai transformuojami į homoskedastinius. Tikrai, padalinę abi puses iš $f^{1/2}(X_t)$ gauname

$$\frac{Y_t}{f^{1/2}(X_t)} = \beta_1 \frac{1}{f^{1/2}(X_t)} + \beta_2 \frac{X_t}{f^{1/2}(X_t)} + \frac{\varepsilon_t}{f^{1/2}(X_t)}.$$

Pažymėję

$$Y_t^* = \frac{Y_t}{f^{1/2}(X_t)}, \; X_{1t}^* = \frac{1}{f^{1/2}(X_t)}, \quad X_{2t}^* = \frac{X_t}{f^{1/2}(X_t)}, \quad \varepsilon_t^* = \frac{Y_t}{f^{1/2}(X_t)},$$

gauname

$$Y_t^* = \beta_1 X_{1t}^* + \beta_2 X_{2t}^* + \varepsilon_t^*.$$

Gautas modelis yra homoskedastinis. Įvertinę modelį mažiausių kvadratų metodu gausime geriausius tiesinius nepaslinktus įvertinius.

4.2 Heteroskedastiškumo testai

Vienas iš būdų tiriamo modelio heteroskedastiškumui nustatyti yra euristiniai samprotavimai, paremti tam tikrais ekonominiais dėsniais. Kitas būdas – grafinis. Jei modelio triukšmo dispersijos yra nevienodos, tai dažniausiai galime pastebėti plokštumoje atidėję taškus $(\widehat{\varepsilon}_t, \widehat{y}_t)$. Paprastai matosi dispersijos didėjimo ar mažėjimo tendencija.

Heteroskedastiškumui nustatyti egzistuoja ir formalių metodų. Keletą iš jų aptarsime.

4.2.1 Goldfeldo–Quandto testas

Tarkime, turime stebėjimų duomenis

$$(Y_t, X_{2t}, \dots, X_{dt}), \quad t = 1, \dots, T.$$

Nagrinėkime tiesinį modelį

$$Y_t = \alpha + \beta X_t + \varepsilon_t, \quad t = 1, \dots, T$$

ir hipotezę

$$H_0 : E\varepsilon_1^2 = E\varepsilon_2^2 = \dots = E\varepsilon_n^2.$$

Šiai hipotezei patikrinti Goldfeldo – Kvandto (Goldfeld – Quandt) testas atliekame šiais žingsniais:

a) kintamuosius $X_1, X_2, \dots, X_T$ sutvarkome didėjimo tvarka: $X_1^* \le X_2^* \le \dots \le X_T^*$;

b) pašaliname p vidurinių iš sutvarkytos eilutės duomenų (rekomenduojama imti $p \sim T/4$);

c) abiems gautoms duomenų grupėms

$$(Y_1^*, X_1^*), \dots, (Y_s^*, X_s^*)$$

ir

$$(Y_{s+p}^*, X_{s+p}^*), \dots, (Y_n^*, X_n^*)$$

surandame atitinkamai dvi regresijos linijas, kurių paklaidų kvadratų sumos yra atitinkamai ESS_1 ir ESS_2. Čia žvaigždutės reiškia, kad duomenys sugrupuoti pagal jų sutvarkymą pirmajame žingsnyje;

Jei paklaidos (ε_t) yra nepriklausomi normaliniai atsitiktiniai dydžiai, tai statistika

$$f_T = \frac{ESS_2}{ESS_1}$$

turi Fišerio skirstinį su $((T-p-4)/2, (T-p-4)/2)$ laisvės laipsniais.

Hipotezę H_0 atmesime, jei $f_n > x_\alpha$, padarydami pirmosios rūšies klaidą su tikimybe α. Čia t_α yra randamas iš lygties

$$P(f_n > x_\alpha) = \alpha.$$

Šį Goldfeldo – Kvandto testą galime apibendrinti ir d egzogeninių kintamųjų atvejui. Tokiu atveju, duomenis reikia sutvarkyti pagal vieną iš nepriklausomų egzogeninių kintamųjų. Atitinkama statistika tada turės Fišerio skirstinį su $((n-p-2d)/2, (n-p-2d)/2)$ laisvės laipsniais.

Vienas pagrindinių šio kriterijaus trūkumų yra tas, kad kritinę sritį galime parinkti tik pagal pirmosios rūšies klaidos tikimybę. Vadinasi, kriterijaus galia nėra didelė.

4.2.2 Broišo-Pegano-Godfrėjaus testas

Nagrinėkime paprasčiausią regresinį modelį

$$Y_t = \beta_1 + \beta_2 X_t + \varepsilon_t. \tag{4.1}$$

Broišo-Pegano-Godfrėjaus (Breusch–Pegan–Godfrey) (BPG) kriterijus heteroskedastiškumui nustatyti remiasi prielaida, kad

$$\sigma_t^2 = \alpha_0 + \alpha_1 Z_t,$$

kai Z_t yra arba egzogeninis kintamasis arba kuris nors pagalbinis kintamasis, paaiškinantis paklaidų dispersijos kitimą. Kitaip tariant nulinė homoskedastiškumo hipotezė nagrinėjama atžvilgiu alternatyvos

$$H_A: \quad \sigma_t^2 = \alpha_0 + \alpha_1 Z_t.$$

Testas remiasi tokia regresinio tipo alternatyvos reprezentacija:

$$\begin{aligned} \varepsilon_t^2 &= \alpha_0 + \alpha_1 Z_t + (\varepsilon_t^2 - \sigma_t^2) = \\ &\alpha_0 + \alpha_1 Z_t + v_t, \quad t = 1, \dots, T. \end{aligned} \tag{4.2}$$

Čia $v_t = \varepsilon_t^2 - \sigma_t^2$ yra atsitiktiniai dydžiai su nuliniu vidurkiu $Ev_t = E\varepsilon_t^2 - \sigma_t^2 = 0$ su kiekvienu $t = 1, \ldots, T$. Dydis Z_t yra arba į modelį įtrauktas paaiškinamasis kintamasis X_t, arba kuris kitas kintamasis, paaiškinantis heteroskedastiškumą. Jei turėtume ε_t, tai (4.2) modelyje įvertinę parametrus α_0 ir α_1, galėtume patikrinti hipotezę $\alpha_1 = 0$. Kadangi dydžiai ε_t nėra stebimi, juos keičiame liekanomis $\widehat{\varepsilon}_t$. Taigi, norėdami patikrinti hipotezę H_0 prieš alternatyvą H_A atliekame šiuos žingsnius:

1. įvertinę modelį, surandame regresijos liekanas $\widehat{\varepsilon}_t$;
2. įvertiname
$$\widehat{\sigma}^2 = T^{-1} \sum_{t=1}^{T} \widehat{\varepsilon}_t^2;$$
3. įvertiname tiesinę regresiją
$$\frac{\widehat{\varepsilon}_t^2}{\widehat{\sigma}^2} = \alpha_0 + \alpha_1 Z_t + v_t$$
ir suskaičiuojame paaiškinamąją variacijos dalį RSS;
4. Testinė statistika yra $RSS/2$. Jei teisinga nulinė hipotezė ir modelis yra gausinis, tai žinoma, kad testinė statistika turi χ_1^2 skirstinį.

4.2.3 Harvėj testas

Nagrinėjamas modelis, kuriame triukšmo dispersijos yra
$$\sigma_t^2 = \exp\{\alpha_0 + \alpha_1 Z_t\}.$$
Tuomet
$$\log(\varepsilon_t^2) = \alpha_0 + \alpha_1 Z_t + \log(\varepsilon_t^2/\sigma_t^2) =$$
$$\alpha_0 + \alpha_1 Z_t + \nu_t.$$
Jei $\varepsilon_t \sim N(0, \sigma_t^2)$, tai
$$E\nu_t = E[\log(\chi_1^2)] = -1.2704$$
ir
$$var(\nu_t) = var[\log(\chi_1^2)] = 4.9348.$$
Toliau galime liekaną centruoti ir gauti regresinį modelį su nulinio vidurkio liekana. Tada ε_t keičiame liekanomis $\widehat{\varepsilon}_t$.

4.2.4 Spirmano testas

Randame regresijos liekanas ir sutvarkome jas mažėjimo tvarka: $\widehat{\varepsilon}_1^* > \widehat{\varepsilon}_2^* > \cdots > \widehat{\varepsilon}_T^*$. Be to, sutvarkome mažėjimo tvarka ir egzogeninius kintamuosius: $X_1^* > X_2^* > \cdots > X_T^*$.

Suskaičiuojame statistiką

$$r_{\varepsilon,X} = 1 - \frac{6\sum_{t=1}^{T} D_t^2}{T(T^2-1)};$$

čia D_t yra atitinkamų porų X ir $\widehat{\varepsilon}$ rangų (vietų numerių, kur jie atsidūrė po atitinkamų perstatų) skirtumas. Jei statistika $r_{\varepsilon,X}$ yra didelė, tai rodo heteroskedastiškumą. Kai paaiškinamųjų kintamųjų yra daugiau, reikia suskaičiuoti analogiškus dydžius su visais kintamaisiais.

4.2.5 Ekonominis pavyzdys

Nagrinėkime kviečių pasiūlą Australijoje, besivystančiame kviečių auginimo regione. Kviečių pasiūlos funkcija priklausys nuo naudojamos technologijos, nuo kviečių kainos (arba prognozuojamos kainos) ir oro sąlygų. Taigi bendru atveju kviečių pasiūlos funkcija būtų

$$Q = f(P, T, W)$$

Pasiūlos priklausomybė nuo kainos yra labai svarbi vyriausybei, reguliuojant kviečių supirkimo kainas ar subsidijas. Mat, jei vyriausybė ruošiasi fiksuoti supirkimo kainą, jai svarbu žinoti tikėtiną pasiūlą, kuri atitiktų siūlomą kainą.

Sudarykime ekonometrinį modelį:

$$q_t = \beta_1 + \beta_2 p_t + \beta_3 t + \varepsilon_t,$$

kai

q_t – išaugintų kviečių kiekis t-metais;

p_t – kviečių kaina t-metais (garantuota ar prognozuojama);

$t = 1, 2, \ldots, 26$ – metai ir kartu trendas, atspindintis kviečių auginimo technologijos pažangą;

ε_t – atsitiktinis faktorius, kuris, be kitų, atspindi ir oro įtaką kviečių derliui.

Kaip įprasta, $\beta_1, \beta_2, \beta_3$ yra modelio parametrai, kuriuos reikia įvertinti. Tam tikslui turime duomenis apie Australijos regionus (3 lentelė).

q	p	t	q	p	t
197.6	1.47	1	240.0	2.42	14
140.1	1.30	2	236.1	2.45	15
162.3	1.59	3	234.5	2.44	16
166.5	1.44	4	239.0	2.26	17
159.5	1.89	5	258.4	2.50	18
195.6	1.49	6	247.9	2.41	19
207.0	1.94	7	272.2	2.83	20
218.4	1.52	8	266.2	2.79	21
239.0	2.15	9	284.1	3.17	22
208.2	2.09	10	283.4	2.83	23
253.4	1.74	11	277.4	2.69	24
278.7	2.51	12	301.0	3.65	25
221.1	2.14	13	281.4	3.36	26

Kviečių kiekių ir atitinkamų kainų duomenys

Norėdami užbaigti modelio specifikavimą, turime padaryti prielaidas paklaidoms. Viena galimybė – atsitiktiniai dydžiai (ε_k) yra nepriklausomi ir vienodai pasiskirstę su nuliniu vidurkiu ir baigtine dispersija. Tačiau yra žinoma, kad po pirmų trylikos metų buvo išvestos kviečių veislės, kurios mažai reaguoja į oro permainas. Todėl tikėtina, kad po trylikos metų, kviečių derlius tapo stabilesnis. Turėdami omenyje, kad pagrindinis paklaidų faktorius yra oras, realistiškesnės būtų šios prielaidos modelio paklaidoms:

$$E\varepsilon_t = 0, \quad t = 1, \dots, 26; \tag{4.3}$$

$$var(\varepsilon_t) = \begin{cases} \sigma_1^2, & \text{kai } \ t = 1, \dots, 13 \\ \sigma_2^2, & \text{kai } \ t = 14, \dots, 26; \end{cases} \tag{4.4}$$

$$cov(\varepsilon_t, \varepsilon_s) = 0, \quad \text{kai } \ t \neq s. \tag{4.5}$$

Be to, galime tarti, kad $\sigma_1^2 > \sigma_2^2$. Aprašytas modelis atitinka grupinį heteroskedastiškumą. Suskaidę duomenis į dvi grupes, įvertiname

$$\widehat{\sigma}_1^2 = 641.64, \qquad \widehat{\sigma}_2^2 = 57.76$$

Gautas reikšmes panaudoję modelio transformacijai, įvertiname koeficientus ir išvedame regresiją

$$\begin{array}{rlll} \widehat{q}_t = & 138.1 + 21.72 p_t + & 3.283 t \\ & (12.7) \quad (8.81) & (0.812) \end{array}$$

Goldfeldo-Kvandto testinė statistika yra

$$GQ = 3.35$$

Su 5 procentų pasikliautinumu, kritinė reikšmė, atitinkanti $(18, 18)$ laisvės laipsnių, yra $F_c = 2.22$. Kadangi $GQ = 3.35 > F_c = 2.22$, tai nulinę hipotezę atmetame, tardami, kad paklaidų dispersija priklauso nuo pajamų lygio.

4.3 Autokoreliacija

Prielaida, kad skirtingų stebėjimų rezultatų paklaidos nekoreliuotos, dažniausiai nebėra teisinga stebint laiko eilutes. Tačiau ir regresiniuose modeliuose dažnai atsitinka, kad ta prielaida nėra pagrįsta.

Pirmiausia pastebėsime, kad paklaidų koreliacija neturi įtakos įvertinimų suderinamumui, bet turi įtakos tų įvertinimų efektyvumui.

Labai dažnai modelio paklaidos (ε_t) tenkina sąryšį

$$\varepsilon_t = f(\varepsilon_{t-1}) + v_t;$$

čia (v_t) yra nepriklausomi ir vienodai pasiskirstę atsitiktiniai dydžiai, o f kuri nors funkcija. Tokiu atveju sakoma, kad modelio paklaidos yra autokoreliuotos.

Dažniausiai pasitaikančios priežastys autokoreliuotumui yra šios:

1. Praleisti paaiškinamieji kintamieji (bloga modelio specifikacija). Dauguma ekonominių kintamųjų linkę autokoreliuoti. Jei toks kintamasis nebuvo įtrauktas į lygtį tai būtinai atsiras autokoreliacija.

2. Bloga modelio specifikacija. Pavyzdžiui, parinkome tiesinį modelį, nors iš tikro yra ciklinė priklausomybė.

3. Statistikiniai duomenys dažnai gaunami interpoliuojant. Tai susiję su praleistais duomenimis.

4. Net ir grynai atsitiktiniai faktoriai, pavyzdžiui, karai, štormai, drebėjimai, streikai ir pan., turės įtakos keliems laiko periodams (pavyzdžiui, gamybos procesui).

4.4 Autokoreliacijos testavimas

4.4.1 Durbino–Vatsono testas

Bene dažniausiai pasitaiko pirmos eilės autoregresinė tiesinio modelio paklaidų struktūra.

Paklaidų koreliuotumui tikrinti egzistuoja keletas testų. Vienas iš jų yra Durbino – Vatsono (Durbin – Watson) DW testas. Jis skirtas regresiniam modeliui

$$Y_t = \beta_1 + \sum_{j=2}^{d} \beta_j X_{jt} + \varepsilon_t,$$

kai

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t.$$

DW testas skirtas nulinei hipotezei

$$H_0 : \rho = 0$$

prieš vieną iš šių alternatyvų:

$$H_{a1} : \rho > 0; \quad H_{a2} : \ \rho < 0; \quad H_{a3} : \quad \rho \neq 0.$$

Testinė statistika yra

$$DW = \frac{\sum_{t=2}^{n}(\widehat{\varepsilon}_t - \widehat{\varepsilon}_{t-1})^2}{\sum_{t=1}^{n}\widehat{\varepsilon}_t^2},$$

kai $(\widehat{\varepsilon}_t)$ yra mažiausių kvadratų regresijos liekanos. Jei $DW = 2$, tai duomenys neprieštarauja hipotezei.

Jei $DW = 0$, tai hipotezę atmetame, kaip nepagrįstą.

Jei $DW = 4$, nulinę hipotezę taip pat atmetame.

Nagrinėkime

$$d = \frac{\sum_{t=1}^{T}(\varepsilon_t - \varepsilon_{t-1})^2}{\sum_{t=1}^{T}\varepsilon_t^2}.$$

Tarę, kad pakankamai dideliems T

$$\sum_{t=2}^{T}\varepsilon_t^2 \approx \sum_{t=2}^{T}\varepsilon_{t-1}^2 \approx \sum_{t=1}^{T}\varepsilon_i^2,$$

galime užrašyti

$$d \approx 2\Big(1 - \frac{\sum_{t=1}^{T}\varepsilon_t\varepsilon_{t-1}}{\sum_{t=1}^{T}\varepsilon_t^2}.\Big)$$

Dabar galime nesunkiai suprasti DW testo motivaciją. Tikrai, jei tarsime, kad ε_t ir ε_{t-1} yra dažniausiai to paties ženklo, tai esant teigiamai pirmosios eilės koreliacijai, sandaugos $\varepsilon_t\varepsilon_{t-1}$ bus teigiamos ir d bus ryškiai mažesnis už du. Jei ε_t it ε_{t-1} dažniau yra skirtingų ženklų, tai suma $\sum_t \varepsilon_t\varepsilon_{t-1}$ darysis neigiama ir d bus ryškiai didesnis už du.

4.5 Modelio vertinamas esant autokoreliacijai

Nagrinėkime modelį

$$Y_t = \beta_1 + \beta_2 X_{2t} + \cdots + \beta_d X_{dt} + \varepsilon_t, \quad t = 1, \ldots, T, \tag{4.6}$$

ir tarkime, kad paklaidos (ε_t) tenkina tokį sąryšį:

$$\varepsilon_t = \rho\varepsilon_{t-1} + \nu_t, t = 1, \ldots, T, \tag{4.7}$$

kai $\nu_1, \ldots, \nu_T$ yra nepriklausomi vienodai pasiskirstę $N(0, \sigma_\nu^2)$ atsitiktiniai dydžiai. Be to, tarkime, kad $\varepsilon_0 = 0$ ir, kad ν_t ir ε_{t-1} yra nepriklausomi. Triukšmas, aprašomas (**??**) lygtimi, vadinamas pirmos eilės autoregresiniu triukšmu.

Tarkime, $|\rho| < 1$. Nesunku suskaičiuoti, kad

$$\begin{aligned} \varepsilon_t &= \rho\varepsilon_{t-1} + \nu_t = \rho(\rho\varepsilon_{t-2} + \nu_{t-1}) + \nu_t = \cdots = \\ &= \nu_t + \rho\nu_{t-1} + \rho^2\nu_{t-2} + \cdots \end{aligned}$$

Kadangi $E\varepsilon_t = 0$, tai

$$\begin{aligned} E\varepsilon_t^2 &= E\nu_t^2 + \rho^2 E\nu_{t-1}^2 + \rho^4 E\nu_{t-2}^2 + \cdots = \\ &= (1 + \rho^2 + \rho^4 + \cdots)\sigma_\nu^2 = \\ &\frac{\sigma_\nu^2}{1 - \rho^2}, \end{aligned}$$

o

$$\begin{gathered} cov(\varepsilon_t, \varepsilon_{t-1}) = \\ E(\nu_t + \rho\nu_{t-1} + \rho^2\nu_{t-2} + \cdots)(\nu_{t-1} + \rho\nu_{t-2} + \rho^2\nu_{t-3} + \cdots) = \\ E(\nu_t + \rho(\nu_{t-1} + \rho\nu_{t-2} + \rho^2\nu_{t-3} + \cdots))(\nu_{t-1} + \rho\nu_{t-2} + \rho^2\nu_{t-3} + \cdots) = \\ = \rho E(\nu_{t-1} + \rho\nu_{t-2} + \rho^2\nu_{t-3} + \cdots)^2 = \\ = \frac{\rho\sigma_\nu^2}{1 - \rho^2}. \end{gathered}$$

Bendru atveju gausime

$$E\varepsilon_t\varepsilon_{t-s} = \frac{\rho^s\sigma_\nu^2}{1 - \rho^2}.$$

Taigi

$$cov(\boldsymbol{\varepsilon}) = E\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' = \sigma_{\boldsymbol{\varepsilon}}^2 \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{T-1} \\ \rho & 1 & \rho & \dots & \rho^{T-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \dots & 1 \end{pmatrix}.$$

Jei ρ žinomas, tai galime, panaudoję paprastas procedūras, modelį suvesti į nekoreliuotų paklaidų modelį. Kadangi

$$\begin{aligned} Y_t \quad &= \beta_1 + \beta_2 X_{2t} + \dots + \beta_d X_{dt} + \varepsilon_t \\ Y_{t-1} \quad &= \beta_1 + \beta_2 X_{2(t-1)} + \dots + \beta_d X_{d(t-1)} + \varepsilon_{t-1}, \end{aligned}$$

tai antrąją lygtį padauginę iš ρ ir atėmę iš pirmosios, gausime

$$Y_t^* = \beta_1(1-\rho) + \beta_2 X_{2t}^* + \dots + \beta_d X_{dt}^* + \nu_t, \tag{4.8}$$

kai

$$Y_t^* = Y_t - \rho Y_{t-1}, \quad X_{jt}^* = X_{jt} - \rho X_{j,t-1}, \quad j = 1, \dots, d.$$

Ypač įdomus yra atvejas, kai $\rho = 1$. Tuomet gautasis modelis (4.8) vadinamas pradinio modelio *pirmaja išvestine*.

Tačiau dažniausiai ρ nėra žinomas. Tuomet naudojame jo įvertinimą, kurį galime rasti, pavyzdžiui, pritaikę tokią procedūrą.

- Randame regresijos liekanas $\widehat{\varepsilon}_t, t = 1, \dots, T$ ir apskaičiuojame

$$\widehat{\rho}_t = \frac{\sum_{t=1}^T \widehat{\varepsilon}_t \widehat{\varepsilon}_{t-1}}{\sum_{t=1}^T \widehat{\varepsilon}_{t-1}^2}.$$

- Su šia $\widehat{\rho}$ reikšme atliekame apibendrintą diferencijavimą:

$$Y_t^* = \beta_1(1-\widehat{\rho}) + \beta_2 X_{2i}^* + \dots + \beta_d X_{dt}^* + \nu_t, \tag{4.9}$$

kai

$$\begin{gathered} Y_t^* = Y_t - \widehat{\rho} Y_{t-1}, \\ X_{2t}^* = X_{2t} - \widehat{\rho} X_{2,t-1}, \dots, \boldsymbol{X}_{kt}^* \\ = X_{dt} - \widehat{\rho} X_{d,t-1}, \quad t = 1, \dots, T. \end{gathered}$$

- Surandame regresijos (4.9) koeficientų β_i įvertinimus $\widehat{\beta}_i$ ir apskaičiuojame liekanas

$$\widehat{\widehat{\varepsilon}_t} = Y_t - \widehat{\beta}_1 - \widehat{\beta}_2 X_{2t} - \cdots - \widehat{\beta}_d X_{dt},$$

$t = 1, \ldots, T$.

- Autoregresinę lygtį (4.8) keičiame lygtimi

$$\widehat{\widehat{\varepsilon}_t} = \rho \widehat{\widehat{\varepsilon}_{t-1}} + \nu_t$$

ir kartojame pirmą žingsnį.

- Šį procesą tęsiame tol, kol naujoji ρ reikšmė nuo ankstesnės skirsis ne daugiau kaip 0.01 arba 10 – 20 kartų.

4.6 PROGNOZAVIMAS ESANT AUTOKORELIACIJAI

Ankstesniuose skireliuose matėme, kad klasikiniu regresiniu modeliu prognozuojame paprasčiausiu interpoliavimo metodu. Jei modelis yra

$$Y_t = \beta_1 + \beta_2 X_t + \varepsilon_t, \quad t = 1, \ldots, T,$$

tai prognozavimas remiasi prielaida, kad kiekvienos naujos stebėjimų poros (X_{T+1}, Y_{T+1}) generavimo mechanizmas yra toks pat. T.y.

$$Y_{T+1} = \beta_1 + \beta_2 X_{T+1} + \varepsilon_{T+1}.$$

Jei modelio paklaidos nekoreliuotos, tai geriausias Y_{T+1} reikšmės įvertinys būtų

$$\widehat{Y}_{T+1} = \widehat{\beta}_1 + \widehat{\beta}_2 X_{T+1}, \tag{4.10}$$

kai $\widehat{\beta}_1$ ir $\widehat{\beta}_2$ yra parametrų atitinkamai β_1 ir β_2 mažiausių kvadratų įvertiniai. Tačiau, jei paklaidos autokoreliuoja, tai $\widehat{\beta}_1$ ir $\widehat{\beta}_2$ nebėra patys geriausi įvertiniai. Todėl ir (4.10) formule apibrėžtas Y_{T+1} įvertinys $\widehat{Y}_{T+1}$ nebėra geriausias. Aišku, modelio parametrus galima įvertinti atsižvelgiant į paklaidų autokoreliaciją. Tačiau tai dar ne viskas. Kai autokoreliacijos nėra, geriausias paklaidos ε_{T+1} įvertinimas yra jo vidurkis, kuris lygus nuliui. Kai paklaidos autokoreliuoja, tai $\varepsilon_{T+1} = \rho\varepsilon_T + \nu_T$. Paklaidos sisteminę komponentę $\rho\varepsilon_T$ galime prognozuoti dydžiu $\widehat{\rho}\widehat{\varepsilon}_T$, kai $\widehat{\rho}$ yra parametro ρ įvertinys, o

$$\widehat{\varepsilon}_T = Y_t - \widehat{\beta}_1 - \widehat{\beta}_2 X_T.$$

Taigi geriausia Y_{T+1} prognozė, kai modelio paklaidos autokoreliuoja, yra

$$\widehat{Y}_{T+1} = \widehat{\beta}_1 + \widehat{\beta}_2 X_{T+1} + \widehat{\rho}\widehat{\varepsilon}_T.$$

4.7 Ekonominis pavyzdys

Dažniausiai autokoreliuotus modelius tenka taikyti sprendžiant uždavinius susijusius su žemės ūkiu. Kaip pavyzdį, panagrinėkime žemės ūkio produktui panaudoto žemės ploto ir to produkto pajamingumo sąryšį. Tegu A nagrinėjama žemės ūkio kultūra užimtą plotą, P – gautos pajamos. Nagrinėkime pastovaus elastiškumo log-log modelį

$$log(A) = \beta_1 + \beta_2 \log(P) + \varepsilon.$$

Turimi duomenys aprašo cukraus pasėlius Bangladeše (Lietuviškų duomenų kol kas nėra). Turime 34 metinius užimto ploto ir cukraus kainos stebėjimus.

Pažymėję $Y_t = \log(A_t)$, $X_t = \log(P_t)$, sudarykime modelį

$$Y_t = \beta_1 + \beta_2 X_t + \varepsilon_t.$$

Kol kas nedarykime jokių prielaidų modelio paklaidoms, o tiesio įvertinkime parametrus mažiausių kvadratų metodu. Gausimme tokį rezultatą:

$$\begin{array}{rlll} \widehat{Y}_t = & 6.111 + & 0.971 X_t, & R^2 = 0.706 \\ & (0.169) & (0.111) & \text{(stand.pakl.)} \end{array}$$

Plotas	Cukranendrių kaina	Plotas	Cukranendrių kaina
A_t	P_t	A_t	P_t
29	0.075258	91	0.205394
71	0.114894	121	0.267396
42	0.101075	162	0.230411
90	0.110309	143	0.368771
72	0.109562	138	0.285076
57	0.132486	230	0.360332
44	0.141783	128	0.322976
61	0.209559	87	0.301266
60	0.188259	124	0.287834
70	0.195946	97	0.401437
88	0.226087	152	0.404692
80	0.145585	197	0.353188
125	0.194030	220	0.410233
232	0.270362	171	0.360418
125	0.235821	208	0.463087
99	0.220826	237	0.401582
250	0.380952	235	0.391660

Duomenys apie cukranendrėmis užimtą plotą ir cukranendrių kainą

Rezultatas rodo, kad abu parametrai ženkliai skiriasi nuo nulio, o pasėlių ploto elastiškumas kainai yra beveik vienetas. Tačiau pagalvokime, ar galime taip lengvai nekreipti dėmesio į paklaidų savybes: heteroskedastiškumą ar autokoreliaciją? Akivaizdu, kad fermeriai planuoja cukranendrių pasėlių plotus atsižvelgdami į savo suvokimą apie būsimą cukraus kainą bei vyriausybės muitų politiką cukraus atžvilgiu. Modelyje nei vienas iš šių kintamųjų nėra įtrauktas. Taigi juos atspindi modelio paklaidos. Kadangi fermerių suvokimas kinta gana lėtai, tikėtina, kad paklaidos bus koreliuotos. Aišku, galime iškart pasitelkti į pagalbą Durbino-Vatsono testą ir patikrinti paklaidų autokoreliaciją. Bet skubėti nepatartina. Pirmiausia neaišku kokia koreliacija galėtų būti apskritai. Taigi pirmiausia patariama pasižiūrėti į regresijos liekanas $\widehat{\varepsilon}_t$, kurios, tam tikra prasme, atspindi paklaid ε_t struktūrą. Žemiau esančioje letelėje pateitos visų 34 liekanų reikšmės. Iš jos gana aiškiai matyti neigiamos koreliacijos tendencija. Mat liekanų ženklai grupuojasi tai į teigiamus, tai į neigiamus. Gretimų liekanų ženklai dažniausiai yra vienodi. Tai yra teigiamos koreliacijos požymis.

Laikas	$\widehat{\varepsilon}_t$	Laikas	$\widehat{\varepsilon}_t$	Laikas	$\widehat{\varepsilon}_t$	Laikas	$\widehat{\varepsilon}_t$
1	-0.233	10	-0.281	19	-0.035	28	-0.209
2	0.251	11	-0.191	20	0.401	29	0.182
3	-0.149	12	0.141	21	-0.180	30	0.147
4	0.528	13	0.308	22	0.034	31	0.021
5	0.312	14	0.605	23	0.317	32	-0.027
6	-0.106	15	0.119	24	-0.162	33	0.242
7	-0.431	16	-0.050	25	-0.481	34	0.258
8	-0.484	17	0.347	26	-0.082		
9	-0.396	18	-0.064	27	-0.651		

Modelio liekanos cukranendrių pavyzdyje

Norėdami patikrinti autokoreliacijos hipotezę, pasinaudokime Durbino-Vatsono kriterijų. Cukranandrių pavyzdyje Durbino-Vatsono statistika yra $d = 1.291$. Ar šis skaičius ženkliai mažesnis už du ar didesnis už nulį? Tikriausiai. Bet būsime tikresni suskaičiavę atitinkamą p-reikšmę:

$$p - \text{reikšmė} = P(d \leq 1.291) = 0.0098.$$

Ši reikšmė mažesnė už priimtiną reikšmingumo lygmenį 0.05. Taigi išvada: modelio paklaidos yra teigiamai autokoreliuotos.

Taigi pataisyti modelį galime padarę prielaidą, kad paklaidos yra autokoreliuotos. T.y.

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

kai (η_t) nepriklausomi ir vienodai pasiskirstę atsitiktiniai dydžiai. Norėdami įvertinti parametrus cukranendrių modelyje, reikia žinoti ρ. Jį galime įvertinti pasinaudoję turimomis liekanpomis $(\widehat{\varepsilon}_t)$ taip:

$$\widehat{\rho} = \frac{\sum_{t=2}^{T} \widehat{\varepsilon}_t \widehat{\varepsilon}_{t-1}}{\sum_{t=2}^{T} \widehat{\varepsilon}_{t-1}} = 0.342.$$

Kitas žingsnis – transformuoti duomenis ir iš jų įvertinti parametrus.Visa tai atlikę, gauname

$$\begin{array}{rcll} \log\left(\widehat{A}_t\right) & = & 6.164 + & 1.007\log(P_t) \\ & & (0.213) & (0.137) \end{array}$$

Suskaičiuokime cukranendrių pasėlių ploto prognozę atitinkančią kainą 0.4. Kadangi

$$\widehat{\beta}_1 = 6.1641, \quad \widehat{\beta}_2 = 1.0066, \quad \widehat{\rho} = 0.342,$$

tai

$$\begin{aligned} &\widehat{\varepsilon}_T = Y_T - \widehat{\beta}_1 - \widehat{\beta}_2 X_T \\ &= \log(A_T) - \widehat{\beta}_1 - \widehat{\beta}_2 \log(P_T) \\ &= 5.4596 - 6.1641 - 1.0066(-0.9374) = 0.239. \end{aligned}$$

Dabar jau galime suskaičiuoti ir reikalingą prognozę:

$$\begin{aligned} &\widehat{Y}_{T+1} = \widehat{\beta}_1 + \widehat{\beta}_2 X_{T+1} + \widehat{\rho}\widehat{\varepsilon}_T \\ &= 6.1641 + 1.0066\log(0.4) + 0.342 \cdot 0.239 = 5.3235. \end{aligned}$$

Atkreipiame dėmesį, kad gauta ne pasėlių ploto prognozė, bet jo logaritmo. Perskaičiavus į plotą, gauname $\widehat{A}_{T+1} = 205$.

5 Bendresni regresiniai modeliai

5.1 Fiktyvių kintamųjų panaudojimas

Klasikiniame tiesiniame regresiniame modelyje

$$Y_t = \beta_1 + \beta_2 X_{1t} + \cdots + \beta_d X_{dt} + \varepsilon_t$$

išreikštiniai parametrai $\beta_1, \beta_2, \ldots, \beta_d$ laikomi nekintamais. Kadangi

$$\beta_k = \text{ vidurkio } \quad EY_t \quad \text{pasikeitimas, kai} \quad X_{kt} \quad \text{pakinta vienu vienetu,}$$
$$\text{o kiti kintamieji lieka pastovūs} = \frac{\partial EY_t}{\partial X_{kt}},$$

tai dydžio X_{kt} pasikeitimas turi tą pačią įtaką visiems stebėjimams. Tačiau labai dažnai nutinka, kad skirtingais laikotarpiais X_{kt} įtakoja endogeninį kintamąjį skirtingai. Tačiau to neaprašo klasikinis tiesinis regresinis modelis. Šiame skyrelyje aptarsime galimus klasikinio tiesinio modelio apibendrinimus.

Fiktyvus kintamasis yra toks paaiškinantysis kintamasis, kuris įgyja tik dvi reikšmes: 0 arba 1. Paprastai, tokie kintamieji naudojami norint skaitmeniškai išreikšti kokybinius rodiklius.

Kad būtų aiškiau, nagrinėkime sąryšį tarp JAV pajamų iš išlaidų 1929–1970 metų laikotarpyje. Tegu C_t reiškia suvartojimą vienam gyventojui t metais, o Y_t – realiosios pajamos vienam gyventojui t metais. Nagrinėkime regresinį modelį

$$C_t = \beta_1 + \beta_2 Y_t + \varepsilon_t, \quad t = 1929, \ldots, 1970.$$

Akivaizdu, kad toks modelis nebus tikslus, nes nagrinėjamas laiko periodas apima ir antąjį pasulinį karą. Gerai žinoma, kad krizių laikotarpiais vartojimas paprastai ribojamas. Todėl kuris nors iš parametrų turėtų reaguoti į tai. Vienas iš būdų atsižvelgti į kokybinę charakteristiką (šiuo atveju į karo laikotarpį) – panaudoti fiktyvų (dichotominį) kintamąjį. Fiktyvus kintamasis D apibrėžiamas taip:

$$D = \begin{cases} 1, & \text{kai teisinga kokybinė charakteristika} \\ 0, & \text{kai kokybinės charakteristikos nėra.} \end{cases}$$

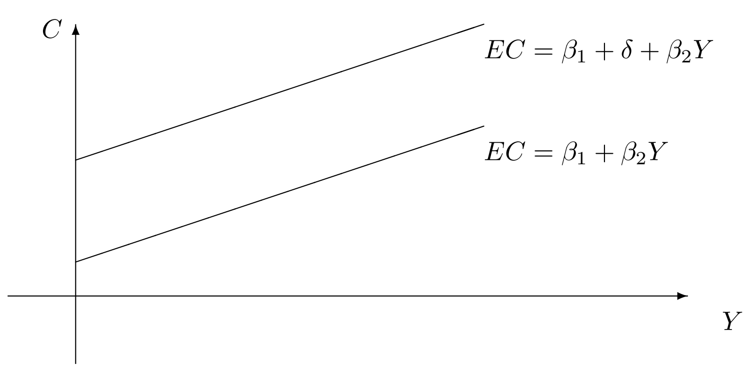
Nordami šią idėją pritaikyti nagrinėjamam pavyzdžiui, išskirkime karo periodo 1941–1946 metus ir apibrėžkime

$$D_t = \begin{cases} 1, & \text{kai } \ t = 1941, \dots, 1946; \\ 0, & \text{kitais atvejais.} \end{cases}$$

Toliau nagrinėkime apibendrintą modelį:

$$C_t = \beta_1 + \beta_2 Y_t + \delta D_t + \varepsilon_t, \quad t = 1929, \dots, 1970.$$

Šiuo modeliu aprašome besikeičiantį laisvąjį narį, palikdami polinkio parametrą nekintamu.



Grafikas 5.1: fiktyvus kintamasis keičia laisvąjį narį

Metai	C	Y	Metai	C	Y
1929	1145	1236	1950	1520	1646
1930	1059	1128	1951	1509	1657
1931	1016	1077	1952	1525	1678
1932	919	921	1953	1572	1726
1933	897	893	1954	1575	1714
1934	934	952	1955	1659	1795
1935	985	1035	1956	1673	1839
1936	1080	1158	1957	1683	1844
1937	1110	1187	1958	1666	1831
1938	1097	1105	1959	1735	1881
1939	1131	1190	1960	1749	1883
1940	1178	1259	1961	1755	1909
1941	1240	1427	1962	1813	1968
1942	1197	1582	1963	1865	2013
1943	1213	1629	1964	1945	2123
1944	1238	1673	1965	2044	2235
1945	1308	1642	1966	2123	2331
1946	1439	1606	1967	2160	2398
1947	1431	1513	1968	2248	2480
1948	1438	1567	1969	2301	2517
1949	1451	1547	1970	2323	2579

Išlaidos ir pajamos vienam žmogui

Turimi duomenys duoda tokį rezultatą:

$$\begin{array}{rccc} \widehat{C}_t = & 101.36- & 204.95D_t+ & 0.86Y_t \\ & (3.98) & (-10.91) & (58.73) \end{array}$$

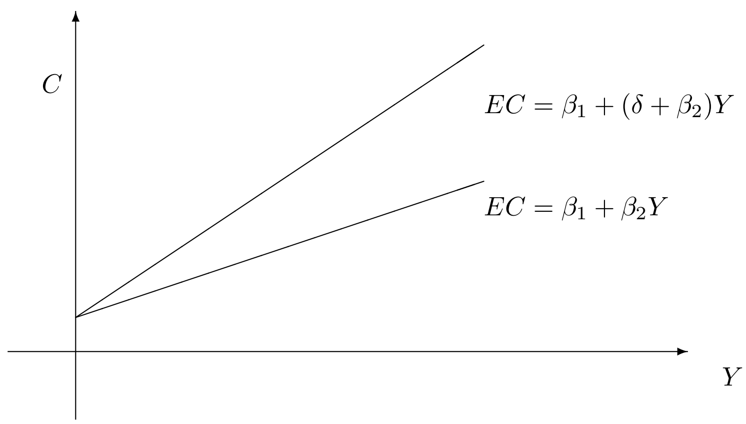
Skliaustuose nurodytos atitinkamos t-testo reikšmės.

Fiktyvų kintamąjį galime panaudoti taip pat ir polinkio kitimui aprašyti. Pavyzdžiui, galime nagrinėti modelį

$$C_t = \beta_1 + \beta_2 Y_t + \gamma(Y_t D_t) + \varepsilon_t.$$

Galimas ir bendras atvejis, apjungiantis tiek fiktyvų laisvą narį, tiek polinkio parametrą:

$$C_t = \beta_1 + \delta D_t + \beta_2 Y_t + \gamma Y_t D_t + \varepsilon_t.$$



Grafikas 5.2: fiktyvus kintamasis keičia polinkį

Yra daugybė pavyzdžių, kai fiktyvaus kintamojo pagalba reikia aprašyti arba kintamą laisvą narį arba polinkio parametrą arba abu iškart.

1. Nagrinėjame kurios nors profesijos atlyginimus W. Paaiškinamieji kintamieji turėtų būti: patirtis (EXP), kvalifikacija ($QUAL$), pareigų atlikimo matas ($OUTPUT$). Lytis taip pat gali turėti įtakos atlyginimui. Norėdami į tai atsižvelgti modelyje, įveskime fiktyvų kintamąjį D, kuris priima reikšmę 0, jei darbuotojas yra vyras ir 1, jei moteris. Tuomet galime nagrinėti tiesinį modelį

$$W = \beta_1 + \delta D + \beta_2 EXP + \beta_3 OUTPUT + \beta_4 QUAL + \varepsilon.$$

   Parametras δ aprašo atlyginimų skirtumą tarp vyro ir moters, kai jau atsižvelgta į patirtį, kvalifikaciją ir sugebėjimus. Rezultatas $\delta \neq 0$ parodytų, kad egzistuoja atlyginimų diskriminacija.

2. Tarkime, tiriame, kaip statistikos magistrų atlyginymas (SAL) priklauso nuo kvalifikacijos (GPA), lyties, ir programos, kurią baigė ar studijuoja magistrantas: statistiką ar ekonometriją. Pastaruosius du kintamuosius aprašysime fiktyviais dydžiais D – priklausomumą nuo lyties ir ($METRICS$) – priklausomumą nuo baigtos programos. Regresinis modelis galėtų būti toks:

$$SAL = \beta_1 + \delta_1 D + \delta_2 METRICS + \beta_2 GPA + \varepsilon.$$

3. Nekilnojamo turto agentūros ekonomistai nagrinėja faktorius įtakojančius namo kainą ($PRICE$). Pagrindiniai paaiškinamieji kintamieji yra: namo plotas ($SIZE$) ir jo amžius (AGE). Be to, domina visa eilė kokybinių namo charakteristikų, kurios galėtų daryti esminę įtaką kainai. Pavyzdžiui, kai kurie namai turi baseinus (galime nagrinėti atitinkamą fiktyvų kintamąjį ($POOL$)). Kainai gali daryti esminę įtaką gali daryti vietovė ($HOOD$). Pavyzdžiui, galime nagrinėti modelį

$$PRICE = \beta_1 + \beta_2 SIZE + \beta_3 LOT + \beta_4 AGE + \delta_1 POOL + \delta_2 HOOD + \varepsilon_t.$$

4. Fiktyvūs kintamieji gali būti panaudoti nustatant regioninę įtaką ekonomikos raidai. Pavyzdžiui, tarkime, tiriame vartojimo priklausomybės nuo pajamų įpatumus skirtingiems regionams. Galime negrinėti tokius fiktyvius kintamuosius: Z – žemaitija; S – suvalkija, D – dzūkija ir A – aukštaitija. Tačiau visi keturi fiktyvūs kintamieji vienu metu įeiti į modelį negali. Mat
$$Z + S + D + A = 1.$$
Vadinasi, tie kintamieji yra tiesiškai surišti. Todėl kurio nors vieno iš jų reikia į lygtį neįtraukti. Modelis galėtų būti toks:
$$V = \beta_1 + \beta_2 I + \delta_1 Z + \delta_2 D + \delta_3 S + \varepsilon.$$

Norint nustatyti ar kokybinis kintamasis daro esminę įtaką, reikia atlikti koeficiento reikšmingumo testus. Pavyzdžiui, nagrinėkime vartojimo modelį

$$C_t = \beta_1 + \delta D_t + \beta_2 Y_t + \varepsilon_t, \quad t = 1929, \ldots, 1970.$$

Norėdami patikrinti, ar karo laikotarpis turėjo esminės įtakos vartojimui, galime pasinaudoti t testu. Jei $\widehat{\delta}$ – parametro δ mažiausių kvadratų įvertis, tai
$$\tau = \frac{\widehat{\delta} - \delta}{se(\widehat{\delta})} \quad \sim \quad t_{T-K},$$
kai modelio paklaidoms teisinga gausiškumo prielaida. Norėdami patikrinti karo laikotarpio įtaką, turime nustatyti koeficiento δ reikšmingumą.

1. $H_0 : \quad \delta = 0;$

2. $H_1 : \quad \delta \neq 0;$

3. Testinė statistika
$$\tau = \frac{\widehat{\delta}}{se(\widehat{\delta})},$$
kai teisinga nulinė hipitezė, turi stjudento skirstinį;

4. Parinkę reikšmingumo lygmenį $\alpha = 0.01$, nulinę hipotezę atmesime, jei $|\tau| \geq 2.708$;

5. Testinės statistikos reikšmė yra
$$\tau = \frac{-204.95}{18.79} = -10.91.$$
Kadangi $|-10.91| \geq 2.708$ nulinę hipotezę atmetame ir darome išvadą, kad karas turėjo reikšmingos įtakos vartojimui.

5.2 TIKIMYBINIO PASIRINKIMO MODELIAI

Labai dažnai tenka spręsti dilemą: arba-arba. Pavyzdžiui, stoti į universitetą ar ne? Suteikti kreditą ar ne? Pirkti namą ar nepirkti? Ekonometristai bando išsiaiškinti, kokios priežastys lemia vieną ar kitą sprendimą. Tam tikslui išvystyti diskretaus pasirinkimo modeliai. Pabandykime juos išsiaiškinti paprastu pavyzdžiu. Kaip paaiškinti individo pasirinkimą vykti į darbą visuomeniniu transportu ar nuosavu automobiliu? Norėdami sukiekybinti pasirinkimą, apibrėžkime fiktyvų kintamąjį:

$$Y = \begin{cases} 1, & \text{jei pasirenkamas nuosavas automobilis} \\ 0, & \text{jei pasirenkamas visuomeninis transportas.} \end{cases} \tag{5.1}$$

Dydis Y yra atsitiktinis ir diskretus. Kadangi galimos reikšmės yra tik dvi, dydis Y pilnai aprašomas tikimybe

$$p = P(Y = 1).$$

Tuomet $1 - p = P(Y = 0)$. Vidurkis $EY = p$. Taigi, išties reikia įvertinti tikimybę p. Kokie faktoriai lemia jo reikšmę? Vienas faktorių gali būti laiko tarp kelionės nuosavu ir visuomeniniu transportu skirtumas. Apibrėžkime

$$X = t_1 - t_2,$$

kai t_1 reiškia laiką, trunkantį nuvykti į darbą autobusu, t_2 – nuosavu automobiliu. *A priori* tikėtina, kad didėjant X, didėja ir tikimybė rinktis nuosavą transportą. Taigi tikėtinas teigiamas sąryšis tarp tikimybės p ir laiko skirtumo X. Kaip pamename, regresija atsitiktinį dydį išreiškia vidurkio ir atsitiktinės paklaidos suma:

$$Y = E(Y) + \varepsilon.$$

Tardami, kad vidurkis EY ir X surišti tiesiniu sąryšiu, turėtume

$$EY = p = \beta_1 + \beta_2 X.$$

Tuo pačiu ir tiesinę regresiją

$$Y = \beta_1 + \beta_2 X + \varepsilon.$$

Tačiau toks modelis būtų visai nepriimtinas. Viena iš priežasčių būtų heteroskedastiškumas. Kitas trūkumas tas, kad įvertinę parametrus, gautas pasirinkimo Y įvertinys būtų

$$\widehat{Y} = \widehat{\beta}_1 + \widehat{\beta}_2.$$

Tačiau tas dydis gali įgyti tiek neigiamas reikšmes, tiek reikšmes didesnes už vieną.

Tam kad pasirinkimo tikimybė p išliktų intervale $[0, 1]$, naudojama netiesinė priklausomybė tarp p ir X. Būtent, tarkime F yra kokia nors pasiskirstymo funkcija. Tuomet tikimybė p modeliuojama sąryšiu

$$p = F(\beta_1 + \beta_2 X).$$

Jei F standartinio normaliojo dydžio pasiskirstymo funkcija, tai gautas modelis vadinamas “probit”, jei eksponentinio – “logit”.

Kaip įvertinti tikimybinį pasirinkimo modelį? Dažniausiai naudojamas maksimalaus tikėtinumo metodas. Tarkime, atsitiktinai atrinkome tris piliečius ir nustatėme, kad du pirmieji važiuoja į darbą autobusu, o trečiasis – nuosavu automobiliu. Taigi $Y_1 = 1, Y_2 = 1, Y_3 = 0$. Be to, nustatėme, kad tuos tris piliečius atatinkančios X reikšmės yra $X_1 = 15, X_2 = 20, X_3 = 5$. Kokia yra tikimybė, kad $Y_1 = 1, Y_2 = 1, Y_3 = 0$? Tardami, kad tie dydžiai yra nepriklausomi, turime

$$P(Y_1 = 1, Y_2 = 1, Y_3 = 0) = P(Y_1 = 1)P(Y_2 = 1)P(Y_3 = 0) =$$
$$F(\beta_1 + 15\beta_2)F(\beta_1 + 20\beta_2)[1 - F(\beta_1 + 5\beta_2)]$$

Gauta funkcija statistikoje vadinama tikėtinumo funkcija. Parametrai β_1 ir β_2 įvertinami suradus tas jų reikšmes, su kuriomis tikėtinumo funkcija yra didžiausia. Toks parametrų vertinimo metodas ir vadinamas didžiausio tikėtinumo metodu.

Norėdami rasti modelio interpretavimą, išnagrinėkime kokį poveikį įvykio $Y=1$ tikimybei padaro kintamojo X pasikeitimas vienu vienetu. Tą pasikeitimą parodo išvestinė

$$\frac{dp}{dX} = f(\beta_1 + \beta_2 X)\beta_2.$$

Taigi poveikis priklauso nuo skirstinio F tankio funkcijos f.

1. Kadangi f yra tankio funkcija, tai $f(\beta_1+\beta_2 X)$ visados teigiamas dydis. Vadinasi poveikio dp/dX ženklas sutampa su β_2 ženklu. Transporto pasirinkimo uždavinyje tikėtina, kad $\beta_2 > 0$. Taigi ir $dp/dX > 0$.

2. Kai X kinta, keičiasi funkcijos $f(\beta_1 + \beta_2 X)$ reikšmės. Standartinio normaliojo dydžio tankis savo didžiausią reikšmę pasiekia koordinačių pradžios taške. Todėl $f(\beta_1 + \beta_2 X)$ didžiausią reikšmę pasiekia taške $\beta_1 + \beta_2 X = 0$. Šiuo atveju $p = F(0) = 0.5$. Tai atitinka vienodą galimybę, tiek rinktis nuosavą transportą tiek visuomeninį. Tuo pačiu, X pasikeitimas čia gali turėti didžiausią įtaką, nes pilietis yra „ribinėje padėtyje" ir gali pakrypti tiek į vieną tiek į kitą pusę.

3. Jei $\beta_1 + \beta_2 X$ reikšmė yra didelė, sakykime, lygi 3, tai tikimybė, kad pilietis pasirinks nuosavą transportą yra labai didelė, artima 1. Tokiu atveju, X pasikeitimas vargu ar padarys kiek žymesnę įtaką apsisprendimui, mat $f(\beta_1 + \beta_2 X)$ šiuo atveju artimas nuliui. Analogiški samprotavimai taikomi ir kitam kraštiniam atvejui, kai $\beta_1 + \beta_2 X$ yra neigiamas, sakykime, mažesnis už -3.

Probit modelio rezultatus galime pritaikyti prognozuodami pasirinkimą. Galimybė prognozuoti diskretų pasirinkimą yra labai svarbi daugelyje ekonomikos sričių. Pavyzdžiui, bankui labai svarbu įvertinti tikimybę, kad klientas yra mokus (grąžins ar negrąžins paskolą). Pritaikę maksimalaus tikėtinumo metodą ir įvertinę parametrus, randame tikimybės įvertinį:

$$\widehat{p} = F(\widehat{\beta}_1 + \widehat{\beta}_2 X).$$

Pasirinkimą vertname pagal šią taisyklę

$$\widehat{Y} = \begin{cases} 1, & \text{jei} \quad \widehat{p} > 0.5 \\ 0, & \text{jei} \quad \widehat{p} \le 0.5 \end{cases}$$

Turime šiuos automobilio ir visuomeninio transporto pasirenkamumą aprašančius duomenis. Kintamasis X aprašo skirtumą "kelionės autobusu trukmė minus kelionės automobiliu trukmė", o kintamasis Y įgyja reikšmę 1, jei kelionei į darbą pasirenkamas automobilis.

Automobiliu	Autobusu	X	Y
52.9	4.4	-48.5	0
4.1	28.5	24.4	0
4.1	86.9	82.8	1
56.2	31.6	-24.6	0
51.8	20.2	-31.6	0
0.2	91.2	91.0	1
27.6	79.9	52.1	1
89.9	2.2	-87.7	0
41.5	24.5	-17.0	0
95.0	43.5	-51.5	0
99.1	8.4	-90.7	0
18.5	84.0	65.5	1
82.0	38.0	-44.0	1
8.6	1.6	-7.0	0
22.5	74.1	51.6	1
51.4	83.8	32.4	1
62.2	90.1	27.9	1
95.1	22.2	-72.9	0
41.6	91.5	49.9	1

Transporto pasirenkamumo duomenys

Didžiausio tikėtinumo metodu ganame

$$\beta_1 + \beta_2 X_t = -0.0644 + 0.0299 X_t$$

Parametrų t reikšmės yra atitinkamai -0.161 ir 2.915. Teigiamas polinkio koeficientas rodo, kad kelionės visuomeniniu transportu trukmės padidėjimas padidina ir tikimybę kelionei į darbą pasirinkti nuosavą automobilį. Koeficientas yra statistiškai reikšmingas. Tuo tarpu laisvasis narys nėra statistiškai reikšmingas. Tačiau jo neigiamas ženklas rodo, kad jei laikas vyktį į darbą yra vienodas tiek visuomeniniu transportu, tiek nuosavu automobiliu, tai labiau tikėtinas visuomeninio transporto pasirinkimas.

Įvertintu modeliu galime pasinaudoti norėdami prognozuoti individo elgesį tuo atveju, kai kelionė į darbą autobusu truktų 30 minučių ilgiau, nei kelionė nuosavu automobiliu. Taikydami probit modelį, turime

$$\widehat{p} = F(\widehat{\beta}_1 + \widehat{\beta}_2 X) = F(-0.0644 + 0.0299 \cdot 30) = 0.798$$

Kadangi įvertintoji tikimybė yra daugiau nei 0.5, tai galime prognozuoti, kad jei kelionės į darbą trukmė visuomeniniu transportu yra pusę valandos ilgesnė nei nuosavu automobiliu, pirmenybė bus teikiama pastarajam.

6 Simultaninių lygčių modeliai

6.1 Paprastas pavyzdys

Nagrinėkime uždarą ekonomiką, be užsienio prekybos. Pajamų (BVP) tapatybė yra

$$Y = C + I + G;$$

čia
C – vartojimo išlaidos;
I –investicijos;
G – vyriausybės išlaidos;
Viskas matuojama realiame laike. Modelio konstravimas, tai įvairios hipotezės (prielaidos) apie komponentes sudarančias BVP. Pavyzdžiui, atžvilgiu varojimo išlaidų galime daryti prielaidą, kad jos priklauso nuo mokesčių sistemos ir palūkanų normos. Taigi, galime imti

$$C = f((1-\tau)Y, r) + \varepsilon.$$

Čia τ mokesčių lygis, ta pati konstanta visai ekonomikai. Pavyzdžiui, vidutinis mokesčių dydis (pvz., 0.3); r – palūkanų norma, taip pat apibendrinta visai ekonomikai. Galime teigti, kad dalinės išvestinės turi tokius ženklus ir reikšmes:

$$0 < \frac{\partial f}{\partial x_1} < 1, \qquad \frac{\partial f}{\partial x_2} < 0.$$

Naudojant ekonominę interpretaciją, funkcijos f dalinė išvestinė pagal pirmąjį argumentą $\partial f/\partial x_1$ reiškia polinkį vartoti, atsižvelgiant į pajamas. Todėl natūrali prielaida, kad didėjant pajamoms didės ir vartojimas. O išvestinės reikšmė mažesnė už vienetą paprasčiausiai reiškia, kad nevartojama daugiau nei turima (nėra išlaidų deficito). Neigiama funkcijos f dalinė išvestinė pagal antrąjį argumentą reiškia, kad didėjančios palūkanos turi neigiamą įtaką vartojimui ir tai yra aišku, nes didesnės palūkanos skatina taupymą.

Lygiai taip pat galime modeliuoti investicijas. Tarkime, investicijos

$$I = g(\Delta Y, r) + \delta$$

yra BVP pokyčio ir palūkanų normos funkcija. Panašiai interpreduodami funkcijos g dalines išvestines, turime tokias prielaidas jų ženklams:

$$0 < \frac{\partial g}{\partial x_1}, \qquad \frac{\partial g}{\partial x_2} < 0.$$

Surinkę viską į vieną vietą, gauname sistemą

$$\begin{cases} Y &= C + I + G \\ C &= f((1-\tau)Y, r) + \varepsilon \\ I &= g(\Delta Y, r) + \delta. \end{cases}$$

Be to, turime informaciją apie išvestinių ženklus ir galimas jų reikšmes. Gavome modelį. Tai aišku tik teorija apie simultaninį dydžių C, I, G sąryšį, kuris, aišku, priklauso nuo dydžių G, r, τ. Kaip nagrinėti tuos tris kintamuosius? Ar reikalinga pavyzdžiui papildoma teorija, paaiškinanti vyriausybės išlaidas G, palūkanų normą r ir mokesčių lygį τ? Jei tokią teoriją kursime, atsiras naujos lygtys ir dar keli nauji kintamieji.

Sudarytas modelis turi dvi elgsenos lygtis – vartojimui ir investicijoms. Taigi, ekonomikos teorijos pagalba nustatėme paaiškinančiuosius kintamuosius ir išvestinių (parametrų) ženklus. Tačiau dar lieka nemažai atvirų klausimų. Keletą panagrinėkime.

Pirmiausia kyla klausimas dėl funkcinės formos. Tik teoriniais samprotavimais neįmanoma nustatyti funkcionalinės formos, kuria susieti kintamieji į vieną sąryšį. Galime nagrinėti daugybę funkcionalinių formų ir dauguma iš jų tenkins nustatytus išvestinių ženklus. Pavyzdžiui, pažymėję $Z = (1-\tau)Y$ (kas lieka iš pajamų, atskaičiavus mokesčius) disponuojamas pajamas, vartojimą galime aprašyti tokiomis funkcijomis:

$$\begin{aligned} C &= \alpha_0 + \alpha_1 Z + \varepsilon \\ C &= A Z^{\alpha_1} \varepsilon \\ C &= \alpha_0 - \alpha_1 Z^{-1} + \varepsilon. \end{aligned}$$

Aišku, kad koeficientus galime parinkti taip, kad kiekviena iš tų funkcijų tenkintų apribojimus dalinių išvestinių ženklams. Tačiau visos šios formos turi skirtingą kiekybinę prasmę. Pirmoji: 100 Lt. pajamų padidėjimas visados duoda tą patį išlaidų padidėjimą, nes $C' = \alpha_1$ nepriklauso nuo Z. Antroji ir trečioji formos aprašo skirtingą vartojimo pokyčio priklausomybę nuo pajamų: $C' = A\alpha_1 Z^{\alpha_1 - 1}\varepsilon$ (Z padidėjimas skatina išlaidų pokytį) arba $C' = \alpha_1 Z^{-2}$. Išvada, kurią galime padaryti iš šio paprasto pavyzdžio, yra tokia:

> *ekonomikos teorija funkcionalinę modelio formą gali tik rekomenduoti.*

Antrasis svarbus klausimas – duomenys ir matavimo vienetai. Toliau – klausimas apie lagų struktūrą, kiekybines ir kokybines pasekmes, teorijos parinkimą.

Ekonometrijos tikslas atsakyti į visus panašius klausimus. Tai pasiekiama per modelio specifikavimą ir specifikavimo patikrinimą (verifikavimą).

Paprastai pradedama nuo paprasčiausios modelio formos. Nagrinėjamame pavyzdyje tai galėtų būti tiesinis modelis

$$\begin{cases} C_t &= \alpha_0 + \alpha_1(1-\tau)Y_t + \alpha_2 r_t + \varepsilon_t \\ I_t &= \beta_0 + \beta_1(Y_{t-1} - Y_{t-2}) + \beta_2 r_{t-1} + \delta_t \\ Y_t &= C_t + I_t + G_t. \end{cases}$$

A priori parametrų ženklai ir reikšmės yra

$$0 \le \alpha_1 < 1, \quad \alpha_2 < 0, \quad \beta_1 > 0, \quad \beta_2 < 0.$$

Be to, t reiškia laiką. Paprastai tai metai, ketvirtis ar mėnuo. Tačiau būtinai laikas diskretus. Pateiktas modelis yra struktūrinė simultaninių lygčių modelio forma. Endogeniniai kintamieji yra C_t, I_t, Y_t. Dydžiai Y_{t-1}, Y_{t-2} – laginiai endogeniniai, G_t, r_t – egzogeniniai, r_{t-1} – laginis egzogeninis.

Tagi, modeliu paaiškiname endogeninių kintamųjų elgesį kitų kintamųjų atžvilgiu. Kokių, priklauso nuo modelio. Sistemą išsprendę endogeninių kintamųjų atžvilgiu gauname redukuotą modelio formą:

$$C_t = \pi_{10} + \pi_{11}G_t + \pi_{12}r_t + \pi_{13}r_{t-1} + \pi_{14}Y_{t-1} + \pi_{15}Y_{t-2} + \delta_t^C \tag{6.1}$$

$$Y_t = \pi_{20} + \pi_{21}G_t + \pi_{22}r_t + \pi_{23}r_{t-1} + \pi_{24}Y_{t-1} + \pi_{25}Y_{t-2} + \delta_t^Y \tag{6.2}$$

$$I_t = \pi_{30} + \pi_{33}r_{t-1} + \pi_{34}Y_{t-1} + \pi_{35}Y_{t-2} + \delta_t^I. \tag{6.3}$$

Pastebėsime, kad kiekviena redukuota lygtis yra regresinė. Jų tyrimui naudojama regresinių modelių teorija.

Parametrai π yra labai svarbūs. Jie aprašo endogeninių kintamųjų pasikeitimus, pasikeitus atitinkamiems išankstiniams kintamiesiems. Nagrinėkime pavyzdžiui G_t padidėjimą. Jei imtume tik struktūrinę formą, vienetu padidėjęs vartojimas tik tiek ir tebadidintų BVP. Bet iš vartojimo funkcijos matosi, kad padidėjęs BVP skatins vartojimą, kuris, savo ruožtu, vėl padidins BVP. Iš redukuotos formos gauname daugiau informacijos:

$$\frac{\partial Y_t}{\partial G_t} = \pi_{21} = \frac{1}{1 - \alpha_1(1-\tau)}.$$

Tai nacionalinių pajamų multiplikatorius (daugiklis) paprastojoje Keynes'o teorijoje. Jei $\tau = 0.25, \alpha_1 = 0.8$ tai $\pi_{21} = 2.5$. Vadinasi, vienetu padidinus vyriausybės lšlaidas, kai kiti parametrai nekinta, nacionalinės pajamos padidėtų 2.5 vienetais. Visi parametrai π yra įtakos multiplikatoriai. Jie parodo, kokią itaką einamuoju momentu daro išsankstiniai kintamieji. Tačiau čia dar ne viskas. Redukuota modelio forma parodo taip pat dinamines endogeninių dydžių savybes. Perrašykime (6.2) lygtį skirtumine forma:

$$\Delta Y_t = \pi_{21}\Delta G_t + \pi_{22}\Delta r_t + \pi_{23}\Delta r_{t-1} + \pi_{24}\Delta Y_{t-1} + \pi_{25}\Delta Y_{t-2}. \qquad (6.4)$$

Čia $\Delta Y_t = Y_t - Y_{t-1}$. Panagrinėkime šią lygtį. Tarkime, G ir r pakankamai ilga laikotarpį buvo nekintantys Y atžvilgiu tam tikrame pusiausvyriniame lygyje. Taigi turime šias sąlygas:

$$\begin{aligned} \Delta G_t &= \Delta G_{t-1} = \cdots = 0 \\ \Delta r_t &= \Delta r_{t-1} = \cdots = 0 \\ \Delta Y_t &= \Delta Y_{t-1} = \cdots = 0 \end{aligned}$$

Toliau tarkime, vyriausybės išlaidos $t+1$-uoju periodu padidintos dydžiu d ir toliau išlaikomos nekintančios, t.y.

$$\Delta G_{t+1} = d, \quad \Delta G_{t+2} = \Delta G_{t+3} = \cdots = 0.$$

Iš skirtuminės (6.4) lygties gauname

$$\Delta Y_{t+1} = \pi_{21} d.$$

Iš tos pačios (6.4) lygties taip pat gauname

$$\Delta Y_{t+2} = \pi_{24}\Delta Y_{t+1} = \pi_{24}\pi_{11} d.$$

Periodui $t+3$ turime

$$\begin{aligned} &\Delta Y_{t+3} = \pi_{24}\Delta Y_{t+2} + \pi_{25}\Delta Y_{t+1} = \\ &\pi_{24}^2\pi_{21} d + \pi_{25}\pi_{21} d. \end{aligned}$$

Taigi pakeitę viename žingsnyje vyriausybės išlaidas, dėl pasirinktos lagų (vėlinimų) sistemos, gauname visą eilę Y pokyčių Iš (6.4) išvedame

$$\begin{aligned} \frac{\partial Y_{t+1}}{\partial G_{t+1}} &= \pi_{11} \quad \text{nulinis lagas, arba įtakos multiplikatorius} \\ \frac{\partial Y_{t+2}}{\partial G_{t+1}} &= \pi_{14}\pi_{11} \quad \text{vieno periodo lagas} \\ \frac{\partial Y_{t+3}}{\partial G_{t+1}} &= (\pi_{14}^2 + \pi_{15})\pi_{11} \quad \text{dviejų periodų lagas.} \end{aligned}$$

Multiplikatoriai ir įvairių laikotarpių lagai vadinami laikotarpių (interim) multiplikatoriais. Jei juos susumuosime laike (tardami, kad sumos konverguoja), gausime visuminius (total) multiplikatorius. Jie aprašo endogeninio dydžio pusiausvirinę reikšmę, viename žingsnyje pakeitus egzogeninio dydžio reikšmę.

Pagaliau panagrinėkime galutinę modelio formą. Iš (6.2) lygties išvedame

$$Y_t - \pi_{14} Y_{t-1} - \pi_{15} Y_{t-2} = \pi_{10} + \pi_{11} G_t + \pi_{12} r_t + \pi_{13} r_{t-1}. \tag{6.5}$$

Tai antrosios eilės nehomogeninė skirtuminė lygtis. Perrašykime ją bendresne forma

$$Y_t - \pi_{14} Y_{t-1} - \pi_{15} Y_{t-2} = f(G, r).$$

Dinaminės sistemos įdomus efektas yra tai, kad kiekvieną endogeninį kintamąjį galime išreikšti antros eilės skirtumine lygtimi su tais pačiais koeficientais. Skirsis tik laisvieji nariai. Tikrai, kadangi

$$I_t = \beta_0 + \beta_1 (Y_{t-1} - Y_{t-2}) + \beta_2 r_{t-1},$$

tai

$$\begin{aligned}
&I_t - \pi_{14} I_{t-1} - \pi_{15} I_{t-2} = \beta_1 (Y_t - \pi_{14} Y_{t-1} - \pi_{15} Y_{t-2}) - \\
&\beta_1 (Y_t - \pi_{14} Y_{t-1} - \pi_{15} Y_{t-2}) + \beta_0 (1 - \pi_{14} - \pi_{15}) + \\
&\beta_2 (r_{t-1} - \pi_{14} r_{t-2} - \pi_{15} r_{t-3})
\end{aligned}$$

Taigi

$$I_t - \pi_{14} I_{t-1} - \pi_{15} I_{t-2} = h(G, r).$$

Pagaliau

$$\begin{aligned}
&C_t - \pi_{14} C_{t-1} - \pi_{15} C_{t-2} = \beta_1 (Y_t - \pi_{14} Y_{t-1} - \pi_{15} Y_{t-2}) - \\
&(I_t - \pi_{14} I_{t-1} - \pi_{15} I_{t-2}) - (G_t - \pi_{14} G_{t-1} - \pi_{15} G_{t-2}) = k(G, r).
\end{aligned}$$

Taigi visi trys endogeniniai kintamieji tenkina antros eilės skirtuminę lygtį su tais pačiais koeficientais. Reikia išspręsti charakteristinę lygtį

$$\lambda^2 - \pi_{14} \lambda - \pi_{15} = 0.$$

Struktūrinė ir redukuota lygtys rodo, kad

$$\pi_{14} = -\pi_{15} = \frac{\beta_1}{1 - \alpha_1 (1 - \tau)} > 0.$$

Pažymėję tą parametrą α, turime skirtuminę lygtį

$$\lambda^2 - \alpha \lambda + \alpha = 0.$$

Jos sprendinys yra

$$\lambda_{1,2} = \frac{1}{2}(\alpha \pm \sqrt{\alpha(\alpha - 4)}).$$

Jei $\alpha < 4$, tai šaknys yra kompleksinės. Tuomet ekonominė sistema yra cikliška. Be to, jei $\alpha < 1$ tai ciklai yra silpnėjantys (damped), o jei $1 < \alpha < 4$ – sprogstantys (explosive). Parametras α, kaip matome, priklauso nuo parametro β_1 – investicijų akseleracijos koeficiento, α_1 – marginalinio polinkio vartoti, τ – mokesčių.

7 Laiko eilučių modeliai

Laiko eilučių modeliais galime prognozuoti ekonominių rodiklių elgesį pagal jų stebėjimus praeityje. Pavyzdžiui, stebime $y(t), 0 \leq t \leq T$. Kaip elgsis $y(t)$, kai $t > T$? Čia $y(t)$ gali būti biržos indeksas, palūkanų norma, prekių indeksas, prekės pardavimo kaina ir pan. Naudodami laiko eilučių teriją nustatome ją atitinkančio proceso $y(t)$ modelį. Įvertinę modelio parametrus, galėsime prognozuoti $y(t)$, kai $t > T$.

Laiko eilute vadiname kokio nors dydžio, tarkime, y stebėjimų laike seką $y_1, y_2, \ldots, y_T$. Stochastinių laiko eilučių modeliai remiasi prielaida, kad tie stebėjimai yra generuoti stochastinio proceso $(Y(t), t \in I)$. Čia I yra laiką atitinkanti parametrų aibė ir tai gali būti pavyzdžiui Z, N. Toliau nagrinėsime tik laiko eilutes indeksuotas sveikaisiais skaičiais.

Stochastinės laiko eilutės $y_1, \ldots, y_T$ pagalba reikia nustatyti ją atitinkančio atsitiktinio proceso $(Y(t), t \in Z)$ charakteristikas.

7.1 Stacionariosios laiko eilutės

Laiko eilutė $y_1, y_2, \ldots, y_T$ vadinama stacionariąja, jei ją atitinkantis procesas $(Y(t), t \in I)$ yra stacionarus plačiąja prasme.

Priminsime, kad procesas $(Y_t, t \in Z)$ vadinamas *stacionariuoju siaurąja prasme*, jei jo baigtiniamačiai skirstiniai yra invariantiniai laiko postumiui, t.y.

$$P_{t_1,\ldots,t_n} = P_{t_1+h,\ldots,t_n+h},$$

su visais $t_1 < t_2 < \cdots < t_n \in Z$ ir $h \in Z$. Skirstinys $P_{t_1,\ldots,t_n}$ apibrėžtas taip:

$$P_{t_1,\ldots,t_n}(A) = P((Y(t_1)\ldots, Y(t_n) \in A),$$

kai A yra erdvės R^n Borelio aibė.

Procesas $(Y(t), t \in Z)$ vadinamas *stacionariuoju plačiąja prasme*, jei

- $EY(t)^2 < \infty$, su kiekvienu $t \in Z$;
- su kiekvienu $t \in Z$, $EY(t) = EY(0)$;
- su visais $s, t, h \in Z$, $r(s,t) = r(s+h, t+h)$.

Čia $r(s,t)$ yra proceso $Y(t)$ kovariacinė funkcija:

$$r(s,t) = E(Y(t) - EY(t))(Y(s) - EY(s)), \quad s,t \in Z.$$

Kadangi stacionariojo plačiąja prasme proceso $Y(t)$ kovariacinė funkcija priklauso tik nuo argumentų skirtumo, t.y. $r(s,t) = r(s-t,0)$, tai ją patogu traktuoti kaip vieno kintamojo funkciją ir sutrumpintai rašyti $r(t)$, $r(t) = r(t,0)$.

Stacionarus siaurąja prasme procesas yra stacionarus ir plačiąja prasme. Atvirkščiai teisinga ne visada. Gauso procesui abi stacionarumo sąvokos sutampa.

Svarbi proceso charakteristika – autokoreliacinė funkcija.

7.1.1 apibrėžimas. *Stacionariojo proceso $(X_t, t \in Z)$ autokoreliacine funkcija vadiname funkciją*

$$\rho(k) = \frac{r(k)}{r(0)}, k \in Z.$$

Turėdami stacionarią laiko eilutę $y_1, y_2, \ldots, y_T$ ją atitinkančio proceso kovariacinę funkciją galime įvertinti taip:

$$\widehat{r}(h) = \sum_{t=1}^{T-h} (y_t - \overline{y})(y_{t+h} - \overline{y}),$$

o autokoreliacinę funkciją –

$$\widehat{\rho}(h) = \frac{\sum_{t=1}^{T-h} (y_t - \overline{y})(y_{t+h} - \overline{y})}{\sum_{t=1}^{T} (y_t - \overline{y})^2}.$$

7.1.2 apibrėžimas. *Stacionarusis procesas $(Z_t, t \in Z)$ vadinamas baltuoju triukšmu su vidurkiu 0 ir dispersija σ^2, jei jo vidurkis lygus nuliui, o kovariacinė funkcija –*

$$r(h) = \begin{cases} \sigma^2, & \textit{kai } h = 0, \\ 0, & \textit{kai } h > 0. \end{cases}$$

Trumpai rašysime $(Z_t) \in BT(0, \sigma^2)$. Baltojo triukšmo autokoreliacinė funkcija yra

$$\rho(h) = \begin{cases} 1, & \text{kai } h = 0, \\ 0, & \text{kai } h > 0. \end{cases}$$

Tirdami, ar laiko eilutė atitinka baltąjį triukšmą naudojame įvairius kriterijus paremtus autokoreliacinės funkcijos įvertinimu $\widehat{\rho}(h)$.

Bartlet'o kriterijus: jei laiko eilutė generuota baltojo triukšmo, tai $\widehat{\rho}(h)-\rho(h)$, kai $h > 0$, turi apytikriai normalinį skirstinį su nuliniu vidurkiu ir dispersija $1/T$.

Box – Pierce'o kriterijus: statistika

$$Q = T \sum_{k=1}^{K} \widehat{\rho}_k^2$$

turi apytikrį χ^2 skirstinį su K laisvės laipsniais.

7.2 Homogeninės laiko eilutės

Praktikoje labai retai laiko eilutės atitinka stacionariuosius procesus. Ypač ekonominių rodiklių elgesį atitinkančios laiko eilutės. Tačiau dažnai pasitaiko, kad laiko eiluė turi sąvybę: vieną ar kelis kartus diferencijuota eilutė yra stacionari. Laiko eilutės turinčios šią savybę vadinamos homogeninėmis. Homogeniškumo eilę nusako atitinkamo diferencijavimo eilė.

7.2.1 apibrėžimas. *Procesas $(Y(t), t \in Z)$ vadinamas k-os eilės homogeniniu procesu, jei procesas $(w_k(t), t \in Z)$ yra stacionarus. Čia*

$$\begin{aligned} &w_k(t) = w_{k-1}(t) - w_{k-1}(t-1), \\ &w_0(t) = Y(t) - Y(t-1), \quad t \in Z. \end{aligned}$$

Procesas $(w_k(t), t \in Z)$ dažnai vadinamas $k-$aja proceso $(Y(t), t \in Z)$ išvestine. Laiko eilutė $y_1, \ldots, y_T$ vadinama k-os eilės homogenine laiko eilute, jei ji atitinka k-os eilės homogeninį procesą.

Labai svarbus nestacionaraus, bet homogeninio stochastinio proceso pavyzdys yra atsitiktinis klaidžiojimas $(Y(t), t \in Z)$. Šis procesas apibrėžtas taip:

$$Y(t) = Y(t-1) + \varepsilon_t, t \in Z; \tag{7.1}$$

čia $(\varepsilon_t) \in BT(0, \sigma^2)$. Dar kitaip šis procesas vadinamas integruotu baltuoju triukšmu.

Patikrinti ar duotoji laiko eilutė atitinka atsitiktinį klaidžiojimą, yra naudojami įvairūs testai.

7.3 Vėlinimo operatorius

Tarkime, turime seką $(x(t), t \in Z)$. *Vėlinimo operatoriumi* vadiname operatorių L, veikiantį pagal taisyklę

$$Lx(t) = x(t-1), \quad t \in Z.$$

Galime apibrėžti velinimo operatoriaus bet kurį laipsnį:

$$L^k x(t) = L(L^{k-1}x(t)) = x(t-k), \quad t \in Z.$$

Be to, susitarsime, kad $L^0 x(t) = x(t)$ su visais $t \in Z$. Ateityje mums pravers kai kurios vėlinimo operatoriaus savybės. Pirmiausia pastebėkime, kad

$$L(a + bx(t)) = a + bLx(t) = a + bx(t-1),$$

kai $a, b \in R$. Ši savybė vadinama operatoriaus tiesiškumu. Tarkime, duotas polinomas

$$\alpha_p(v) = 1 + \alpha_1 v + \cdots + \alpha_p v^p.$$

Tuomet turi prasmę $\alpha_p(L)$,

$$\alpha_p(L) = 1 + \alpha_1 L + \cdots + \alpha_p L^p,$$

veikiantis pagal taisyklę

$$\alpha_p(L)x(t) = x(t) + \alpha_1 x(t-1) + \cdots + \alpha_p x(t-p).$$

Be to, galime apibrėžti ir eilutes

$$\psi_\infty(L) = L^0 + \psi_1 L + \psi_2 L^2 + \cdots$$

Jei operatoriai $\psi_p(L)$, ir $\phi_q(L)$ yra tokie, kad

$$\psi_p(L)\phi_q(L) = 1 = L^0,$$

tai $\phi_q(L)$ vadinamas $\psi_p(L)$ atvirkštiniu operatoriumi ir dažnai žymimas $\psi_p^{-1}(L)$. Pavyzdžiui, operatoriui $\psi_1(L) = 1 - \psi L$, kai $|\psi| < 1$ atvirkštinis yra

$$\psi_1^{-1}(L) = 1 + \psi L + \psi^2 L^2 + \cdots .$$

Tai visai nesunku įrodyti.

7.4 Autoregresiniai modeliai

7.4.1 Savybės

Laiko eilutė $y_1, \ldots, y_T$ aprašoma parametro p autoregresiniu modeliu, jei ją atitinkantis procesas $(Y(t), t \in Z)$ tenkina skirtuminę lygtį

$$Y(t) = \phi_1 Y(t-1) + \cdots + \phi_p Y(t-p) + v + \varepsilon_t, \quad t \in Z. \tag{7.2}$$

Čia $(\varepsilon_t, t \in Z) \in BT(0, \sigma^2)$, $v \in R$. Trumpai žymėsime $(y_t, t = 1, \ldots, T) \in AR(p)$. Procesas, tenkinantis skirtuminę lygtį (7.4) vadinamas $AR(p)$ procesu. Pritaikę vėlinimo operatorių L ir pažymėję polinomą

$$\phi_p(z) = 1 - \phi_1 z - \cdots - \phi_p z^p,$$

(7.4) lygtį galime užrašyti kompaktišku būdu taip:

$$\phi_p(L) Y(t) = \varepsilon_t, \quad t \in Z.$$

- Panagrinėkime $AR(1)$ procesą $(Y(t), t \in Z)$:

$$Y(t) = \phi_1 Y(t-1) + \varepsilon_t, \quad t \in Z. \tag{7.3}$$

4 teorema. *Jei* $|\phi_1| \neq 1$, *tai egzistuoja stacionarus lygties* (7.3) *sprendinys* $(Y(t), t \in Z)$. *Be to, jei* $|\phi_1| < 1$, *tai*

$$Y(t) = \sum_{k=0}^{\infty} \phi_1^k \varepsilon_{t-k},$$

jei $|\phi_1| > 1$, *tai*

$$Y(t) = \sum_{k=1}^{\infty} \phi_1^{-k} \varepsilon_{t+k}.$$

Abiem atvejais eilutės konvergavimas suprantamas kvadratinio vidurkio prasme.

Galima įrodyti, kad (7.3) lygtis, kai $|\phi| = 1$ stacionaraus sprendinio neturi.

Kai $|\phi| < 1$, tai, kai matėme, (7.3) lygties sprendinį galime užrašyti ir taip:

$$Y(t) = (1 - \phi B)^{-1}\varepsilon_t.$$

Suskaičiuokime stacionariojo $AR(1)$ proceso $(Y(t), t \in Z)$ koreliacinę funkciją $r(h), h \in Z$. Remdamiesi 1 teorema, kai $|\phi_1| < 1$ gauname

$$r(0) = EY^2(t) = \sum_{k=0}^{\infty} \phi_1^{2k} E\varepsilon_{t-k}^2 = \frac{\sigma^2}{1-\phi_1^2},$$

o kai $h > 0$,

$$r(h) = EY(t+h)Y(t) = E\sum_{k=0}^{\infty}\phi^k\varepsilon_{t-k}\sum_{k=0}^{\infty}\phi^k\varepsilon_{t+h-k} =$$
$$\sum_{i,j=0}^{\infty}\phi_1^i\phi_1^j E\varepsilon_{t-i}\varepsilon_{t+h-j} = \sum_{i=0}^{\infty}\phi_1^i\phi_1^{h+i}\sigma^2 =$$
$$\sigma^2\phi_1^h\sum_{i=1}^{\infty}\phi_1^{2i} = \frac{\sigma^2\phi_1^h}{1-\phi_1^2}.$$

Analogiškai, jei $|\phi_1| > 1$, tai

$$r(0) = EY^2(t) = \sum_{k=1}^{\infty}\phi_1^{-2k}E\varepsilon_{t+k}^2 = \frac{\sigma^2}{\phi_1^2 - 1}.$$

o

$$r(h) = \frac{\sigma^2\phi_1^{-h}}{\phi_1^2 - 1},$$

kai $h \neq 0$. Iš gautų rezultatų randame $AR(1)$ proceso autokoreliacinę funkciją

$$\rho(h) = \begin{cases} \phi_1^{|h|}, & \text{kai } |\phi_1| < 1, \\ \phi_1^{-|h|}, & \text{kai } |\phi_1| > 1, \end{cases} \quad h \neq 0.$$

Matome, kad autoregresinio proceso autokoreliacinė funkcija yra eksponentiškai nykstanti. Nesunku matyti, kad aurokoreliacinė funkcija $\rho(h)$ tenkina skirtuminę lygtį

$$\rho(h) = \phi\rho(h-1), \quad h = 1, 2, \ldots$$

•• Tarkime, $(Y(t), t \in Z)$ yra $AR(p)$:

$$Y(t) = \phi_1 Y(t-1) + \cdots + \phi_p Y(t-p) + \varepsilon_t, \quad t \in Z. \tag{7.4}$$

Šita lygtis turi stacionarų sprendinį tada ir tik tada, kai charaktringasis polinomas

$$1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p$$

neturi šaknų vienetiniame skritulyje $\{z \in C : |z| = 1\}$. Tuomet egzistuoja operatoriaus $\phi(L) = 1 - \phi_1 L - \cdots - \phi_p L^p$ atvirkštinis operatorius.

Suskaičiuokime to proceso kovariacinę funkciją. Nesunku matyti, kad kovariacinė funkcija tenkina šias lygtis:

$$\begin{aligned}
&r(k) = EY(t-k)Y(t) = \\
&EY(t-k)(\phi_1 Y(t-1) + \cdots + \phi_p Y(t-p) + \varepsilon_t) = \\
&\phi_1 r(k-1) + \cdots + \phi_p r(k-p) + EY(t-k)\varepsilon_t = \\
&\begin{cases} \phi_1 r(k-1) + \cdots + \phi_p r(k-p), & \text{kai } k \neq 0 \\ \phi_1 r(k-1) + \cdots + \phi_p r(k-p) + \sigma^2, & \text{kai } k = 0. \end{cases}
\end{aligned}$$

Paėmę $k = 0, 1, \ldots, p$ gauname šią lygčių sistemą:

$$\begin{cases} r(0) = \phi_1 r(1) + \phi_2 r(2) + \cdots + \phi_p r(p) + \sigma^2 \\ r(1) = \phi_1 r(0) + \phi_2 r(1) + \cdots + \phi_p r(p-1) \\ \cdots \\ r(p) = \phi_1 r(p-1) + \phi_2 r(p-2) + \cdots + \phi_p r(0) \end{cases} \tag{7.5}$$

Tai yra pradinių sąlygų sistema. Ją išsprendę randame $r(0),\ r(1), \ldots, r(p)$. Šios pradinės sąlygos reikalingos norint išspręsti skirtuminę lygtį

$$r(k) = \phi_1 r(k-1) + \phi_2 r(k-2) + \cdots + \phi_p r(k-p), \tag{7.6}$$

kai $k > p$. Gerai žinoma, kad bendrasis (7.6) lygties sprendinys yra

$$r(k) = \sum_{i=1}^{m} \sum_{j=0}^{r_i - 1} d_{ij} k^j z_i^{-k}, \quad k \geq 0.$$

Čia $z_1, \ldots, z_m$ yra skirtingi polinomo $1 - \phi_1 z - \cdots - \phi_p z^p$ nuliai, $r_1, \ldots, r_m$ atitinkamų nulių kartotinumas ($\sum_{i=1}^m r_i = p$), o d_{ij} – laisvieji parametrai, randami iš pradinių sąlygų.

Iš (7.5) sistemos išvedame taip vadinamas Jule-Valkerio (Yule-Walker) lygtis $AR(p)$ proceso autokoreliacinei funkcijai:

$$\begin{cases} \rho(1) = \phi_1 + \phi_2 \rho(1) + \cdots + \phi_p \rho(p-1) \\ \rho(2) = \phi_1 \rho(1) + \phi_2 + \cdots + \phi_p \rho(p-2) \\ \cdots \\ \rho(p) = \phi_1 \rho(p-1) + \phi_2 \rho(p-2) + \cdots + \phi_p \end{cases} \tag{7.7}$$

Matriciniu pavidalu sistemą galime perrašyti taip

$$\boldsymbol{\rho} = P\boldsymbol{\phi}$$

ir galime išspręsti:

$$\boldsymbol{\phi} = P^{-1}\boldsymbol{\rho}.$$

Pakeitę $\boldsymbol{\rho}$ į $\widehat{\boldsymbol{\rho}}$ gauname vadinamus momentinius autoregresinio modelio parametrų įvertinimus.

Suskaičiuokime $AR(2)$ proceso

$$Y(t) = 0.6Y(t-1) - 0.9Y(t-2) + \varepsilon_t, \quad t \in Z$$

koreliacinę funkciją. Kai $k > 2$ turime skirtuminę lygtį

$$r(k) = 0.6r(k-1) - 0.9r(k-2).$$

Šios lygties charakteristinis polinomas yra $\lambda^2 - 0.6\lambda + 0.09$ o jo nuliai – $\lambda = 0.3$. Vadinasi, skirtuminės lygties sprendinys yra

$$r(k) = (d_1k + d_2)0.3^k.$$

Norėdami rasti parametrus d_1, d_2, turime išspręsti sistemą

$$\begin{cases} r(0) = 0.6r(1) - 0.9r(2) + 1 \\ r(1) = 0.6r(0) - 0.9r(1) \\ r(2) = 0.6r(1) - 0.9r(0) \end{cases}$$

7.4.2 Autoregresinio modelio įvertinimas

Tardami, kad stochastinė laiko eilutė $y_1, \dots, y_T$ atitinka autoregresinį procesą, reikia nustatyti to proceso eilę p ir įvertinti parametrus $\phi_1, \dots, \phi_p$.

- Pirma tarkime, kad eilę p žinome. Tuomet

$$y_t = \phi_1 y_{t-1} + \dots + \phi_p y_{t-p} + \varepsilon_t,$$

kai $(\varepsilon_t) \in BT(0, \sigma^2)$. Čia $t = p+1, \dots, T$. Bet tai yra tiesinis regresinis modelis, kurį galime užrašyti taip:

$$\boldsymbol{Y}_p = \boldsymbol{X}_p\boldsymbol{\phi}_p + \boldsymbol{\varepsilon};$$

čia $\boldsymbol{Y} = (y_{p+1}, \ldots, y_T)^\tau$, $\boldsymbol{\varepsilon} = (\varepsilon_{p+1}, \ldots, \varepsilon_T)^\tau$,

$$\boldsymbol{X}_p = \begin{pmatrix} y_p & y_{p-1} & \cdots & y_1 \\ \vdots & \vdots & \ddots & \vdots \\ y_{T-1} & y_{T-2} & \cdots & y_{T-p} \end{pmatrix}.$$

Pritaikę mažiausių kvadratų metodą, gauname

$$\widehat{\boldsymbol{\phi}}_p = (\boldsymbol{X}_p^\tau \boldsymbol{X}_p)^{-1} \boldsymbol{X}_p^\tau \boldsymbol{y}_p.$$

•• Praktikoje retai atsitinka, kad autoregresinio proceso eilė p yra žinoma.

Jo įvertinimui dažniausiai naudojame *dalinės autkoreliacinės funkcijos* savybes. Pirmiausia išsiaiškinkime pačią sąvoką.

Koreliacija tarp dviejų dydžių dažnai atsiranda todėl, kad abu tie dydžiai koreliuoja su trečiuoju. Laiko eilutėje $(y_t, t = 1, \ldots, T)$ koreliaciją tarp y_t ir y_{t-k} didele dalimi lemia abiejų tų dydžių koreliacija su $y_{t-1}, \ldots, y_{t-k+1}$. Tokiai koreliacijai nustatyti ir yra naudojama dalinė autokoreliacija $p(k), k = 1, \ldots, T$. Dalinės autokoreliacijos reikšmė $p(k)$ prilyginama koeficientui ϕ_{kk} šioje autoregresinėje lygtyje:

$$y_t = \phi_{k1} y_{t-1} + \cdots + \phi_{kk} y_{t-k} + \varepsilon_t.$$

Autokoreliacijos koeficientas $p(1)$ yra regresijos koeficientas ϕ_{11} lygtyje

$$y_t = \phi_{11} y_{t-1} + \varepsilon_t.$$

Iš Jule-Walkerio lygčių AR(1) modeliui matome, kad $\phi_{11} = \rho(1)$. Tai yra, dalinės autokoreliacijos koeficientas $p(1)$ yra lygus autokoreliacijos koeficientui $\rho(1)$. Tai nieko stebėtino, nes tarp dydžių y_t ir y_{t-1} nėra jokių tarpinių dydžių. Koeficientas $p(2)$ yra lygus koeficientui ϕ_{22} regresinėje lygtyje

$$y_t = \phi_{21} y_{t-1} + \phi_{22} y_{t-2} + \varepsilon_t.$$

Iš Jule-Walkerio sistemos AR(2) procesui matome, kad

$$\rho(1) = \phi_{21} + \rho(1)\phi_{22}$$
$$\rho(2) = \rho(1)\phi_{21} + \phi_{22}$$

Iš šios sistemos randame

$$p(2) = \phi_{22} = \frac{\rho(2) - \rho^2(1)}{1 - \rho^2(1)}.$$

Bendruoju atveju taip pat galime pasinaudoti Jule-Walkerio lygčių sistema

$$\boldsymbol{\rho} = R\boldsymbol{\phi};$$

čia

$$\boldsymbol{\rho} = (\rho(1), \dots, \rho(k))', \quad \boldsymbol{\phi} = (\phi_{k1}, \dots, \phi_{kk})',$$

$$R = \begin{pmatrix} 1 & \rho(1) & \dots & \rho(k-1) \\ \rho(1) & 1 & \dots & \rho(k-2) \\ \vdots & \vdots & \ddots & \vdots \\ \rho(k-1) & \rho(k-2) & \dots & 1 \end{pmatrix}.$$

Iš šios sistemos randame

$$p(k) = \phi_{kk} = \frac{|R_k|}{|R|}.$$

čia $|A|$ reiškia matricos A determinantą,

$$R_k = \begin{pmatrix} 1 & \rho(1) & \dots & \rho(k-2) & \rho(1) \\ \rho(1) & 1 & \dots & \rho(k-3) & \rho(2) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho(k-1) & \rho(k-2) & \dots & 1 & \rho(k) \end{pmatrix}.$$

Pagal dalinės autokoreliacijos apibrėžimą ir ką tik išvestas formules matome, kad

$$\begin{aligned} &\text{AR(1) modeliui:} \;\; p(1) = \rho(1), \quad p(k) = 0, \text{ kai } \; k > 1; \\ &\text{AR(2) modeliui:} \;\; p(1) = \rho(1), \; p(2) = \frac{\rho(2) - \rho^2(1)}{1 - \rho^2(1)}, \quad p(k) = 0, \text{ kai } \; k > 2; \\ &\text{AR(p) modeliui:} \;\; p(1) \neq 0, \dots, p(p) \neq 0, \quad p(k) = 0, \text{ kai } \; k > p; \end{aligned}$$

Šis faktas pasitarnauja nustatant autoregresinio modelio eilę.

Kitas, gal kiek ir paprastesnis būdas nustatyti autoregresinio proceso eilę yra toks: įvertiname parametrus, tardami, kad procesas yra eilės k ir tikriname hipotezę apie parametro $\widehat{\theta}_k$ reikšmingumą.

7.5 Slenkančio vidurkio modeliai

7.5.1 Savybės

Laiko eilutė $y_1, \dots, y_T$ aprašoma parametro q slenkančiojo vidurkio modeliu $MA(q)$, jei jį atitinkantis procesas $(Y(t), t \in Z)$ tenkina lygtį

$$Y(t) = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q},$$

arba

$$Y(t) = (1 - \theta_1 L - \cdots - \theta_q L^q)\varepsilon_t,$$

kai $(\varepsilon_t, t \in Z) \in BT(0, \sigma^2)$.

Slenkančio vidurkio modeliai dažnai gaunami kaip begelinės eilės autoregresiniai modeliai.

- Tirkime MA(1) procesą:

$$Y_t = \varepsilon_t - \theta \varepsilon_{t-1}, \quad t = 1, \dots, T.$$

Nesunku įsitikinti, kad

$$\rho(1) = \frac{-\theta}{1 + \theta^2}$$

ir

$$\rho(k) = 0, \quad \text{kai} \quad k > 1.$$

Tarkime, turime begalinį AR procesą

$$Y(t) = -\phi Y(t-1) - \phi^2 Y(t-2) - \phi^3 Y(t-3) - \cdots + \varepsilon_t,$$

$t \in Z$. Pritaikę vėlinimo operatorių

$$(1 + \phi L + \phi^2 L^2 + \cdots)Y(t) = \varepsilon_t.$$

Tarkime, $|\phi| < 1$. Prisiminę, kad

$$(1 + \phi L + \phi^2 L^2 + \cdots)^{-1} = (1 - \phi L)$$

(1) lygtį galime perrašyti taip:

$$Y(t) = (1 - \phi L)\varepsilon_t.$$

Taigi, MA(1) procesą aprašytą lygtimi (2), galime išreikšti begaliniu autoregresiniu modeliu.

•• Nagrinėkime MA(2) procesą:

$$Y_t = \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2}, \quad t \in Z.$$

Tą procesą galime užrašyti

$$Y_t = (1 - \theta_1 B - \theta_2 B^2)\varepsilon_t.$$

Nesunku suskaičiuoti, kad

$$\rho(1) = \frac{-\theta_1 + \theta_1\theta_2}{1 + \theta_1^2 + \theta_2^2}, \quad \rho(2) = \frac{-\theta_2}{1 + \theta_1^2 + \theta_2^2}$$

ir

$$\rho(k) = 0, \quad \text{kai} \quad k > 2.$$

Norėdami MA(2) procesą išreikšti begaliniu autoregresiniu procesu turime išspręsti lygtį

$$(1 - \pi_1 L - \pi_2 L^2 - \cdots)(1 - \theta_1 L - \theta_2 L^2) = 1.$$

Tai galime padaryti sulygindami koeficientus prie operatoriaus L laipsnių:

$$\begin{aligned}
&-\pi_1 - \theta_1 = 0 \quad \longrightarrow \pi_1 = -\theta_1\\
&-\pi_2 + \theta_1\pi_1 - \theta + 2 = 0 \longrightarrow \pi_2 = \theta_1\pi_1 - \theta_2^2 = -\theta_1^2 - \theta_2\\
&-\pi_j + \theta_1\pi_{j-1} + \theta_2\pi_{j-2} = 0 \quad \longrightarrow \pi_j = \theta_1\pi_{j-1} + \theta_2\pi_{j-2}, \quad j > 2.
\end{aligned}$$

Tam, kad MA(2) procesą galėtume išreikšti begaliniu autoregresiniu procesu, reikia kad polinomo $1 - \theta_1 z - \theta_2 z^2$ šaknys būtų už vienetinio skritulio.

• • • Panagrinėkime benrąjį atveją. Galime nesunkiai suskaičiuoti autokoreliacinę funkciją:

$$\rho(k) = \frac{-\theta_k + \theta_1\theta_{k+1} + \cdots + \theta_{q-k}\theta_q}{1 + \theta_1^2 + \cdots + \theta_q^2}, \quad k = 1, \ldots, q$$

ir

$$\rho(k) = 0, \quad \text{kai} \quad k > q.$$

Kyla natūralus klausimas, kuriais atvejais $MA(q)$ procesas užrašomas begaliniu AR procesu? Pasirodo, kad tam pakankama yra tokia sąlyga: polinomo

$$\Theta_q(z) = 1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q, \quad z \in C$$

šaknys yra už vienetinio kompleksinės plokštumos rutulio, t.y. $|z| > 1$ su kiekviena to polinomo šaknimi $z \in C$.

7.5.2 Slenkančio vidurkio modelio įvertinimas

Tarkime, laiko eilutė $y_1, \dots, y_T$ atitinka slenkančio vidurkio modelį. Kaip nustatyti to modelio eilę q? Tarkime, $(Y(t), t \in Z) \in MA(q)$. Nesunku suskaičiuoti to proceso kovariacinę funkciją:

$$cov(Y(t), Y(t+k)) = \begin{cases} \sigma^2 \sum_{i=0}^{q-k} \theta_i \theta_{i+k}, & \text{kai } k = 0, 1, \dots, q; \\ 0, & \text{kai } k > q. \end{cases}$$

Čia $\theta_0 = 0$. Autokoreliacinė funkcija yra

$$\rho(k) = \begin{cases} \sum_{i=0}^{q-k} \theta_i \theta_{i+k} / \sum_{i=0}^{q} \theta_i^2, & \text{kai } k = 0, 1, \dots, q; \\ 0, & \text{kai } k > q. \end{cases}$$

Taigi, slenkančiojo vidurkio parametras q yra toks didžiausias k, su kuriuo $\rho(k) \neq 0$. Pagal turimus stebėjimus, įvertinę autokoreliacinę funkciją, įvertiname q ir patikriname reikalingas hipotezes.

Toliau panagrinėkime, kaip įvertinti slenkančio vidurkio proceso parametrus $\theta_1, \dots, \theta_q$. Vienas iš būdų yra analogiškas mažiausiųjų kvadratų metodui. Būtent, reikia minimizuoti paklaidų kvadratų sumą

$$S(\theta) = \sum_{t=1}^{T} (y_t - \theta_1 \varepsilon_{t-1} - \dots - \theta_q \varepsilon_{t-q})^2.$$

Kadangi triukšmas ε_t nėra stebimas, jį reikia įvertinti, panaudojant duomenis $y_1, \dots, y_T$.

7.6 Autoregresiniai slenkančio vidurkio modeliai

Daugelio laiko eilučių negalėtume aprašyti vien slenkančiojo vidurkio arba vien autoregresiniais modeliais. ARMA modeliai yra tų dviejų modelių apjungimas. Sakysime, kad laiko eilutė $(y_1, \dots, y_T)$ aprašoma ARMA(p, q) modeliu, jei ją atitinkantis procesas $(Y(t), t \in Z)$ tenkina lygtį

$$Y(t) = \phi_1 Y(t-1) + \dots + \phi_p Y(t-p) + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \dots - \theta_q \varepsilon_{t-q}, \quad (7.8)$$

kai $(\varepsilon_t, t \in Z) \in BT(0, \sigma^2)$. Pritaikę vėlinimo operatorių L $(ARMApq)$ lygtį galime užrašyti šitaip:

$$(1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p)Y(t) = (1 - \theta_1 L - \cdots - \theta_q L^q)\varepsilon_t.$$

dar trumpiau

$$\phi_p(L)Y(t) = \theta_q(L)\varepsilon_t, \tag{7.9}$$

kai

$$\begin{aligned}\phi_p(t) &= 1 - \phi_1 t - \cdots - \phi_p t^p,\\ \theta_q(t) &= 1 - \theta_1 t - \cdots - \theta_q t^q.\end{aligned}$$

- Panagrinėkime $AR(1,1)$ procesą

$$(1 - \phi L)Y_t = (1 - \theta L)\varepsilon_t, \quad t \in Z.$$

Užrašykime tą procesą begalinio slenkančio vidurkio procesu:

$$Y_t = \varepsilon_t + \psi_1 \varepsilon_{t-1} + \psi_2 \varepsilon_{t-2} + \cdots$$

Begalinio polinomo ψ koeficientus rasime suskaičiavę $\psi(L) = (1 - \theta L)(1 - \phi L)^{-1}$. Sulyginę koeficientus prie atitinkamų L laipsnių sąryšje

$$(1 - \phi L)(1 + \psi_1 L + \psi_2 L^2 + \cdots) = 1 - \theta L$$

randame

$$\begin{aligned}
L^1 :&\quad \psi_1 - \phi = 0 \quad \Longrightarrow \quad \psi_1 = \phi - \theta\\
L^2 :&\quad \psi_2 - \phi\psi_1 = 0 \quad \Longrightarrow \quad \psi_2 = (\phi - \theta)\phi\\
&\vdots\\
L^j :&\quad \psi_j - \phi\psi_{j-1} = 0 \quad \Longrightarrow \quad \psi_2 = (\phi - \theta)\phi^{j-1}, \quad j > 2.
\end{aligned}$$

Tą patį procesą galime užrašyti ir begaliniu autoregresiniu procesu

$$Y_t = \pi_1 Y_{t-1} + \pi_2 Y_{t-2} + \cdots + \varepsilon_t, \quad t \in Z.$$

Lygybėje

$$(1 - \theta L)(1 - \pi_1 L - \pi_2 L^2 - \cdots) = 1 - \phi L$$

sulygindami koeficientus prie atatinkamų L laipsnių, gauname

$$\begin{aligned}
L^1 :&\quad -\pi_1 - \theta = -\phi \quad \Longrightarrow \quad \pi_1 = \phi - \theta\\
L^2 :&\quad -\pi_2 + \theta\pi_1 = 0 \quad \Longrightarrow \quad \pi_2 = (\phi - \theta)\theta\\
&\vdots\\
L^j :&\quad -\pi_j + \theta\pi_{j-1} = 0 \quad \Longrightarrow \quad \pi_2 = (\phi - \theta)\theta^{j-1}, \quad j > 2.
\end{aligned}$$

Suskaičiuokime autokoreliacinę funkciją. Lygtį

$$Y_t = \phi Y_{t-1} + \varepsilon_t - \theta\varepsilon_{t-1}$$

padauginkime iš Y_{t-k} ir suskaičiuokime gautos lygybės abiejų pusių vidurkius:

$$EY_tY_{t-k} = \phi EY_{t-1}Y_{t-k} + E\varepsilon_t Y_{t-k} - \theta EY_{t-k}\varepsilon_{t-1}.$$

Kai $k > 1$ $E\varepsilon_t Y_{t-k} = 0$ ir $EY_{t-k}\varepsilon_{t-1} = 0$. Todėl

$$r(k) = \phi r(k-1), \quad \text{kai} \quad k > 1.$$

Kadangi

$$E\varepsilon_t Y_t = E[\varepsilon_t(\varepsilon_t + \psi_1\varepsilon_{t-1} + \psi_2\varepsilon_{t-1} + \cdots)] = \sigma^2$$

ir

$$E\varepsilon_{t-1}Y_t = E[\varepsilon_{t-1}(\varepsilon_t + \psi_1\varepsilon_{t-1} + \psi_2\varepsilon_{t-1} + \cdots)] = \psi_1\sigma^2 = (\phi-\theta)\sigma^2$$

tai,

$$\begin{aligned} r(0) &= \phi r(1) + \sigma^2 - \theta(\phi-\theta)\sigma^2 \\ r(1) &= \phi r(0) - \theta\sigma^2. \end{aligned}$$

Išsprendę šia sistema surandame

$$r(0) = \frac{1+\theta^2-2\theta\phi}{1-\phi^2}\sigma^2, \quad r(1) = \frac{(1-\theta\phi)(\phi-\theta)}{1-\phi^2}\sigma^2.$$

Iš čia randame ir autokoreliacinę funkciją:

$$\begin{aligned} \rho_k &= \phi\rho_{k-1}, \quad k > 1; \\ \rho_1 &= \frac{(1-\theta\phi)(\phi-\theta)}{1+\theta^2-2\theta\phi}. \end{aligned}$$

Dalinės autokoreliacijos vienintelė nenulinė reikšmė yra $p_1 = \rho_1$.

$ARMA$ proceso autokoreliacinė funkcija elgiasi panašiai kaip autoregresinio modelio, o dalinė autokoreliacinė funkcija – kaip slenkančiojo vidurkio.

7.7 ARIMA MODELIAI

Kaip jau minėjome, praktikoje dažnai pasitaiko kad laiko eilutė nėra stacionari, bet jos skirtumų arba aukštesnės eilės skirtumų eilutė yra stacionari. Ta aplinkybė pasitarnauja apibrėžiant ARIMA modelius. Sakysime, kad laiko eilutė $(y_1, \dots, y_T)$ tenkina $ARIMA(p, d, q)$ modelį, jei $(w_d(t), t \in Z)$ yra $ARMA(p, q)$ modelis, kai

$$w_d(t) = \Delta^d y(t), \quad t \in Z.$$

Kaip ir anksčiau, Δ yra skirtuminis operatorius

$$\Delta y(t) = y(t) - y(t-1),$$

o $\Delta^s = \Delta(\Delta^{s-1})$. Turėdami w_t prie laiko eilutės y_t grįžti galime sumuodami w_t. Tai yra

$$y_t = \sum^d w_t,$$

kai $\sum^d$ yra sumavimo operatorius. Lengva įsitikinti, kad

$$y_t = y_0 + w_1 + \cdots + w_t.$$

Sakysime, kad laiko eilutė $(y_t, t = 0, \dots, T)$ aprašoma $ARIMA(p, d, q)$ modeliu, jei ją atitinkantis procesas $(Y(t), t \in Z)$ tenkina lygtį

$$\phi_p(L)\Delta^d Y(t) = \theta_q \varepsilon_t, \quad t \in Z. \tag{7.10}$$

Prie šio modelio galime prieiti ir kiek kitu keliu. (7.8) lygtis turi stacionarų sprendinį, jei polinomas $\phi_p(z)$ neturi šaknų vienetinaime skritulyje. Tarkime, kad šis polinomas turi lygiai d šaknų $z = 1$, o visos kitos jo šaknys yra už vienetinio skritulio. Tuomet

$$\phi_p(z) = \tau_{p-d}(z)(1 - z)^d.$$

Čia τ_{p-d} yra $p - d$ eilės polinomas neturintis šaknų vienetiniame skritulyje. Dabar (ARMA2) lygtį galime perrašyti taip:

$$\tau_{p-d}(L)(1 - L)^d Y(t) = \theta_q \varepsilon_t. \tag{7.11}$$

Pastebėję, kad $I - B = \Delta$, o $(I - B)^d = \Delta^d$, gauname, kad procesas $(Y(t), t \in Z)$, tenkinantis (ARMA3) lygtį yra (ARIMA(p-d,d,q) procesas.

Kaip nustatyti ARIMA modelio eilę, t.y. parametrus (p, d, q)? Turint laiko eilutę $(y_t, t = 1, \ldots, T)$ pirmiausia reikia nustatyti homogeniškumo eilę d. Tam panaudojamas tas faktas, kad stacionaraus proceso autokoreliacinė funkcija $\rho(k)$ artėja prie nulio kai k didėja. Kai d yra nustatytas, parametrų (p, q) nustatymui galime naudoti koreliacijos koeficientus ir dalinių koreliacijų koeficientus, bet jau d kartų diferencijuoto proceso.

Tarkime, kad laiko eilutė $(y_1, \ldots, y_T)$ aprašyta modeliu $ARIMA(p, d, q)$. Tai yra, nustatyti parametrai p, d ir q, su kuriais duotoji laiko eilutė atitinka atsitiktinį procesą $(Y(t), t \in R)$ tenkinantį lygtį

$$\phi_p(L)\Delta^d Y(t) = \theta_q(L)\varepsilon_t; \tag{7.12}$$

čia $(\varepsilon_t) \in BT(0, \sigma^2)$. Duotosios laiko eilutės pagalba reikia įvertinti parametrus $\phi_1, \ldots, \phi_p$ ir $\theta_1, \ldots, \theta_q$. Tai atliekame minimizuodami paklaidas. Pažymėkime šitaip:

$$\tilde{\varepsilon}_t = \theta_q^{-1}(L)\phi_p(L)y_t. \tag{7.13}$$

Sudarykime funkciją

$$S(\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q) = \sum_{t=1}^{T} \tilde{\varepsilon}_t^2.$$

Tuomet

$$(\widehat{\phi}_1, \ldots, \widehat{\phi}_p, \widehat{\theta}_1, \ldots, \widehat{\theta}_q) = argmin\, S(\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q).$$

A priedas. Daugiamatės analizės elementai

A.1 Tiesinės algebros elementai

A.1.1 Matricos

Sutvarkytą $p \times d$ skaičių lentelę vadiname matrica. p žymi jos eilučių skaičių, d – stulpelių skaičių. Matricas žymėsime $\boldsymbol{A}, \boldsymbol{B}, ..., \boldsymbol{X}, ...$:

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1d} \\ a_{21} & a_{22} & \dots & a_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \dots & a_{pd} \end{pmatrix}.$$

Jei svarbu pažymėti tik matricos elementus, rašysime $\boldsymbol{A} = (a_{kj})$. Jei svarbi ir matricos eilė, tuomet rašysime $\boldsymbol{A} = (a_{kj}, k = 1, \dots, p; j = 1, \dots, d)$. $d \times 1$ matrica vadinama vektoriumi stulpeliu, o $1 \times p$ matrica – vektoriumi eilute. Vektorius visada reikš vektorių stulpelį, Juos žymėsime $\boldsymbol{a}, \boldsymbol{b}, \dots, \boldsymbol{x}, \cdots$:

$$\boldsymbol{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_d \end{pmatrix}.$$

Aibė visų n-mačių vektorių žymėsime R^n.

Matricos $\boldsymbol{A}$ vektorius stulpelius pažymėję $\boldsymbol{a}_1, \dots, \boldsymbol{a}_d$,

$$\boldsymbol{a}_k = \begin{pmatrix} a_{1k} \\ a_{2k} \\ \vdots \\ a_{pk} \end{pmatrix}$$

ją galime užrašyti vektoriumi eilute

$$\boldsymbol{A} = (\boldsymbol{a}_1, \boldsymbol{a}_2, \ dots, \boldsymbol{a}_d).$$

Matricos $\boldsymbol{A}$ vektorius eilutes pažymėję $\boldsymbol{b}_1, \dots, \boldsymbol{b}_p$,

$$\boldsymbol{b}_k = (a_{k1}, \dots, a_{kd}),$$

ją galime užrašyti vektoriumi stulpeliu

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \\ \vdots \\ \boldsymbol{b}_p \end{pmatrix}.$$

Tos pačios eilės matricas galima sudėti. Jei $\boldsymbol{A} = (a_{jk}), \boldsymbol{B} = (b_{jk})$ tai

$$\boldsymbol{A} + \boldsymbol{B} = (a_{jk} + b_{jk}).$$

Jei $\boldsymbol{A} = (a_{jk})$ yra $m \times n$ eilės, o matrica $\boldsymbol{B} = (b_{k,i})$ – $n \times p$ eilės, tai tas matricas galima sudauginti. Jų sandauga yra matrica

$$\boldsymbol{C} = \boldsymbol{AB} = (c_{ji}), \quad \text{kai} \quad c_{ji} = \sum_{k=1}^{n} a_{jk} b_{ki}.$$

Matricos $\boldsymbol{A} = (a_{jk})$ ir skaičiaus α sandauga vadinama matrica $\alpha\boldsymbol{A} = (\alpha a_{jk})$.

Matricos $\boldsymbol{A} = (a_{jk}, j = 1, \dots, d; k = 1, \dots, n)$ transponuota matrica yra

$$\boldsymbol{A}^\tau = (a_{kj}, k = 1, \dots, n; j = 1, \dots, d).$$

Nesunku įsitikinti, kad teisingos šios formulės:

$$(\boldsymbol{A}^\tau)^\tau = \boldsymbol{A}, \quad (\boldsymbol{A} + \boldsymbol{B})^\tau = \boldsymbol{A}^\tau + \boldsymbol{B}^\tau$$
$$(\boldsymbol{AB})^\tau = \boldsymbol{B}^\tau \boldsymbol{A}^\tau, \quad (\boldsymbol{ABC})^\tau = \boldsymbol{C}^\tau \boldsymbol{B}^\tau \boldsymbol{A}^\tau.$$

Matrica $\boldsymbol{A}$ vadinama simetrine, jei $\boldsymbol{A}^\tau = \boldsymbol{A}$.

Kvadratinė matrica $\boldsymbol{A}$ vadinama idempotentine, jei $A = A^2 = A^3 = \cdots$

Nulinė matrica yra tokia, kurios visi elementai yra nuliai. Matrica $\boldsymbol{\Lambda}$ vadinama diagonaline, jei

$$\boldsymbol{\Lambda} = \begin{pmatrix} \lambda_1 & 0 & 0 & \dots & 0 \\ 0 & \lambda_2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \lambda_n \end{pmatrix}.$$

Diagonalinę matricą žymėsime $\boldsymbol{\Lambda} = diag(\lambda_1, \dots, \lambda_n)$. Matrica

$$\boldsymbol{I} = diag(1, \dots, 1)$$

vadinama tapatingąja arba vienetine matrica.

A.1.2 Matricos pėdsakas, determinantas, rangas

Matricos pėdsakas yra jos diagonalinių elementų suma:

$$tr(\boldsymbol{A}) = \sum_j a_{jj}.$$

Teisingos šios savybės:

$$\begin{aligned} &tr(\boldsymbol{AB}) = tr(\boldsymbol{BA}),\\ &tr(\boldsymbol{I}_n) = n,\\ &tr(\alpha\boldsymbol{A}) = \alpha tr(\boldsymbol{A}),\\ &tr(\boldsymbol{A}^\tau) = tr(\boldsymbol{A}),\\ &tr(\boldsymbol{A}+\boldsymbol{B}) = tr(\boldsymbol{A}) + tr(\boldsymbol{B}). \end{aligned}$$

$n \times n$ matricos $\boldsymbol{A}$ determinantas $det(\boldsymbol{A})$ yra toks skaičius, kad

1. kai $n = 1$, tai $det(\boldsymbol{A})| = a_{11}$;
2. kai $n > 1$:
$$det(\boldsymbol{A}) = \sum_{j=1}^{n} a_{ij}(-1)^{j+i}\, det(\boldsymbol{A}_{ij}),$$
kai $\boldsymbol{A}_{ij}$ yra $(n-1)\times(n-1)$ eilės matrica gaunama iš matricos $\boldsymbol{A}$ išbraukus joje i-tą eilutę ir j-tą stulpelį. Determinantas $det(\boldsymbol{A}_{ij})$ vadinamas matricos $\boldsymbol{A}$ $n-1$ eilės minoru.

Teisingos šios savybės:

$$\begin{aligned} &det(\boldsymbol{AB}) = det(\boldsymbol{A})\, det(mB),\\ &det(diag(\lambda_j)) = \prod_j \lambda_j,\\ &det(\boldsymbol{I}_n) = 1,\\ &det(\alpha\boldsymbol{A}) = \alpha^n\, det(\boldsymbol{A}),\\ &det(\boldsymbol{A}^\tau) = det(\boldsymbol{A}). \end{aligned}$$

Be to,

- matricos stulpelių ar eilučių perstatos determinanto nekeičia;
- determinantas nesikeičia jei prie vienos eilutės (atitinkamai stulpelio) pridedame bet kurią kitų eilučių (atitinkamai stulpelių) tiesinę kombinaciją;

- matricos, kurios du stulpeliai ar dvi eilutės yra lygios, determinantas lygus nuliui;
- determinantas lygus nuliui tada ir tik tada, kai eilutės ar stulpeliai yra tiesiškai surišti.

Vektoriai $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_d \in R^n$ vadinami tiesiškai nesurištais, jei

$$\alpha_1 \boldsymbol{a}_1 + \alpha_2 \boldsymbol{a}_2 + \cdots + \alpha_d \boldsymbol{a}_d = 0$$

tada ir tik tada, kai $\alpha_1 = \cdots = \alpha_d = 0$.

Tarkime, matrica $\boldsymbol{A}$ yra $m \times n$ eilės. Matricos $\boldsymbol{A}$ stulpelių (eilučių) rangu vadinamas maksimalus tiesiškai nesurištų stulpelių skaičius.

Matricos eilučių ir stulpelių rangai visada lygūs ir tas skaičius vadinamas tiesiog matricos rangu.

Teisingos šios savybės:

1. $rank(A) \leq \min\{m, n\}$,
2. $rank(\boldsymbol{AB}) \leq \min\{rank(\boldsymbol{A}), rank(\boldsymbol{B})\}$,
3. jei $\boldsymbol{B}$ yra kvadratinė $n \times n$ eilės matrica, tai $rank(\boldsymbol{AB}) = rank(\boldsymbol{A})$,
4. jei $\boldsymbol{B}$ yra kvadratinė $m \times m$ eilės matrica, tai $rank(\boldsymbol{BA}) = rank(\boldsymbol{A})$,
5. $rank(\boldsymbol{A}) = rank(\boldsymbol{A}\boldsymbol{A}^\tau) = rank(\boldsymbol{A}^\tau \boldsymbol{A})$. Beje, $\boldsymbol{A}\boldsymbol{A}^\tau$ yra $m \times m$ eilės, o matrica $\boldsymbol{A}^\tau \boldsymbol{A}$ – $n \times n$ eilės.

A.1.3 Atvirkščioji matrica

Tarkime, $\boldsymbol{A}$ – kvadratinė $n \times n$ eilės matrica.

Matrica $\boldsymbol{A}$ yra neišsigimusi, jei ji yra maksimalaus rango, t.y. $rank(\boldsymbol{A}) = n$.

Matricos $\boldsymbol{A}$ atvirkštine vadinama tokia matrica $\boldsymbol{A}^{-1}$, kad $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{I}$.

Kiekvienai neišsigimusiai matricai egzistuoja atvirkštinė.

Teisingos šios savybės:

- $det(\boldsymbol{A}^{-1}) = 1/\, det(\boldsymbol{A})$,
- $(\boldsymbol{A}^{-1})^{-1} = \boldsymbol{A}$,
- $(\boldsymbol{A}^{-1})^\tau = (\boldsymbol{A}^\tau)^{-1}$,
- jei egzistuoja $\boldsymbol{A}^{-1}$ ir $\boldsymbol{B}^{-1}$, tai $(\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$.

A.1.4 Tikrinės reikšmės ir tikriniai vektoriai

Tarkime, $\boldsymbol{A}$ yra kvadratinė $n \times n$ eilės matrica. Vektorius $\boldsymbol{a} \in R^n \neq 0$ vadinamas tikriniu matricai $\boldsymbol{A}$, o kompleksinis skaičius λ – tikrine reikšme, jei

$$\boldsymbol{A}\boldsymbol{a} = \lambda \boldsymbol{a}.$$

Lygtis

$$det(\boldsymbol{A} - \lambda \boldsymbol{I}) = 0$$

vadinama charakteringąja lygtimi. Jos šaknys yra tikrinės matricos $\boldsymbol{A}$ reikšmės.

Matricos $\boldsymbol{A}$ ir $\boldsymbol{B}$ vadinamos ekvivalenčiomis, jei egzistuoja tokia matrica $\boldsymbol{C}$, kad

$$\boldsymbol{B} = \boldsymbol{C}^{-1}\boldsymbol{A}\boldsymbol{C}.$$

Ekvivalenčios matricos turi tas pačias tikrines reikšmes, nes

$$\begin{aligned}
&det(\boldsymbol{B} - \lambda \boldsymbol{I}) = det(\boldsymbol{C}^{-1}\boldsymbol{A}\boldsymbol{C} - \lambda \boldsymbol{I}) = det(\boldsymbol{C}^{-1}\boldsymbol{A}\boldsymbol{C} - \lambda \boldsymbol{C}^{-1}\boldsymbol{I}\boldsymbol{C}) = \\
&det(\boldsymbol{C}^{-1}(\boldsymbol{A} - \lambda \boldsymbol{I})\boldsymbol{C}) = det(\boldsymbol{C}^{-1}\, det(\boldsymbol{A} - \lambda \boldsymbol{I})\, det(\boldsymbol{C}) = \\
&det(\boldsymbol{A} - \lambda \boldsymbol{I})\, det(\boldsymbol{C})\, det(\boldsymbol{C})^{-1} = det(\boldsymbol{A} - \lambda \boldsymbol{I}).
\end{aligned}$$

Skirtingas tikrines matricos reikšmes atitinka tiesiškai nesurišti tikriniai vektoriai.

Jei matrica $\boldsymbol{A}$ turi n skirtingų realiųjų tikrinių reikšmių, tai ją galima išreikšti

$$\boldsymbol{A} = \boldsymbol{C}^{-1}\boldsymbol{\Lambda}\boldsymbol{C},$$

kai $\boldsymbol{\Lambda}$ - diagonalinė, o matrica $\boldsymbol{C}$ neišsigimusi.

A.1.5 Simetrinės matricos

Matrica $\boldsymbol{A}$ vadinama simetrine, jei $\boldsymbol{A} = \boldsymbol{A}^\tau$.

Su bet kuria matrica $\boldsymbol{A}$, matrica $\boldsymbol{A}^\tau \boldsymbol{A}$ – simetrinė.

Simetrinės matricos tikriniai vektoriai atitinkantys skirtingas tikrines reikšmės yra ortogonalūs.

Matrica, kurios stulpeliai yra tarpusavyje ortogonalūs, vadinama ortogonaląja. Jei $\boldsymbol{O}$ – ortogonalioji matrica, tai

$$\boldsymbol{O}^\tau \boldsymbol{O} = \boldsymbol{I}$$

ir

$$O^\tau = O^{-1}.$$

Ortogonaliosios matricos determinantas visada lygus arba $+1$, arba -1. Kiekvieną simetrinę matricą mA atitinka tokia ortogonalioji matrica $\boldsymbol{O}$, kad

$$\boldsymbol{O}^\tau \boldsymbol{A}\boldsymbol{O} = diag(\lambda_1, \dots, \lambda_n);$$

čia $\lambda_1, \dots, \lambda_n$ – matricios $\boldsymbol{A}$ tikrinės reikšmės.

A.1.5.1 Teigiamai apibrėžtos matricos

Simetrinė $n \times n$ matrica $\boldsymbol{A}$ vadinama simetrine, jei

$$\boldsymbol{x}^\tau \boldsymbol{A}\boldsymbol{x} > 0$$

su kiekvienu nenuliniu $\boldsymbol{x} \in R^n$.

Kvadratinė $n \times n$ matrica vadinama neneigiamai apibrėžta, jei su visais $\boldsymbol{x} \in R^n$

$$\boldsymbol{x}^\tau \boldsymbol{A}\boldsymbol{x} \geq 0.$$

Pavyzdžiui, matrica $\boldsymbol{A}^\tau \boldsymbol{A}$ visada neneigiamai apibrėžta.

Simetrinių matricų aibėje galima apibrėžti dalinį sutvarkymą. Sakysime, kad $\boldsymbol{A} \geq 0$, jei $\boldsymbol{A}$ neneigiamai apibrėžta, ir $\boldsymbol{A} > 0$, jei ji teigiamai apibrėžta. Sąryšis $\boldsymbol{A} \geq \boldsymbol{B}$ (atitinkamai $\boldsymbol{A} > \boldsymbol{B}$) reiškia, kad $\boldsymbol{A} - \boldsymbol{B} \geq 0$ (atitinkamai $\boldsymbol{A} - \boldsymbol{B} > 0$).

A.2 Vektorinio argumento funkcijų analizė

Funkcijos $f : R^n \to R$ išvestine taške $\boldsymbol{x} = (x_1, \dots, x_n)^\tau)$ vadinamas vektorius eilutė

$$f'(x) = \Big(\frac{\partial f(\boldsymbol{x})}{\partial x_1}, \dots, \frac{\partial f(\boldsymbol{x})}{\partial x_n}\Big).$$

Funkcijos $F = (F_1, \dots, F_m)^\tau : R^n \to R^m$ išvestine taške $\boldsymbol{x} = (x_1, \dots, x_n)^\tau$ vadinama $m \times n$ matrica (dar vadinama Jakobio matrica)

$$F'(\boldsymbol{x}) = \begin{pmatrix} \partial F_1(\boldsymbol{x})/\partial x_1 & \dots & \partial F_1(\boldsymbol{x})/\partial x_n \\ \vdots & \ddots & \vdots \\ \partial F_m(\boldsymbol{x})/\partial x_1 & \dots & \partial F_m(\boldsymbol{x})/\partial x_n \end{pmatrix}$$

Pateiskime kelet svarbesnių pavyzdžių.

1. Jei $f(\boldsymbol{x}) = \boldsymbol{a}^\tau \boldsymbol{x}$, kai $\boldsymbol{a} = (a_1, \ldots, a_n)^\tau$, tai
$$f'(\boldsymbol{x}) = \boldsymbol{a}^\tau.$$

2. Jei $f(\boldsymbol{x}) = \boldsymbol{x}^\tau \boldsymbol{A} \boldsymbol{x}$, kai $\boldsymbol{A}$ – $n \times n$ matrica, tai
$$f'(\boldsymbol{x}) = \boldsymbol{x}^\tau (\boldsymbol{A} + \boldsymbol{A}^\tau).$$
Atskiru atveju, kai $\boldsymbol{A}$ simetrinė,
$$f'(\boldsymbol{x}) = 2\boldsymbol{x}^\tau \boldsymbol{A}.$$

3. Jei $F(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}$, kai $\boldsymbol{A}$ yra $m \times n$ matrica, tai
$$F'(\boldsymbol{x}) = \boldsymbol{A}.$$

B priedas. Būtinosios tikimybinės sąvokos

Dauguma ekonominių dydžių yra iš prigimties atsitiktiniai. Todėl bet kuriuose ekonominiuose duomenyse neišvengiamai slypi neapibrėžtumai. Įprasta, kad pradedant kalbą apie neapibrėžtumus ir jų tyrimą, pradedama magiškuoju trejetu $(\Omega, \mathcal{F}, P)$, kuris vadinamas tikimybine erdve (statistikoje – statistikiniu eksperimentu). Aibė Ω aprašo viską, kas tiriamame reiškinyje iššaukia neapibrėžtumus. $\mathcal{F}$ aibės Ω poaibių σ-algebra. Jos elementai aprašo galimus įvykius esant neapibrėžtumams Ω. P yra tikimybinis matas, apibrėžtas įvykiams $A \in \mathcal{F}$. $P(A)$ vadiname įvykio A tikimybe. Dažniausiai magiškojo trejeto $(\Omega, \mathcal{F}, P)$ prigimtis nėra esminis dalykas. Ypač, jei jis siejamas su kokiu nors eksperimentu. Tada svarbiau nagrinėti to eksperimento metu stebimą dydį.

B.1 Atsitiktiniai dydžiai ir jų skirstiniai

Atvaizdis $X : \Omega \to R$ vadinamas atsitiktiniu dydžiu, jei

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}$$

su bet kuria Borelio aibe $B \subset R$ (pakanka su bet kuria aibe, pavidalo $[a, b), (-\infty, b), [a, +\infty), (-\infty, +\infty), \quad a, b \in R$) Atsitiktinio dydžio X pasiskirstymo funkcija vadinama funkcija

$$F(x) = P(\omega : X(\omega) \leq x), \quad x \in R.$$

Ekonometrijoje paprastai nagrinėjami diskretūs ir tolydūs atsitiktiniai dydžiai. Diskretūs – įgyjantys tik baigtinį arba suskaičiuojamą skaičių reikšmių. Jie pilnai aprašomi įgyjamomis reikšmėmis ir atitinkamomis tų reikšmių įgyjimo tikimybėmis. Daugelis ekonominių dydžių yra diskretūs. Pavyzdžiui, vaikų skaičius šeimoje, apsipirkimų per mėnesį skaičius ir pan.

Jei X yra diskretus atsitiktinis dydis su reikšmėmis $x_1, x_2, \ldots$ ir

$$P(X = x_k) = p_k, \quad k = 1, 2, \ldots$$

tai jo pasiskirstymo funkcija yra

$$F(x) = \sum_{k:x_k \le x} p_k, \quad x \in R.$$

Tolydus atsitiktinis dydis fiksuotą reikšmę įgyja tik su nuline tikimybe. Jo atžvilgiu galima aprašyti tik tikimybę, kad reikšmės priklausys kuriam nors intervalui. Tam patogiausia pasinaudoti tankio funkcija. Tankio funkcija vadinama bet kuri neneigiama integruojama funkcija, kurios integralas lygus vienam. Tai yra, funkcija $f(x), x \in R$ yra tankio funkcija, jei

$$f(x) \ge 0 \quad \text{su visais} \quad x \in R$$

ir

$$\int_{-\infty}^{\infty} f(x)dx = 1.$$

Tolydų atsitiktinį dydį X atitinka tankio funkcija f, jei jo pasiskirstymo funkcija yra

$$F(x) = \int_{-\infty}^{x} f(t)dt, \quad x \in R.$$

Atsitiktiniai dydžiai $X_1, \ldots, X_n$ vadinami nepriklausomais, jei su bet kuriomis Borelio aibėmis $B_1, \ldots, B_n$

$$P(\omega : X_1(\omega) \in B_1, \ldots, X_n(\omega) \in B_n) =$$
$$P(\omega : X_1(\omega \in B_1) \cdots P(\omega : X_n(\omega) \in B_n).$$

Atsitiktiniai dydžiai X_1, X_2 yra vienodai pasiskirstę (žymėsime $X_1 \sim X_2$), jei jų pasiskirstymo funkcijos sutampa, t.y.

$$P(\omega : X_1(\omega) \le x) = P(\omega : X_2(\omega) \le x) \quad \text{su visais} \quad x \in R.$$

Du atsitiktiniai dydžiai X_1 ir X_2 vadinami ekvivalenčiais (žymėsime $X_1 \sim_{b.t.} X_2$, jei

$$P(\omega : X_1(\omega) \ne X_2(\omega)) = 0.$$

Viena iš svarbiausių atsitiktinio dydžio charakteristikų yra jo vidurkis. Diskretaus atsitiktinio dydžio X vidurkis yra

$$EX = \sum_x xf(x) = \sum_k x_k f(x_k) = \sum_k x_k P(X = x_k),$$

kai $x_1, x_2, \ldots$ yra įgyjamos reikšmės, o $f(x_k)$ atitinkamos tų reikšmių tikimybės.

Tolydaus atsitiktinio dydžio X, aprašomo tankio funkcija f, vidurkis yra

$$EX = \int_{-\infty}^{\infty} x f(x) dx.$$

Dauguma funkcijų nuo atsitiktinio dydžio yra taip pat atsitiktiniai dydžiai. Tarkime, $g : R \to R$ tokia funkcija (pavyzdžiui tolydi), kad $g(X)$ yra atsitiktinis dydis. Tuomet jos vidurkis yra

$$Eg(X) = \int_{-\infty}^{\infty} g(x) f(x) dx,$$

kai X yra tolydus su tankio funkcija f ir

$$Eg(X) = \sum_k g(x_k) f(x_k),$$

kai tas dydis yra diskretus.

Atsitiktinų dydžių tiesinė kombinacija yra atsitiktinis dydis. Tai yra, jei $X_1, \ldots, X_n$ – atsitiktiniai dydžiai, $\lambda_1, \ldots, \lambda_n \in R$, tai $\sum_{k=1}^{n} \lambda_k X_k$ yra atsitiktinis dydis. Jo vidurkį apskaičiuojame pagal formulę

$$E \sum_k \lambda_k X_k = \sum_k \lambda_k E X_k,$$

jei atitinkami vidurkiai egzistuoja.

Kita labai svarbi atsitiktinio dydžio charakteristika yar dispersija

$$var(X) = \sigma^2 = E(X - EX)^2.$$

Ji aprašo vidutinį kavdratinį atsitiktinio dydžio nuokripį nuo savo vidurkio.

B.2 Atsitiktiniai vektoriai

Atvaizdis $X = (X_1, \ldots, X_d) : \Omega \to R^d$ vadinamas atsitiktiniu vektoriumi, jei

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) = (X_1(\omega), \ldots, X_d(\omega)) \in B\} \in \mathcal{F}$$

su bet kuria Borelio aibe $B \subset R^d$ (pakanka su bet kuria aibe $B_1 \times \cdots \times B_d$, kai $B_1, \ldots, B_d$ yra realiųjų skaičių Borelio aibės arba aibės pavidalo

$[a, b), (-\infty, b), [a, +\infty), (-\infty, +\infty), \quad a, b \in R)$ Atsitiktinio vektoriaus X pasiskirstymo funkcija vadiname funkciją

$$F(x_1, \dots, x_d) = P(\omega : X_1(\omega) \le x_1, \dots, X_d \le x_d), \quad x = (x_1, \dots, x_d) \in R^d.$$

Atsitiktinis vektorius $X = (X_1, \dots, X_d)$ vadinamas tolydžiu, jei egzistuoja tokia integruojama teigiama funkcija $f(x_1, \dots, x_d)$, su kuria

$$F(x_1, \dots, x_d) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} f(t_1, \dots, t_d) dt_1 \cdots dt_d.$$

Atsitiktinis vektorius $X = (X_1, \dots, X_d)$ yra diskretus, jei visos jo komponentės $X_1, \dots, X_d$ – diskretūs atsitiktiniai dydžiai. Atsitiktinio vektoriaus $X = (X_1, \dots, X_d)$, kurio pasiskirstymo funkcija yra $F(x_1, \dots, x_d)$ kiekviena komponentė X_k yra atsitiktinis dydis. Jo marginalinė pasiskirstymo funkcija yra

$$F_k(x_k) = F(+\infty, \cdots, +\infty, x_k, +\infty, \dots, +\infty), \quad x_k \in R.$$

Tolydaus atsitiktinio dydžio X_k marginalusis tankis yra tokia tankio funkcija $f_k(t)$ su kuria

$$F_k(x_k) = \int_{-\infty}^{x_k} f_k(t) dt.$$

Marginalinį tankį galime surasti pasinaudoję formule

$$f_k(x_k) =$$
$$\int_R \cdots \int_R f(x_1, \dots, x_{k-1}, x_k, x_{k+1}, \dots, x_d) dx_1 \cdots dx_{k-1} dx_{k+1} \dots dx_d.$$

Jei atsitiktiniai dydžiai $X_1, \dots, X_d$ yra nepriklausomi, tai vektoriaus

$$(X_1, \dots, X_d)$$

tankio funkcija yra lygi marginalinių tankio funkcijų sandaugai:

$$f(x_1, \dots, x_d) = f_1(x_1) \cdots f_d(x_d).$$

Jei $X = (X_1, \dots, X_n)$ yra atsitiktinis vektorius, tai jo vidurkis yra vektorius $EX = (EX_1, \dots, EX_n)$.

Dviejų atsitiktinių dydžių kovariacija

$$cov(X, Y) = E(X - EX)(Y - EY) = EXY - EXEY.$$

Diskretiems dydžiams X, Y

$$Eg(X, Y) = \sum_x \sum_y g(x, y) f(x, y),$$

tankio funkcija $f(x, y) = P(X = x, Y = y)$. Koreliacija

$$\rho = \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}}.$$

B.3 Sąlyginiai skirstiniai

Įvykio $A \in \mathcal{F}$ sąlyginė tikimybė su sąlyga, kad įvyko įvykis $B \in \mathcal{F}$ apskaičiuojama pagal formulę

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Norėdami apibrėžti sąlyginę tikimybę, kai sąlyga susdaryta iš bet kokio įvykių skaičiaus, turime pritaikyti kitus argumentus. Lengviausia tai padaryti pradėjus nuo sąlyginio vidurkio sąvokos. Tarkime, $\mathcal{F}_0 \subset \mathcal{F}$ yra duota σ-algebra. Atsitiktinio dydžio X sąlyginis vidurkis atžvilgiu σ-algebros $\mathcal{F}_0$ yra toks $\mathcal{F}_0$-matus atsitiktinis dydis $E(X|\mathcal{F}_0)$, kuriam

$$E(E(X|\mathcal{F}_0)\chi_F) = EX\chi_F, \quad \text{su kiekviena aibe } \; F \in \mathcal{F}_0.$$

Čia χ_F yra aibės F indikatorinė funkcija:

$$\chi_F(\omega) = \begin{cases} 1, & \text{kai} \quad \omega \in F; \\ 0, & \text{kai} \quad \omega \notin F. \end{cases}$$

Sąlyginis vidurkis $E(X|Y) = E(X|\mathcal{F}_0)$, kai $\mathcal{F}_0$ yra mažiausia σ-algebra atžvilgiu kurios yra matus atsitiktinis dydis Y. Įvykio $A \in \mathcal{F}$ sąlyginė tikimybė atžvilgiu $\mathcal{F}_0$ yra

$$P(A|\mathcal{F}_0) = E(\chi_A|\mathcal{F}_0).$$

Svarbu įsidėmėti, kad sąlyginis vidurkis ir sąlyginė tikimybė atžvilgiu kurios nors σ-algebros yra atsitiktinis dydis.

Jei atsitiktinis vektorius (Y, X) yra aprašomas tankio funkcija $f(y, x)$, tai sąlygine tankio funkcija, kai fiksuotas dydis $X = x$ yra

$$f(y|x) = \frac{f(y,x)}{f_X(x)},$$

kai $f_X(x)$ – atsitiktinio dydžio X marginalioji tankio funkcija. Tuomet

$$P(a < Y \le b|X = x) = \int_a^b f(y|x)dy, \quad E(Y|X = x) = \int_{-\infty}^{\infty} yf(y|x)dy.$$

B.4 Svarbiausi atsitiktiniai dydžiai

- *Bernulio atsitiktinis dydis*

Tai pats paprasčiausias diskretus atsitiktinis dydis. Atsitiktinis dydis X turintis tik dvi galimas reikšmes, 0 ir 1 vadinamas Bernulio atsitiktiniu dydžiu. Tikimybė, kad tas dydis įgis reikšmę 1 lygi p, o $P(X = 0) = 1 - p$. Bernuli atsitiktinis dydis aprašo vieno kurio nors įvykio „sėkmę" – „nesekmę". Tai gali būti, tarkime, vartotojo sprendimas pirkti kurią nors prekę; banko sprendimas apie kredito išdavimą; darbdavio sprendimas apie priėmimą į darbą ir t.t.

- *Puasono*

Puasono atsitiktinis dydis Y – diskretus atsitiktinis dydis, kurio reikšmės yra natūralieji skaičiai $1, 2, 3, \dots$ ir atitinkamos tikimybės

$$f_\lambda(\{k\}) = P(X = k) = \frac{\lambda^k}{k!}e^{-\lambda}$$

su kiekvienu $k \in N$.

- *Tolygus*

Atsitiktinis dydis X yra tolygus intervale $[a, b]$, jei jo tankio funkcija yra

$$f(u) = \begin{cases} 1/(b-a), & \text{kai} \quad a \le u \le b; \\ 0, & \text{kitur.} \end{cases}$$

- *Normalinis*

Atsitiktinis dydis X yra normalinis su parametrais a ir σ^2, jei jo tankio funkcija turi pavidalą

$$f(u) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-(u-a)^2/2\sigma^2\}, \quad u \in R.$$

Jei $a = 0$, o $\sigma^2 = 1$ tai X vadinamas standartiniu normaliuoju.

- *Atsitiktinis dydis* χ^2

$Z_1, Z_2, \ldots, Z_m$ – nepriklausomi standartiniai normaliniai atsitiktiniai dydžiai. Tuomet atsitiktinis dydis

$$V = Z_1^2 + Z_2^2 + \cdots + Z_m^2 \sim \quad \chi_m^2$$

yra χ^2 su m laisvės laipsniais. Jo skirstinį taip pat priimta vadinti χ^2 su m laisvės laipsniais. chi_m^2 atsitiktinio dydžio tankio funkcija yra

$$f(x) = e^{-x/2} x^{d/2-1} \frac{1}{2^{d/2}\Gamma(d/2)}, \quad x > 0.$$

Čia Γ yra gamma funkcija,

$$\Gamma(n) = \int_0^\infty e^{-t} t^{n-1} dt.$$

- *Stjudento arba t*

Jei Z yra standartinis normalinis, o V yra χ_m^2 ir jie yra nepriklausomi, tuomet

$$t = \frac{Z}{\sqrt{V/m}} \sim t_m$$

yra Stjudento arba t atsitiktinis dydis su m laisvės laipsnais. Jo tankis yra

$$f(x) = \frac{\Gamma((n+1)/2)}{\Gamma(n/2)\Gamma(1/2)} \Big(1 + \frac{x^2}{n}\Big)^{-(n+1)/2}, \quad x \in R.$$

- *Atsitiktinis dydis F*

Jei $V_1 \sim \chi_m^2$, o $V_2 \sim \quad \chi_n^2$ ir yra nepriklausomi, tai

$$F = \frac{V_1/m}{V_2/n} \sim \quad F_{m,n}$$

yra F (Fišerio) atsitiktinis dydis su (m, n) laisvės laipsniais.

B.5 DAUGIAMATIS GAUSO SKIRSTINYS

Atsitiktinis vektorius $Y = (Y_1, \ldots, Y_d)^\tau$ vadinamas neišsigimusiu gausiniu (normaliuoju), jei jo tankio funkcija yra

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\Big\{ - frac12(\boldsymbol{x} - \boldsymbol{a})^\tau \Sigma^{-1}(\boldsymbol{x} - \boldsymbol{a})\Big\};$$

čia $\boldsymbol{a} \in R^d$ – bet kuris vektorius, Σ – simetrinė teigiamai apibrėžta $d \times d$ matrica. Vektorius $\boldsymbol{a}$ ir matrica Σ yra parametrai ir sutrumpintai rašoma $Y \sim N(\boldsymbol{a}, \Sigma)$ arba $Y \sim N_d(\boldsymbol{a}, \Sigma)$, jei reikia pažymėti dimensiją d. Atsitiktinis vektorius Y vadinamas standartiniu gausiniu (normaliniu), jei jis yra gausinis su parametrais $\boldsymbol{a} = 0$ ir $\Sigma = \boldsymbol{I}$.

Išvardinsime keletą gausinių vektorių savybių,

1. Jei $Y \sim N(\boldsymbol{a}, \Sigma)$, tai $EY = \boldsymbol{a}$, $cov(Y) = \Sigma$.

2. Jei Y, Z – nepriklausomi gausiniai vektoriai, tai ir vektorius $(Y, Z)^\tau$ – gausinis.

3. Jei vektorius $(Y, Z)^\tau$ gausinis ir jo komponentės Y ir Z nekoreliuoti atsitiktiniai vektoriai, tai jos yra ir nepriklausomos.

4. Tarkime, $\boldsymbol{B} : R^d \to R^p$ – tiesinė transformacija, kurią atitinka matrica $\boldsymbol{B}$ ir vektorius $\boldsymbol{b} \in R^p$. Jei $Y \sim N(\boldsymbol{a}, \Sigma)$, tai

   $$\boldsymbol{B}Y + \boldsymbol{b} \;\sim\; N(\boldsymbol{B}\boldsymbol{a} + \boldsymbol{b}, \boldsymbol{B}\Sigma\boldsymbol{B}^\tau).$$

   Kitais žodžiais tariant, afininė transformacija gausinį atsitiktinį vektorių perveda į gausinį. Atskiru atveju gauname, kad bet kuri gausinių atsitiktinių vektorių tiesinė kombinacija yra gausinis atsitiktinis vektorius, o ortogonalioji standartinio gausinio vektoriaus transformacija vėl yra standartinis gausinis vektorius.

5. Tarkime, $Y \sim N(\boldsymbol{a}, \Sigma)$. Kadangi matrica Σ simetrinė ir teigiamai apibrėžta, tai visos jos tikrinės reikšmės $\lambda_1, \ldots, \lambda_d$ yra neneigiamos ir egzistuoja tokia ortogonalioji matrica $\boldsymbol{O}$, kad

   $$\Lambda = diag(\lambda_1, \ldots, \lambda_d) = \boldsymbol{O}^\tau \Sigma \boldsymbol{O}.$$

   Be to, vektorius $X = \boldsymbol{O}^\tau Y - \boldsymbol{O}^\tau \boldsymbol{a}$ yra gausinis su nuliniu vidurkiu ir kovariacija Λ. Dar daugiau, atsitiktinio vektoriaus X komponentės yra nepriklausomi gausiniai atsitiktiniai dydžiai.

6. Bet kuris gausinis atsitiktinis vektorius gali būti gaunamas iš standartinio normalinio vektoriaus afininės transformacijos pagalba.

7. Tarkime, $\boldsymbol{\varepsilon}$ – standartinis normalinis atsitiktinis vektorius, $X = \boldsymbol{A\varepsilon} + \boldsymbol{a}$, $Y = \boldsymbol{B\varepsilon} + \boldsymbol{b}$, kai $\boldsymbol{A} : R^d \to R^q$, $\boldsymbol{B} : R^d \to R^s$ – tiesininės transformacijos. Tuomet

$$cov(X, Y) = \boldsymbol{AB}^\tau.$$

Taigi atsitiktiniai vektoriai X ir Y nepriklausomi tada ir tik tada, kai $\boldsymbol{AB}^\tau = 0$.

8. Tarkime, $\boldsymbol{M}$ – idempotentinė $d \times d$ matrica rango r, $\boldsymbol{\varepsilon}$ – erdvės R^d standartinis normalinis atsitiktinis vektorius. Tuomet atsitiktinis dydis $\boldsymbol{\varepsilon}^\tau \boldsymbol{M\varepsilon}$ turi χ^2_r skirstinį su r laisvės laipsniais.

9. Jei $Y \sim N(\boldsymbol{a}, \Sigma)$ ir matrica σ neišsigimusi, tai

$$X = (Y - \boldsymbol{a})^\tau \Sigma^{-1}(Y - \boldsymbol{a}) \quad \sim \quad \chi^2_d.$$

10. Tarkime, $Z = (X, Y)^\tau$ normailnis atsitiktinis vektorius, $EX = \boldsymbol{a}_x, EY = \boldsymbol{a}_y$ ir matrica V_Y neišsigimusi. Tuomet

$$E(X|Y) = cov(X, Y)V_Y^{-1}(Y - \boldsymbol{a}_y) + \boldsymbol{a}_x.$$

Literatūros sąrašas

[1] Carter Hill, William Griffiths, George Judge (1997). *Undergraduate Econometrics*. John Wiley & Sons, Inc.

[2] Goldberger A. (1990). *A cours in Econometrics*. Cambridge University Press

[3] Greene, W.H. (1997) *Econometric Analysis*, 3rd edition, Prentice-Hall.

[4] Johnston J, and DiNardo J. (1997) *Econometric methods*, 4th edition. McGraw-Hill.

[5] Hamilton J.D. (1994). *Time Series Analysis*. Princeton University Press

[6] Kennedy P. (1998). *A Guide to Econometrics*. Blackwell Publishers

[7] Melenberg B. (2001). *Lecture Notes. Orientation Econometrics*. Tilburg University Press