

eman ta zabal zazu



Universidad
del País Vasco

Euskal Herriko
Unibertsitatea

Gretl-en Bidezko Erregresio Analisia

Egileak:

M. Victoria Esteban

M. Paz Moral

Susan Orbe

Marta Regúlez

Ainhoa Zarraga

Marian Zubia

Ekonomia Aplikatua III Saila

Ekonometria eta Estatistika

Ekonomia eta Enpresa Zientzien Fakultatea

UPV/EHU

Aurkibidea

1	Gretl eta Ekonometria	1
1.1	Sarrera	2
1.2	Zer da Ekonometria?	2
1.3	Ikerketa ekonometriko bat	6
1.4	Datuak eta bere erabilera	7
1.4.1	Datuen jatorriak	8
1.4.2	Software ekonometrikoa	9
1.5	Gretlerako sarrera	10
1.5.1	Aldagai baten analisi deskribatzailea	15
1.5.2	Aldagaien arteko erlazioak	20
2	Erregresio linealeko eredu bakuna	25
2.1	Sarrera	26
2.2	Oinarrizko hipotesiak	29
2.3	Karratu Txikiaren Arruntetako metodoa	33
2.4	Doikuntzaren egokitasuna	46
2.5	Esanguratasun analisia eta konfidantza tarteak	47
2.6	Laburpena. Eraitzen aurkezpena	50
3	Erregresio linealeko eredu orokorra	51
3.1	Sarrera	52
3.2	Karratu Txikiaren Arruntetako estimazioa	53
3.2.1	Koefizienteak	57
3.2.2	Desbideratze tipikoak eta konfidantza tarteak	59
3.2.3	Banakako eta baterako esanguratasunak	61
3.3	Doikuntzaren ontasuna eta ereduaren sailkapena	63

4	Murrizketa linealen kontrasteak eta aurreanak	71
4.1	Sarrera	72
4.2	Murrizketa linealen kontraste orokorra	72
4.3	Murrizketei baldintzaturiko Karratu Txikiaren Arruntetako estimatzailea . .	74
4.4	Murrizketa linealen kontrasteak Gretl erabiliz	76
4.5	Puntuzko eta tarte-zko aurreanak	80
4.6	Puntuzko eta tarte-zko aurreanak Gretl erabiliz	81
5	Zehazpen errorea	87
5.1	Sarrera	88
5.2	Aldagai nabariaren omisioaren ondorioak	88
5.3	Aldagai ez nabariaren barnerapenaren ondorioak	94
6	Kolinealtate anizkoitza	97
6.1	Sarrera	98
6.2	Kolinealtate anizkoitz zehatza	98
6.3	Kolinealtate anizkoitz altua	100
7	Aldagai koalitatiboak	105
7.1	Sarrera	106
7.2	Aldagai koalitatibo bakarra duen eredua	106
7.3	Aldagai koalitatibo bi edo gehiago dituen eredua	116
7.4	Egitura aldaketa	119
A	Eranskina: Errepasoak	123
A.1	Probabilitatearen errepasoak	123
A.1.1	Zorizko aldagaia edo aldagai aleatorioa	123
A.1.2	Zorizko bi aldagai edo gehiago	125
A.1.3	Zenbait probabilitate banaketa	128
A.2	Inferentzia estatistikoaren errepasoak	130
A.2.1	Estimazioa	131
A.2.2	Hipotesien kontrasteak	135
B	Eranskina: Gretl euskeraz instalatzeko jarraibideak	141

Irudiak

1.1	Sakabanatze diagrama	3
1.2	Gretl-eko pantaila nagusia	10
1.3	Datuen edizioa	11
1.4	Datuen informazioa	12
1.5	Datuen erakuspena	13
1.6	Datuen ezaugarriak editatzen	14
1.7	Datu-fitxategiaren erakuspena	14
1.8	Maiztasun erlatiboko histograma	15
1.9	Ikonoen ikuspegia	16
1.10	Asimetria motak	19
1.11	X-Y grafikoa	21
1.12	Sakabanatze diagrama	22
2.1	Datuen bilakaera	27
2.2	Datu-fitxategiaren aukerapena	27
2.3	Datuen erakuspena	28
2.4	(Y, X) ren sakabanatze diagrama	29
2.5	Bilboko etxebizitzaren prezioa daukaten azalera bizigarriarekiko	30
2.6	Eredua: $Y_i = \alpha + \beta \times 5 + u_i$, non $S_X^2 = 0$ den	31
2.7	Perturbazioen bariantza laginean zehar	31
2.8	Perturbazioen bariantza aldagai azaltzailearekiko	32
2.9	Erregresio eredu bakuna	33
2.10	Populazio eta lagin erregresio funtzioak	35
2.11	Karratu Txikiaren Arruntetako estimatzailea	37
2.12	Ereduaren zehazpena	37
2.13	Ikono bezala gordetzen	38

2.14	Estimazio emaitzak ikonoetatik berreskuratzen	39
2.15	Ikonoen ikuspegia	39
2.16	Estimatutako vs benetako aldagai azaldua	40
2.17	Estimazio emaitzen erakuspena	40
2.18	Hondarren grafikoa	41
2.19	Estimazio emaitzen erakuspena	41
2.20	Estimatutako balioak gordetzen	42
2.21	Gordetako datuen erakuspena	42
2.22	Erabaki araua	47
3.1	Hondarrak laginean zehar	55
3.2	Hondarrak sqft aldagai azaltzailearekiko	55
3.3	Benetako eta estimatutako prezioak laginean zehar	56
3.4	Benetako eta estimatutako prezioak vs sqft	56
4.1	Murrizketa linealen kontrasteak	77
4.2	Murrizketa linealen zehazpena	77
4.3	Murrizketa linealen kontrastearen emaitza	78
4.4	Baterako esanguratasunaren murrizketen zehazpena	79
4.5	Baterako esanguratasun kontrastearen emaitza	79
4.6	Balioak editatzea	82
4.7	Behaketak erantsi	82
4.8	Eransteko behaketen kopurua	82
4.9	Datu berrien ikuspena	83
4.10	Ibiltartearen ezarpena	83
4.11	Lagin tartearen ezarpena	84
4.12	Aurresanak	84
4.13	Aurresan tartearen ezarpena	85
4.14	Puntuzko eta tartezko aurresanen balioak	85
4.15	Aurresanen grafikoa	86
5.1	D Ereduko hondarren behaketen grafikoa	91
5.2	D Ereduko hondarren grafikoa SQFT aldagaiarekiko	92
5.3	A Ereduko hondarren behaketen grafikoa	94

5.4	A Ereduko hondarren grafikoa SQFT aldagaiarekiko	94
7.1	Jatorri aldaketa	113
7.2	Jatorri eta malda aldaketa	115
A.1	<i>Normalaren</i> dentsitate funtzioa eta histograma.	124
A.2	Banaketa normalaren adibideak	125
A.3	Aldagai biko banaketa normala.	126
A.4	Chi karratu banaketaren dentsitate funtzioa	129
A.5	Snedecorren-F banaketaren dentsitate funtzioa	129
A.6	Studenten-t banaketaren dentsitate funtzioa	130
A.7	Estimatzailen alborapena	133
A.8	Estimatzailen banaketen adibideak	134
A.9	Alde biko kontrastearen eskualde kritikoak	137
A.10	Alde bateko kontrastearen eskualde kritikoak	139

Taulak

1.1	Etxebizitzaren prezioen datuak	3
1.2	Etxebizitzaren prezioaren maiztasunak	16
1.3	Prezio aldagaiaren estatistiko nagusiak	17
1.4	Aldagaien estatistiko nagusiak	20
1.5	Aldagaien koerlazio matrizea	23
2.1	<i>“2gaia-datuak.gdt”</i> fitxategiko datuak	28
2.2	Aldagai guztien estatistiko nagusiak	43
2.3	Aldagai guztien koerlazio matrizea	44
2.4	KTako estimatzailearen bariantza eta kobariantza matrizea	46
2.5	Tartezko estimazioa	49
3.1	KTako estimatzaileen bariantza eta kobariantza matrizea	59
3.2	Koefizienteen tartezko estimazioa	60
3.3	Estimatu eta konparatuko diren zehazpen desberdinak	65
3.4	Zehazpen desberdinen estimazio emaitzen laburpena.	67
5.1	A ereduko aldagai azaltzaileen koerlazio matrizea	90
5.2	Engle kontrastearen emaitzak	93
6.1	A ereduko aldagaien koerlazio matrizea	101
6.2	Kolinealitate kontrastearen emaitzak	103
7.1	Pool fikzio-aldagaiaren balioak	107
7.2	Price aldagaiaren estatistiko nagusiak	107
7.3	Igerilekua duten etxebizitzaren <i>price</i> aldagaiaren estatistiko nagusiak	108
7.4	Igerilekua ez duten etxebizitzaren <i>price</i> aldagaiaren estatistiko nagusiak	108

1 Gaia

Gretl eta Ekonometria

Aurkibidea

1.1	Sarrera	2
1.2	Zer da Ekonometria?	2
1.3	Ikerketa ekonometriko bat	6
1.4	Datuak eta bere erabilera	7
1.4.1	Datuen jatorriak	8
1.4.2	Software ekonometrikoa	9
1.5	Gretlerako sarrera	10
1.5.1	Aldagai baten analisi deskribatzailea	15
1.5.2	Aldagaien arteko erlazioak	20

1.1 Sarrera

Kurtso hau errealitate ekonomikoari buruzko informazio estatistikoa interpretatu eta ikasi nahi dutenentzat zuzenduta dago. Erabiliko den oinarritzko erreminta eredu ekonometrikoa izango da, ekonomiaren eskema teorikoak eta datuen analisirako teknika estatistikoak bat egiten baititu. Eredu baten egitura oso konplexua izan daiteke baina kurtso honetan oinarritzko ereduetan zentratuko gara, hau da **erregresio eredu lineal orokorretan**.

Bestalde, kurtso honen izaera guztiz aplikatua denez, adibide praktikoen bitartez, estatistika eta ekonometriako oinarritzko kontzeptuak aztertuko ditugu. Horrela, ikasleak kasu praktikoak burutuko ditu software ekonometrikoarekin eta lortutako emaitzak interpretatzen ikasiko du. Erabiliko dugun pakete ekonometrikoa Gretl izango da, erraza, erabilera askekoa eta ikerketa ekonometrikoaren sarrerako liburu askok aztertzen dituzten datu-base anitzeko sarbidea duena.

Lehen gai hau horrela egituraturik dago: bigarren atalean ekonometriaren sarrera aurkitzen da eta hirugarrenak ikerketa ekonometrikoko baten adibide bat jasotzen du, eredu ekonometrikoko baten elementu nagusiak aztertuz. Laugarrenean datu ekonomikoen ezaugarriak, jatorriak eta datu hauek bildu eta prozesatzeko programa informatikoak laburbiltzen dira. Bostgarren atalean, erabiliko dugun Gretl softwarearen sarrera egingo dugu eta saio praktikoko baten bitartez bere erabilpeneko lehen urratsak emango ditugu.

1.2 Zer da Ekonometria?

Orokorrean, ekonomia izaerako erabakiak hartzerakoan, onuragarria izaten da informazioa datu kuantitatibo eran eskuhar izatea. Horrela, unibertsitate ikasketak aukeratzerakoan adibidez, bakoitzaren lehentasun pertsonalen arabera egin dezakegu edota aukeratutako adarreko itxarondako soldatagatik edo lanpostu bat bilatzeko izango dugun erraztasunagatik. Etxebizitza baten salerosketa baldin bada ordea, etxe-merkatuaren egoera nolakoa den interesatzen zaigu. Horretarako etxebizitzaren prezioei buruzko datuak bildu daitezke eta baita prezio horietan eragin dezaketen zenbait ezaugarriei buruzkoa ere, adibidez, tamaina, erreforma beharra duen, erabilitako etxebizitza den, etab..

Suposa dezagun tokiko egunkari bateko iragarkien atalean salmentan dauden hiriguneko 50 etxebizitzari buruzko datuak azaltzen direla:

- Etxebizitzaren prezioa mila eurotan.
- Etxebizitzaren bizigarritasun azalera metro karratutan.
- Etxebizitzaren egoera: erreforma beharra izatea edo sartzeko prest egotea.

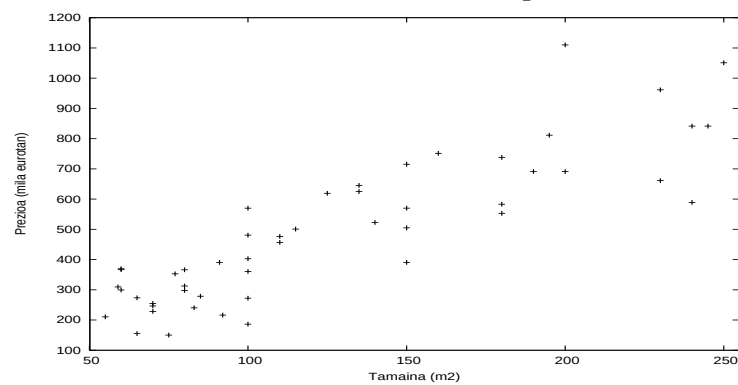
Datu hauek 1 Taulan agertzen dira, jarraian dagoen grafikoan etxebizitza bakoitzaren prezioa eta tamaina agertzen direlarik. Informazio hau kontuan izanik, erreformatutako 100 m^2 -tako etxebizitza bat 525000 eurotan eskeiniz gero, garestia deritzogula esango genuke, bere prezioa ezaugarri hauetako etxebizitza baten batezbesteko prezioa baino handiagoa baita.

$$\frac{402,6 + 360,607 + 570 + 480,809}{4} = 453,504 \text{ mila euro}$$

1.1 Taula: Etxebizitzaren prezioen datuak

Adierazlea	Tamaina	Prezioa	Erreformatzeke	Adierazlea	Tamaina	Prezioa	Erreformatzeke
1	55	210,354	ez	26	110	476,600	ez
2	59	309,520	ez	27	110	456,769	ez
3	60	366,617	ez	28	115	500,643	ez
4	60	299,304	bai	29	125	619,000	ez
5	60	369,650	ez	30	135	645,253	ez
6	65	273,460	bai	31	135	625,000	ez
7	65	155,000	bai	32	140	522,800	bai
8	70	228,384	ez	33	150	390,660	ez
9	70	246,415	ez	34	150	504,850	bai
10	70	255,000	bai	35	150	715,204	ez
11	75	150,253	bai	36	150	570,000	bai
12	77	352,800	ez	37	160	751,265	ez
13	80	366,000	bai	38	180	583,000	bai
14	80	298,000	bai	39	180	738,000	ez
15	80	312,530	ez	40	180	552,931	bai
16	83	240,400	ez	41	190	691,200	ez
17	85	278,569	bai	42	195	811,400	ez
18	91	390,658	ez	43	200	691,000	bai
19	92	216,364	bai	44	200	1110,000	ez
20	100	402,600	ez	45	230	961,620	ez
21	100	272,300	bai	46	230	661,000	ez
22	100	360,607	ez	47	240	841,417	ez
23	100	570,000	ez	48	240	588,992	bai
24	100	480,809	ez	49	245	841,400	bai
25	100	186,314	bai	50	250	1051,000	ez

1.1 Irudia: Sakabanatze diagrama



Hala ere, zer esan dezakegu erreformatu beharreko 90 m^2 -tako etxebizitza izango balitz? Edo erreformatutako 50 m^2 -takoa izango balitz? Ekonometria batek galdera hauei erantzuten lagundu diezaguke. Datuen adierazpen grafikoan oinarrituz etxebizitza baten tamaina eta prezioaren arteko erlazio edo *funtzio egonkor* bat ikusi daiteke. Erlazio hau *eredu* baliagarri batean gauzatu daiteke planteatutako galderei erantzunez. Eredu bat eta datu batzuekin, teknika ekonometrikoek etxebizitzaren tamaina edo egoerak bere prezioarengain duten eragina neurtu ditzakete. Erantzuna honakoa izan daiteke adibidez: *erreformatu beharreko 90 m^2 -tako etxebizitza baten batezbesteko prezioaren estimazioa 297350 euro da, nahiz eta prezioa 152711 eta 441989 euro bitartean egotea posible izan daiteken %90eko konfidantza mailarekin. Gainera, erreformatutako etxebizitza izango balitz, batezbesteko prezioa 100000 eurotan handituko litzateke, zenbatekoa 398580 izanik eta 210521 eta 556639 euro bitarteko prezioak eginkorrak izanik.*

Ekonometria, aldagai ekonomikoen arteko erlazioak neurtu eta kuantifikatzeko estatistika erabiltzen duen Ekonomiaren adar bat da. Ekonomia Teoria, matematika, estatistika eta konputu metodoak erabiltzen dituen disziplina anizkoitzeko materia da. Ramanathan (2002) esaten duen bezala:

*“En términos sencillos, la **econometría** se ocupa de la aplicación de métodos estadísticos a la economía. A diferencia de la estadística económica, que es principalmente datos estadísticos, la econometría se distingue por la unificación de teoría económica, instrumentos matemáticos y metodología estadística. En términos más generales, la econometría se ocupa de (1) estimar relaciones económicas, (2) confrontar la teoría económica con los datos y contrastar hipótesis relativas al comportamiento económico, y (3) predecir el comportamiento de variables económicas.”*

Zertarako balio du Ekonometriak?

Ikerketa ekonometriko baten helburua fenomeno ekonomikoak hobeto ulertzea da eta ondorioz, interesatzen zaigun fenomenoaren etorkizuneko garapena auresatea. Oinarrizko erraminta **eredua** da, aldagai ekonomikoen arteko erlazioak ulertzeko lagungarria da, zenbait neurri edo ekonomia politiken eraginak ebaluatzeko erabilgarria baita. Ekonometriaren erabilgarritasunaren zenbait adibide honakoak dira:

- Aktiboen merkatuko analista edo ikertzaile batek aktibo baten prezioa eta aktibo hori eskaintzen duen enpresaren ezaugarri desberdinen arteko erlazioa ikertu eta kuantifikatu nahi du ekonomiaren egoera orokorrarekin batera.
- Iberdrolaren burugoak elektrizitate eskaeran zeintzu faktorek eragiten duten ikertu nahi izango du.
- Eroski taldeak bere salmentetan eta mozkinetan publizitate maila desberdinek duten eragina kuantifikatu nahi izango du.
- Enpresa batek bere salmentetan publizitate kanpaina batek duen eragina estimatu nahi izango du eta erabilitako komunikabidearen arabera desberdintasun nabariak dauden

edo ez argitu. Aurreko kanpainen emaitzen ikerketa galdera horiei erantzuteko baliagarria da eta bide batez, etorkizuneko kanpainaren plangintzarako lagungarria.

- Ekonomia Ministeritzako Ikaskuntza Sailak eta Espainiako Bankuak edo Europako Banku Zentralak moneta eta fiskal politikek langabezian, inflazioan, esportazio eta inportazioan, interes tipoan, ... duten eragina aztertu nahi izango dute.
- Erakunde batek sexuarekiko soldata bereizketa zuzentzeko politikak ezarri nahi baditu, arazoan eragiten duten faktoreak ezagutu beharko ditu lehendabizi eta ondoren, ezarri daitezken neurriak ikertu, neurri horien ondorioak zeintzuk diren aztertuz.
- Herrialde-Gobernu bati populazioaren garapena aurrikustea interesatzen zaio gizarte zerbitzuen beharra eta ondorioztatzen den finantzaketa planifikatzeko. Bere finantzaketa gaitasunaren informazio zehatza eduki behar izango du eta hortaz, zerga biltzearen aurreanak ere izan beharko ditu.
- Pertsona batek prestamu bat kontratatu nahi badu, interes tipoen garapena zein den interesatzen zaio.

Azken urteetan metodo ekonometrikoen hedapena eta erabilpena handiagoa izan da, besteak beste, datuen gertutasun handiago eta kalitate hobekoagatik eta konputu metodoen garapenagatik. Ekonometriaren erabilera ez da ekonomia alorrera bakarrik murrizten, orokorrean, Giza-Zientzietan aplikagarriak diren datuen ikerketa prozedurak eskeintzen ditu. Adibidez:

- Zigorren gogortzeak (heriotz zigorraren ezarpena adibidez) kriminaltasun tasa murrizten duen edo ez analizatzeko.
- Bide segurtasun neurrien eraginkortasuna (segurtasun gerrikoa jantzi beharra adibidez) trafiko istripuetako hilkortasun tasa murrizten duen ala ez ikertzeko.
- Kirol lehiaketen emaitzak auresateko, Ingalaterrako selekzioak sartuko dituen gol kopurua munduko futbol txapelketan adibidez.
- Neurri konkretu batek bozemailetan izango duen eragina aztertzeko, adibidez, leku publikoetan erretzearen debekuak, sexu berdineko pertsonen arteko ezkontzak, etab.
- Botuetan leku-hauteskundeak, herrialde-hauteskundeak edo europar-hauteskundeak izatean diferentziarik izango den aztertzea.
- Tabako eta alkohol kontsumoaren publizitatean neurri murriztaileek produktu horien kontsumoa urritzen duen aztertzea.

Ekonometriaren hasiera XX. mendeko hirugarren hamarkadan jarri daiteke, Depresio Handiarekin bat izatea kasuala ez delarik; honen ondorioz, garai hartako ekonomilariak ikusten zituzten ziklo ekonomikoak auresan eta ekidin nahi zuten. Ekonomilari horien artean, krisi horiek arintzeko, ekonomia ihardueretan gobernuaren partehartzearen defendatzaile bezala Keynes nabarmentzen da. Horrela, lehen ekonometrik arazo makroekonomikoei erantzun nahian ibili ziren, gobernuak ekonomia politiken ezarpenean aholkatzeko asmoz.

Hasiera batean, jada zientzia naturaletan erabilitako metodo estatistikoak aplikatu ziren datu ekonomikoetan. Hala ere, metodo hauek ezin zuten mimetikoki berregin esparru ekonomikoa eta aldagai sozioekonomikoek zituzten ezaugarri bereziei egokitu behar ziren. Guzti hau dela eta, ekonometrian bi arlo hazi ziren: *ekonometria teorikoa* eta *ekonometria aplikatua*. Lehenaren helburua datuen ikerketa eta bilakatze metodoak garatzea eta bere propietateak zehaztea da. Ekonometria aplikatua berriz, metodo hauek aplikatzean datza, praktikan interesgarriak diren arazoei erantzuteko.

Kurtso honen helburuetariko bat, amaieran proiektu aplikatu bat egiteko gai izatea denez, ezinbestekoa da oinarritzko metodo eta erraminten ezagueran denbora dedikatzea, aplikazio praktiko egoki bat egiteko alde aurreko betebeharra baita.

1.3 Ikerketa ekonometriko bat

Gaur egun, etxean ordenagailu bat duen pertsona batek proiektu ekonometriko txiki bat egiteko gai da, ikerketa ekonometriko bateko ondorengo etapak burutuz (Heij, De Boer, Franses, Kloek eta van Dijk (2004)):

- *Analisiaren helburua.* Interesatzen zaiguna zehaztean datza eta horretarako erantzun nahi ditugun galderak zehazki planteatu behar dira. Adibidez, herri bateko etxermerkatuaren egoera ikertu nahi bada, hurrengo galdera egin dezakegu: zein da herrialde horretako etxebizitzaren prezioa eta zeintzu faktorek eragiten dute prezioan? Ekonomia Teoria lagungarria da arazoa fokatzeko, zein aldagai tartekatzen diren eta zein den beraien arteko erlazioa zehazteko.
- Ikerketarako nabariak diren *datu estatistiko*en *bilketa*. Aurreko adibidean erraza da etxebizitzaren prezioa, bere tamaina eta prezioan eragiten duten beste ezaugarrien datuak jasotzea (ikusi 1 Taula). Ikerketaren emaitzak datuen kalitatearen arabera egongo dira. Hala ere, analisirako aipagarriak diren datuak lortzea ez da beti erraza izaten eta noski, arazoak izan ditzakegu: datu baten egoneza, aldagai baten definizio aldaketa, datu nahikorik ez izatea edo aldagai baten informazioa ez izatea.
- *Ereduaren zehazpena eta estimazioa.* Ekonomia Teorian eta lehen etapan planteatutako galderen loturatik **eredu ekonometriko** bat lortzen da. Adibide gisa, etxebizitza baten batezbesteko prezioa (Y) bere tamainaren (X) menpean dagoela pentsa dezakegu. Teoria hau jasotzen duen eredu ekonometriko posible bat, hurrengo erlazio lineala izan daiteke:

$$Y \sim N(\alpha + \beta X, \sigma^2).$$

Hau da, tamaina konkretu bat izanik (100 m^2 adibidez), etxebizitzaren prezioa bere batezbestekoaren ($\alpha + \beta 100$) inguruan banatzen da σ^2 bariantzako banaketa normal baten arabera. Eredua zehazterakoan aldagaien arteko erlazioaren funtzioa aukeratu dugu eta baita aldagai azalduaren (Y) izaera estokastikoa ere. Helburua, gure galderei erantzuteko gai den eredu nabari eta erabilgarri bat lortzea da.

Hurrengo pausua analisirako interesgarriak diren banaketako parametro ezezagunak estimatzea da, hau da, α eta β . Estimazioa, nabaria den informazio guztia eta datuak

erabiltzean datza benetako parametro ezezagunen informazioa lortzeko. Estimazioko emaitzen interpretazioan garrantzi handikoa da parametroen balioak *ez ditugula* eza-gutzen kontuan izatea eta ondorio bezala “%95eko konfidantzarekin, erregaiei gainera zergen igoerak gasolina kontsumoan ez du eragiten” motako baieztapenak bakarrik egin ahal izango ditugula.

Estimazio metodo asko daude. Bata edo bestearen hautaketa, aukeratutako eredu ekonometrikoreen propietateen arabera dago. Eredu oker baten hautapenak estimazioen baliozkotasunean eragiten du. Kurtso honetan, praktikan erabili daitezkeen balibide sinple eta errezenak erabiliko ditugu: erregresio linealeko eredua eta *Karratu Txikiaren Arruntetako* estimazio metodoa.

- *Ereduaren analisia*. Aukeratutako eredua datuen portaera jasotzeko egokia den ikustea da. Adibidez, etxebizitzaren tamainak bere prezioan eragiten duela jasotzeko egokia den, bi aldagaien arteko erlazioa lineala zuzena den, etab. Eredua zuzen zehaztuta dagoen baloratzeko, hau da, egindako balizkoak baliodunak diren jakiteko, kontrasteetan oinarrituko gara. Horrela, kontrasteetan lortutako emaitzetan oinarrituz, baliteke eredua aldatu beharra izatea.
- *Ereduaren aplikazioa*. Eredu zuzena lortu ondoren, interesgarriak diren galderak erantzuteko erabiltzen da.

Proiektu ekonometrikorik bat burutzeko ezinbestekoa da datuak lortzea eta analisi ekonometrikorako software zehatz bat maneiatzea. Horregatik, jarraian, bi atal hauek sakonduko ditugu.

1.4 Datuak eta bere erabilera

Nola lortzen dira datu ekonomikoak? Datu ekonomikoak ez dira kontrolatutako saiakuntzetatik eratortzen. Ekonomikariek, Giza-Zientzia esparruko beste ikertzaile askok bezala, errealitateko behaketetatik lortzen dituzte datuak. Kontrolatutako saiakuntza batean ordea, ikertzaileak ikerketaren baldintza guztiak kontrolatzen ditu. Ongarrien eragina ikertzeko adibidez, kantitate desberdinak aplikatu ditzakegu lursailetan, hezetasuna edota landare bakoitzak jasotzen duen argi kopurua ere kontrolatuz. Gainera, saiakuntza hau nahi adina errepikatu daiteke baldintza berdinak mantenduz edo batzuk aldatuz, besteak beste, hezetasuna eta argi kopurua. Jakina denez, aukeratutako kantitate kopurua berdina izan arren, saiakuntza desberdinetan ez dugu emaitza berdina espero (adibidez landareen hazkuntza), erabilitako hazia desberdina delako edo neurketa erroreak ematen direlako. Esperimentu hauen arteko diferentzia natural hauek, *lagineko aldaketak* bezala ezagutzen dira.

Kontrolatutako saiakuntzetatik lortutako datuak (laborategietan adibidez) Natur Zientzietan ohikoak izaten dira eta *datu esperimental* bezala ezagutzen dira. Gizartean emandako prozedura baten ondorioz lortutako datuak berriz, ez daude pertsona bat edo askogatik kontrolatuak eta *datu ez esperimentalak* bezala ezagutzen dira. Ezaugarri hau faktore garrantzitsu bat da ekonometriako tekniken garapenerako eta kontuan izan beharko da emaitzen interpretazioa egiterakoan.

Datu ekonomikoak mota desberdinetakoak izan daitezke eta burutuko dugun analisisan eragina izango du. Lehen sailkapen batek aldagai *koantitatibo* eta *koalitativoen* datuak bereizten ditu. Lehenengoen zenbakizko balioak hartzen dituzte, adibidez, etxebizitza baten prezioa eta tamaina. Bigarrenak ordea, kategoria edo atributoetan agertzen dira, adibidez, genero, lanbidea edo etxebizitzaren egoera. Kurtso honetako lehen sei gaitan datu kuantitatiboetan oinarrituko gara eta zazpigarren gaian, faktore azaltzaileetarikoa bat kualitatiboa delaren kasua sakonduko dugu.

Bigarren sailkapen bat *denbora serietako datuak* eta *gurutzatutako datuak* izango da. Lehenak, jarraikako denbora momentuetan jasotako behaketak dira (normalean erregularrak), urteak, hiruhilabetealdiak, hilabeteak, etab.. Adibidez, Nazio-Kontabilitateko hiruhilabeteke Barne Produktu Gordina (BPG), Gizarte Segurantzako hileroko kidetza kopurua edo IBEX35-aren eguneroko balioa. Bigarrenak ordea, banako desberdinek denbora momentu batean hartzen dituzten behaketak dira, adibidez, Europar Elkarteke (EE) herri bakoitzaren 2004 urteko langabezia, populazioa, industria sektore bakoitzaren batezbesteko soldata 2005 urtean edo iazko irailean familia multzo batek egindako gastua testu liburuetan. Posible da ere denbora serieko eta gurutzatutako datuen konbinazio bat izatea, adibidez, 2003-04 eta 2004-05 ikasturteetan ekonometriako ikasleek lortutako kalifikazioa. Denboran zehar inkesta egin duten banakoak berdinak direnean, adibidez langabezia tasa eta BPG 1990tik 2004 arte EEko 25 herrialdeetan, *taula-datuak* edo *datu longitudinalak* bezala ezagutzen dira.

Kurtso honetan erabiliko ditugun teknikak, gurutzatutako datuak aztertzeke balio dute eta denborazko serieko datuetan ere aplikatu daitezke. Hala ere, askotan azken datu hauen ezaugarriak direla eta, teknika konplexuagoak erabili behar dira. Goi mailako teknika hauek kurtso honetatik kanpo gelditzen dira, oinarritzko tekniken erabilpenean oinarrituko garelako.

Hirugarren sailkapen bat bateratzearen mailaren funtzioan ezartzen da. Horrela, *datu mikroekonomikoak* edo *mikrodatuak* izango ditugu. Hauek, banakoak, familia edo enpresa bezalako agente ekonomikoen portaera jasotzen dute, adibidez, INEk prestatutako Populazio Aktiboko Inkesta http://www.ine.es/prodyser/micro_epa.htm. Bestalde, *datu makroekonomikoak* edo *makrodatuak* ditugu. Hauek, banakako agenteen baterakuntzaren ondorioz sortutako herri, eskualde edo nazioei buruzko datuak dira, adibidez, Nazio-Kontabilitateko emaitzak.

1.4.1 Datuen jatorriak

Askotan datuak bilatu eta biltzea ez da erraza izaten, batzuetan egoerari egokitzen diren datuak lortzea eta maneiatzea lan handia izaten baita. Hala ere, azken urteetan datuen bilketa asko hobetu da, batez ere erakunde askok bere datu-baseetan sartzen uzten dutelako *World Wide Web* (WWW) orrian. Datu makroekonomikoak argitaratzen dituzten zenbait erakunde honakoak dira:

- Euskal Estatistika Erakundea (EUSTAT): <http://www.eustat.es>.
- Espainiako Bankoa: <http://www.bde.es> → Estadísticas. Besteak beste *Boletín estadístico mensual* eta *Boletín de coyuntura mensual*.
- Nazio Estatistika Erakundea (INE): <http://www.ine.es> → Inebase edo Banco tempus. Eskuragarri daude adibidez, Biztanleria Aktiboaren Inkestako emaitzak, Nazio

Kontabilitatekoak eta Hileroko Aldizkarikoak. Gainera, estekaturik estatistika zerbitzuko beste web orri batzuk aurkitzen dira.

- **EUROSTAT.** Europako Batzordeko Estatistika-bulegoa da eta estatu kideek jasotako datuak analizatzeaz eta baieztatzeaz arduratzen da. Eurost-aren betebeharra datuak bateratzea eta metodologia homogeen bat erabiliz konparagarriak direla ziurtatzea da. Taula estatistikoak, estatistika-aldezkarriak, informazio-aldezkarriak, lan-papereak etab., ondorengo helbidean lortu daitezke:
<http://europa.eu.int/comm/eurostat>.
- **Ekonomia Lankidetzaren eta Garapenerako Antolakundea (ELGA):** Bertan Nazioarteko Merkataritza argitalpeneneko zenbait serie eskuragarri daude. Bere web orria:
<http://www.oecd.org>.
- **Nazioarteko Moneta Funtza, NMF (Fondo Monetario Internacional, FMI):**
<http://www.imf.org>. Herrialde askotako datuak lortzeko bere **Nazioarteko Finantza Estatistikak** (hilekoa eta urtekoa) argitalpena kontsultatu daiteke.

Ekonometriarako eskuliburu askok, argibide moduan testuan sakontzen diren datu-base bateko CD bat barneratzen dute. Kurtso honetan Ramanathan (2002) liburuko datuak erabiliko ditugu gehienbat.

1.4.2 Software ekonometrikoa

Ordenagailuen garapenak datu kopuru handi bat gordetzeko aukera eskeintzen du eta aldi berean, bere maneia errazten du. Gaur egun, analisi ekonometrikoko bat burutzeko pake asko daude, jarraibide erraz batzuk erabiliz eragiketa konplexuak egitea posible izanik. Datuak papelean bakarrik eskuragarriak badira, datuak sartzeko, prestatzeko eta eragiketa errazak egiteko kalkulu-orriak (adibidez EXCEL) erabili daitezke. Hala ere, orokorrean, programa ekonometrikoko zehatzak erabiltzea komenigarri izaten da. Ekonometriako kurtsoetan ohikoenak honakoak dira:

- **Gretl.** Allin Cottrell (Wake Forest Unibertsitatea) garatua. Software askea da eta erabiltzeko erraza: *<http://gretl.sourceforge.net>* orritik deskargatu daitekeen gaztelaniako bertsio bat du. Orri honetan ere, Ekonometriako zenbait testu barneratzen dituen datu-base zabal bat dago erabilgarri, besteak beste, Ramanathan (2002), Wooldridge (2003) eta Greene (2003) eta Espainiako Bankua bezalako erakunde publiko batzuen datuak ere barneratzen ditu.
- **EViews.** Quantitative Micro Softwareak garatuak. Analisi ekonometrikoko teknika asko biltzen ditu. Programaren informazioa jasotzeko web orria:
<http://www.eviews.com>.
- **SHAZAM.** British of Columbia Unibertsitateak (Canada) garatua eta eredu ekonometrikoko asko estimatzeko teknikak biltzen ditu. Informazio gehiago hurrengo web orrian lortu daiteke: *<http://shazam.econ.ubc.ca>*.

- **RATS**, *Regression Analysis of Time Series*-ren akronimoa da. Analisi ekonometriako-rako teknika ugari biltzen ditu, denborazko serieko ikerketan arduraldi berezia izanik. Bere web orria honakoa da: <http://www.estima.com>
- **R**, estatistikako eragiketa eta grafikoak egiteko software askea da. Lengoaia bat hau da exekuzio inguru bat, bertan komandaok idatzi eta exekutatzeko aukera ematen du. Bere web orria honakoa da: <http://www.r-project.org>

Kurtso honen helburuetariko bat, ikaslea programa ekonometrikoekin ohitzea da. Bere baktuntasuna eta ailegaerraztasuna kontuan izanik, kurtso honetan ikasleari Gretl erabiltzeko gomendatuko zaio. Horrela, ikasgai bakoitzean proposaturiko praktikak Gretlekin askatu beharko ditu. Horretarako *Gretl Web* orrian PC batean instalatzeko argibideak ikusi daitezke: *gretl_install.exe* jaitsi eta exekutatu behar da. Ondoren, euskarazko bertsioa instalatzeko *B Eranskinean* dituzuen pausuak jarraitu.

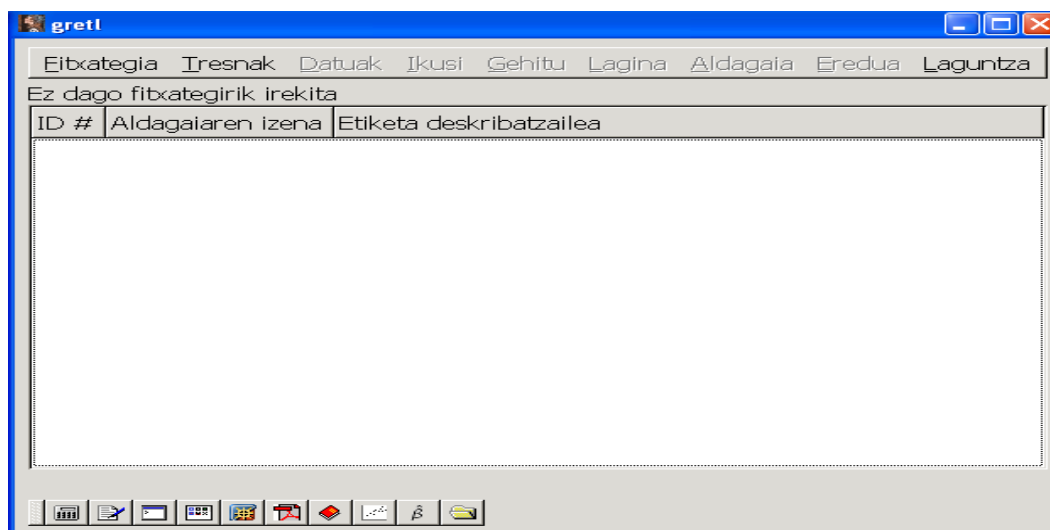
Beste datu-base batzuk http://gretl.sourceforge.net/gretl_data.html web orrian eskuratu daitezke. Bestalde, *textbook datasets* aukeran, jarraian azaltzen ditugun testuliburuetao aplikazioetan erabilitako datuak agertzen dira: Davidson eta Mackinnon (2004), Gujarati (1997), Stock eta Watson (2003), Verbeek (2004) eta Wooldridge (2006).

1.5 Gretlerako sarrera

Atal honetan, Gretl programarekin lehen hurbilketa izango dugu. Bertan, 1 Taulako datuak barneratuz datu-fitxategi bat osatuko dugu eta analisi deskribatzaile bat egingo dugu.

Fitxategiaren prestakuntza. Gretl exekutatzean hurrengo leihatila azaltzen da:

1.2 Irudia: Gretl-eko pantaila nagusia



Oraindik ez dugunez fitxategirik kargatu, menu nagusiko aukera batzuk gris argi batean agertzen dira, erabilgarri ez baitaude. Praktika honetan analizatuko ditugun datuak ez daude

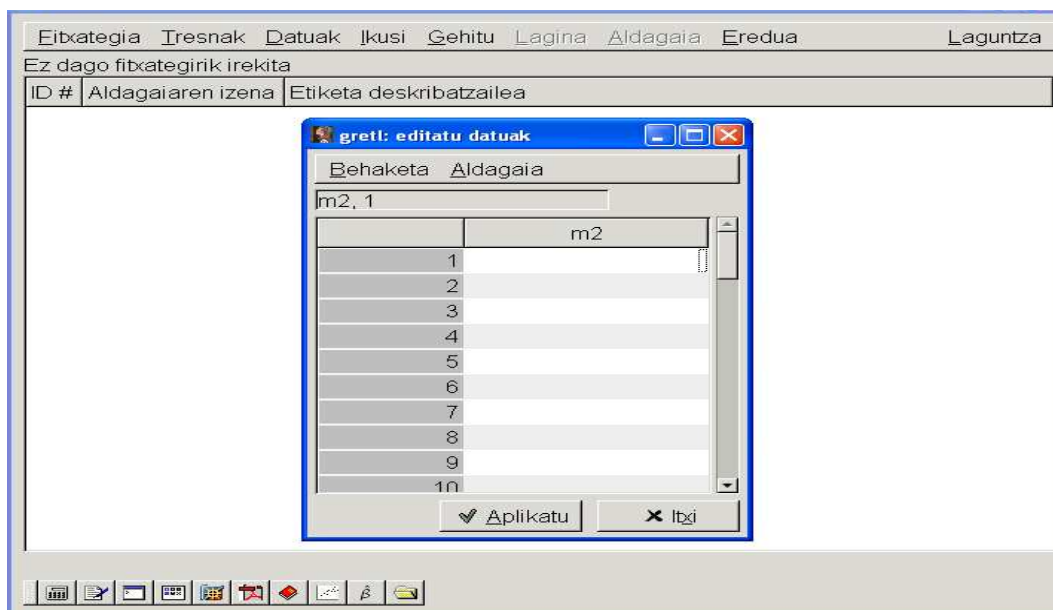
Gretleko datu-basean sartuta eta beraz, ondorengo aukerara joango gara,

Fitzategia $\rightarrow$ *Datu-fitzategi berria*

Programak eskatzen duen informazioa osatzen joango gara:

- *Behaketen kopurua*, 1 taulan azaldu dugunez, 50 etxebizitza ditugu.
- Nolako datuak ditugun. Gure adibidekoak *gurutzatutakoak* dira.
- Aurreko pausua zuzen egin bada, datu multzoaren egitura baieztatuko dugu *Aurrera* eta *Ados* klikatuz. Nolako datuak diren azaltzen den lehiatila berreskuratu nahi bada, *Atzera* klikatu behar da, baina aukera honek ez du behaketa kopuruan izan den akatsa zuzentzen uzten.
- Azken lehiatilan *Bai* klikatuko dugu datuak barneratzen hasi nahi badugu.
- Ondorengo lehiatilan *lehenengo aldagaiari ezarriko diogun izena* jarriko dugu, adibidez, *m2*. Kontuz, ezin baitaiteke $\tilde{n}$ hizkia, tilde edota 15 karaktere baino gehiago erabili. Jarraian *Ados* klikatu eta kalkulu-orri bat irekiko da, pantailan honakoa azalduz:

1.3 Irudia: Datuen edizioa



m2 aldagaiaren datuak sartzeko, saguarekin dagokion gelaxkara joan (adibidez lehenengora) eta saguko eskubiko tekleri eman; 55 zifra tekleatu ondoren, *intro* zapalduko dugu. Datu bat sartzerakoan akats bat egiten badugu, adibidez bigarren behaketa (59 m^2) ez badugu sartu, ondorengo errenkadan jarri (gure adibidean 60 m^2) eta *Behaketak* $\rightarrow$ *Erantsi behaketak* klikatuz, errenkada txuri berri bat irekiko da aurrekoaren gainean. Gogoratu, lan saioko aldaketak gordetzeko *Aplikatu* klikatu behar dela.

Aldagai gehiago erantsi ditzakegu kalkulu-orriko menuko *Aldagaia* → *Gehitu* aukerarekin. Adibidez, *Erreforma* izeneko aldagai berri bat sortuko dugu. Aldagai hau kualitatiboa denez, *erreformatzeke* = *bai* egoerari 0 balioa elkartuko diogu eta *erreformatzeke* = *ez* egoerari 1 balioa. Datu guztiak sartu ondoren, *Aplikatu* klikatu eta kalkulu-orria itxiko dugu. Azken aldaketak ez baldin baditugu gorde, kalkulu-orria ixterakoan koadro berri bat azalduko da aldaketen berrespena eskatuz. Osatutako serieak pantailan horrela azalduko dira:

1.4 Irudia: Datuen informazioa

ID #	Aldagaiaren izena	Etiketa deskribatzailea
0	const	Berez eraikitako konstantea
1	m2	etxebizitzaren tamaina metro karratuetan
2	Errefor	0 balioa etxebizitza erreformayau behar bada

Komenigarria izaten da artxibo edo fitxategi batean jadanik sartutako datuak gordetzea, menu nagusiko *Fitxategia* → *Gorde datuak* erabiliz. Hurrengo koadroan direktorioa erantsiko diogu eta datuen fitxategiaren izena, adibidez, *etxebizitzak*. Datuak *gdt* luzapenarekin gordeko ditu eta hurrengo saio batean fitxategia irekitzeko, ireki Gretl programa (euskarazkoa) eta *Ireki datuak* aukeratu ondoren *Erabiltzaile fitxategia...* hautatu eta nahi den datu-fitxategia ireki¹.

Askotan, datuak EXCEL bezalako beste kalkulu-orri batean egoten dira gordeta. Adibidez, *etxebizitzak.xls* artxiboan, 1 Taulako *m2* eta *prezioa* aldagaiak azaltzen dira. Horrela, *Prezio* aldagaiaren datuak Gretleko artxibora eranstea oso erraza da.

- Konprobatu Gretleko kalkulu-orria beste datu batzuekin irekita ez dagoela.
- Menu nagusiko *Fitxategia* → *Erantsi datuak* → *Excel...* erabiliz.
- Eman *etxebizitzak.xls* EXCEL artxiboaren izena eta kokapena.
- Datuak inportatu behar diren hasierako gelaxka zehaztu. Kasu honetan, *prezioa* aldagaia B1 gelaxkan (izena dagoena) hasten da eta *zutabea 2*, *errenkada 1*-tik inportatuko

¹ Oharra: datu-fitxategia bi aldiz klikatzen baduzue, orduan gaztelaniazko Gretl programa irekiko zaizue.

ditugu datuak. Bestalde, m^2 eta *Prezioa* aldagaiak eransteko, *zutabea 1, errenkada 1-n* hasiko ginateke. Ondoren datuak *gurutzatutakoak* direla jarri, aurrera egin eta *Ados* klikatu.

Azkenik, datuen inportazioa ondo burutu den konprobatzeko, *Datuak* → *Aukeratu guztia* atalera joan eta kalkulu-orria aktibatuko dugu *Datuak* → *Editatu balioak* aukerarekin edota datuak pantailan agertuko dira *Datuak* → *Erakutsi balioak*, egin daiteke. Gure adibidean, ondorengo leihatila agertuko da:

1.5 Irudia: Datuen erakuspena

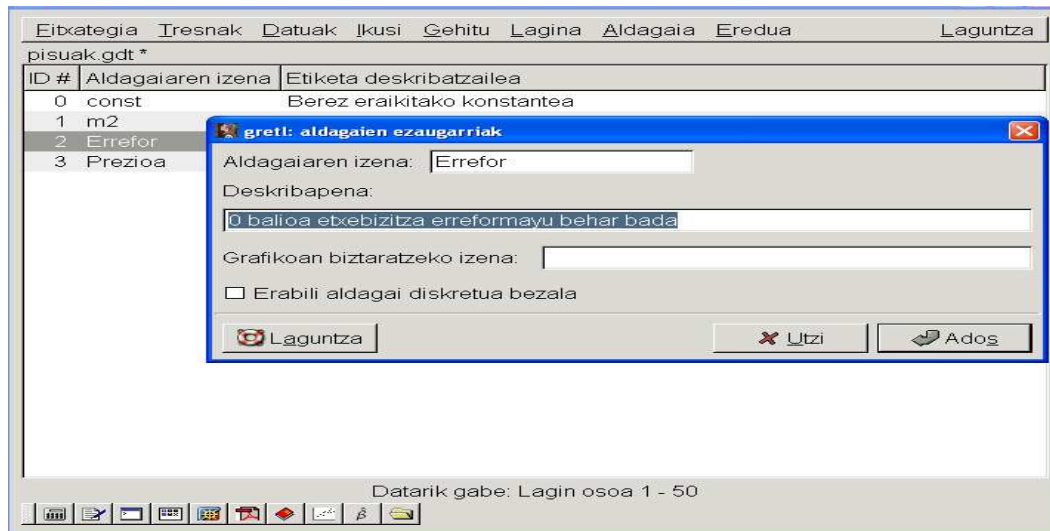
Obs	m2	Errefor	Prezioa
1	55	1	210,354
2	59	1	309,520
3	60	1	366,617
4	60	0	299,304
5	60	1	369,650
6	65	0	273,460
7	65	0	155,000
8	70	1	228,384
9	70	1	246,415
10	70	0	255,000

Azalpen oharrak. Datuak hurrengo batean erabiltzeko fitxategi batean gordetzen direnean, egindako lanaren azalpen oharrak eranstea komenigarri izaten da. Gretlen informazio gehigarria jartzeko posibilitatea dago. Informazio orokorra jartzeko:

Datuak → *Editatu informazioa* aukeran, azalpen testua eransteko gelaxka bat irekitzen da eta bertan nahi dena idatzi dezakegu, adibidez: “Etxebizitzaren prezioen analisia”.

Informazio zehatzari dagokionez, aldagai bakoitzaren deskribapen labur bat jarri dezakegu aldagaiaren izenarekin batera, *Etiketa deskribatzaile* bezala agertuz. Adibide gisa, *Erreforma* aldagaiaren informazio ohar bat jarriko dugu. Saguarekin aldagaia markatuko dugu eta *Aldagaia* → *Editatu ezaugarriak* aukera klikatu eta *Deskribapena* ataleko hurrengo gelaxkan 0 balioa etxebizitzak erreforma beharra du idatziko dugu. Azkenik, *Ados* klikatu ondorengo leihatilan agertzen den bezala.

1.6 Irudia: Datuen ezaugarriak editatzen

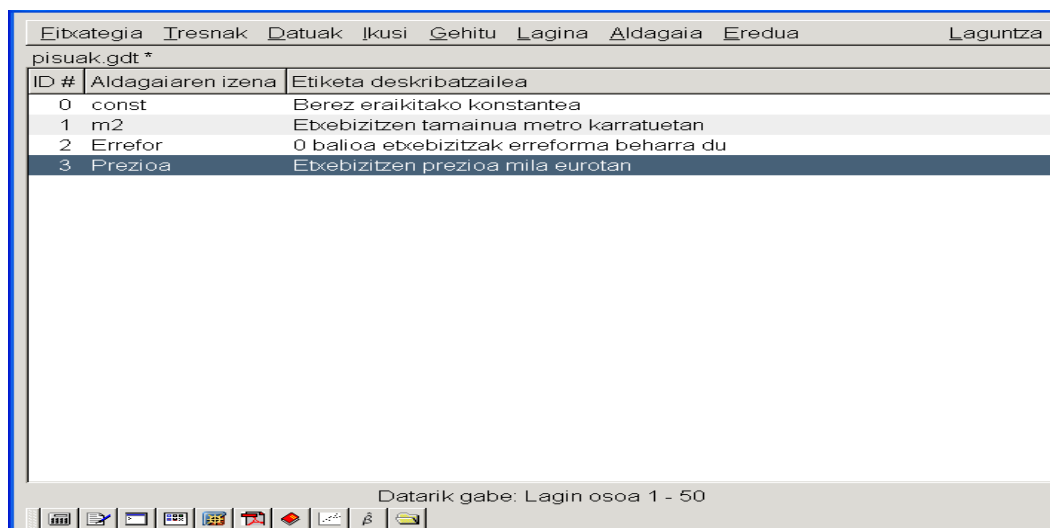


Etiketa deskribatzaileak erabilgarriak dira datuen jatorria ezagutzeko edo dagokien neurri unitateak jakiteko. Adibidez, *Prezioa* eta *m2* aldagaiei ondorengo etiketak ezarriko dizkiegu:

Ald. izena	Deskribapena	Grafikoan biztaratzeko izena
<i>Prezioa</i>	Etxebizitzaren prezioa mila eurotan	Prezioa (mila euro)
<i>m2</i>	Etxebizitzaren tamaina metro karratutan	Azalera (m2)

Prozedura ongi egin bada, pantailan ondorengoa agertuko litzateke:

1.7 Irudia: Datu-fitxategiaren erakuspena



Azkenik, *Datuak* → *Irakurri informazioa* aukerarekin datu multzoaren egituraren informazioa ikusi daiteke erantsitako aldagai azaltzaileekin batera.

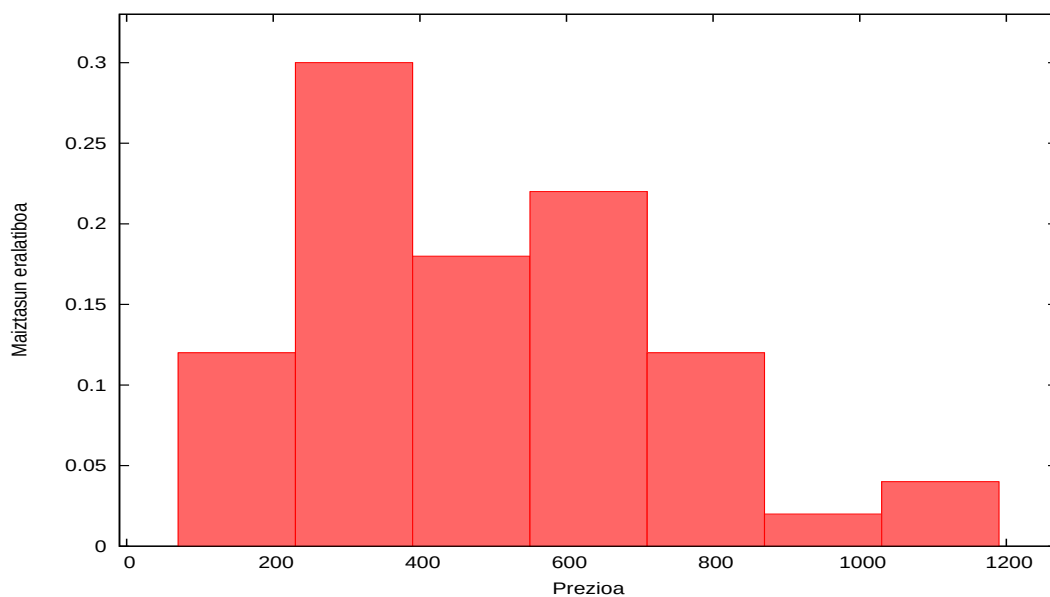
1.5.1 Aldagai baten analisi deskribatzailea

Ikerketa ekonometrikoek datu-base zabalak behar izaten dituzte. Lehen etapa batean, datuen ikuspegi orokor bat egin behar da, analisi deskribatzaile baten helburua datu multzoa laburtzea delarik, ikerketarako ezaugarri eta informazio nabariena ateratzeko. Lehendabizi aldagai bakoitzaren informazioa laburbilduko dugu eta ondoren, aldagaien arteko erlazioen irudiak jasoko ditugu. Horretarako, *grafikoak* eta *estatistiko deskribatzaileak* erabiltzen dira (ikusi Peña eta Romo (1997)).

Gretlen, aldagai baten analisi deskribatzaileko elementu nagusiak *menu laguntzailean* agertzen dira, aldagaiaren gainean saguko eskubiko tekla klikatuz edo menu nagusiko *Aldagaia* aukera klikatuz.

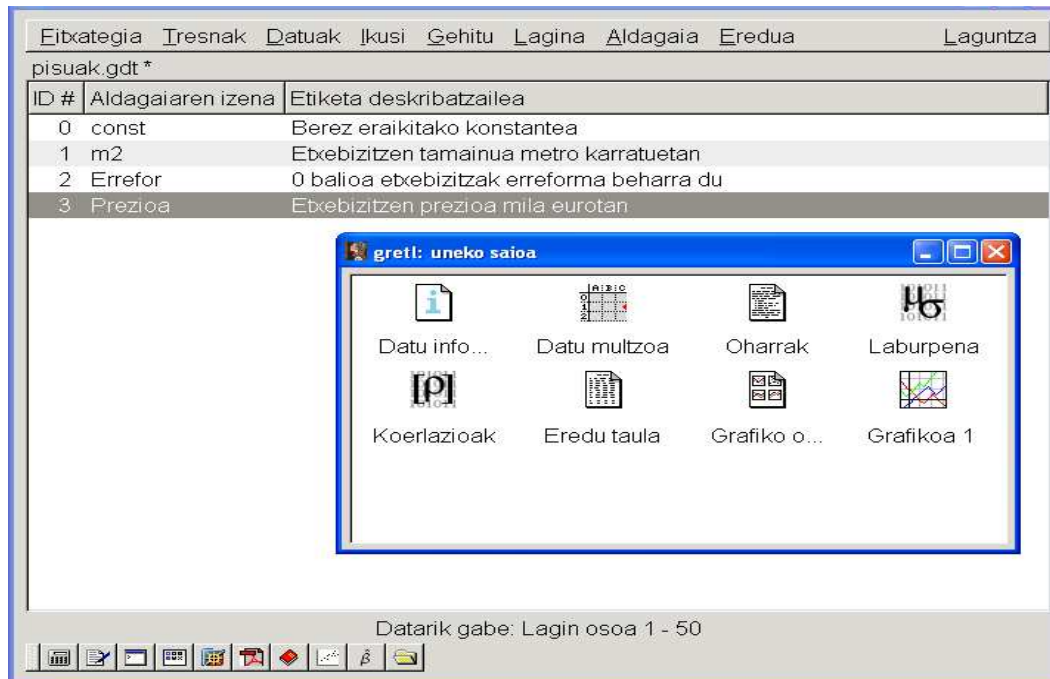
Aldagai ekonomiko baten gurutzatutako datuak laburbiltzeko gehien erabiltzen den grafikoa **histograma** da, menu laguntzailean *Maiztasun grafikoa* aukeran agertzen delarik. Barra diagrama bat da, ardatz horizontalean aldagaien balioak tarteka banatuak agertzen direlarik. *Prezioa* aldagaiaren gainean saguko eskubiko tekla klikatuz eta *Maiztasun grafikoa* klikatuz, ondorengoa lortzen da:

1.8 Irudia: Maiztasun erlatiboko histograma



Grafikoan klikatuz, menu laguntzaile batek irteten da eta aldaketak egitea posible da (*Editatu*) edo formato desberdinetan gordetzea (postscript, pdf, etab.). Gretleko saioan zehar, *Gorde ikono bezala saioan* aukerarekin grafikoa gordetzen dugu eta honela, berriro berreskuratu daiteke behean eskubialdean dagoen laugarren sinboloan (*saioaren ikono ikuspegia*) eta gero *Grafikoa 1*-ean klikatuz.

1.9 Irudia: Ikonoen ikuspegia



Histograman adierazitako maiztasun banaketak ikusteko, dagokion aldagaia markatu eta *Aldagaia* → *Maiztasunen grafikoa* aukerara zuzenduko gara. Gure adibideko *Prezioa* aldagaiaren maiztasun banaketari dagokiona honakoa litzateke:

1.2 Taula: Etxebizitzaren prezioaren maiztasunak

Prezioa -ren maiztasun banaketa, behaketa 1-50 tarte kopurua = 7,
batezbestekoa = 489,858, D.T. = 237,416

tartea	erdiko pt	maiztasuna	erl.	met.	
< 230,23	150,25	6	12,00%	12,00%	****
230,23 - 390,19	310,21	15	30,00%	42,00%	*****
390,19 - 550,15	470,17	9	18,00%	60,00%	*****
550,15 - 710,11	630,13	11	22,00%	82,00%	*****
710,11 - 870,06	790,08	6	12,00%	94,00%	****
870,06 - 1030,0	950,04	1	2,00%	96,00%	
>= 1030,0	1110,0	2	4,00%	100,00%	*

Banaketa normala izatearen hipotesi hutsaren kontrastea:
Chi-karratua(2) = 6,825 p-balioarekin 0,03296

Lehen zutabea *Prezioa* aldagaiaren balioen tarteak agertzen dira eta bigarrenean, tartearen erdiko puntua edo **tartearen marka**. Hirugarren zutabea *maiztasunak* agertzen dira. Hauek tarte baten **maiztasun absolutuak** dira, hau da, tarte horretako prezioa duten etxebizitzak kopurua. Adibidez, 1 Taulan 230,23 eta 390,190 euro tarteko prezioa duten etxebizitzak 15 dira. Laugarren zutabea (*erl.*) tarte bakoitzaren **maiztasun erlatiboak** agertzen dira, hau da, tarte bakoitzaren zatikia. Maiztasun hauekin osatu da aurreko histograma. Adibidez, [230,23 ; 390,190) tarteko 15 etxebizitzek, etxebizitza guztien %30a osatzen dutenez eta tarte guztien zabalera berdina denez, histogramaren bigarren zutabearen altuera 0,3 izango da. Tarte baten maiztasun erlatiboari, aurreko tarteena gehitzen badiogu, tarte horretarainoko **bateratutako maiztasun metatua** lortuko genuke, balio hauek bostgarren zutabea agertzen direlarik (*met.*). Horrela, ikertzen ari gareen adibideko etxebizitza multzoan, %42 etxebizitzek 390190 euro baino prezio baxuagoa dute.

Aldagai baten zenbakizko deskribapenari dagokionez, aldagaian ezarritako saguko eskubiko botoia klikatuz *Estatistiko deskribatzaileak* aukera agertzen da edo menu nagusiko *Aldagaia* → *Estatistikoen laburpena* aukeran. *Prezioa* aldagaiari dagokiona hurrengo taulan agertzen da:

1.3 Taula: Prezio aldagaiaren estatistiko nagusiak

Estatistikoen laburpena, 1 - 50 behaketak erabiliz
'Prezioa' aldagaiarentzat (50 behaketa baliagarriak)

Batezbestekoa	489,86
Mediana	466,68
Minimoa	150,25
Maximoa	1110,0
Desbideratze tipikoa	237,42
Aldakuntza Koefizientea (A.K.)	0,48466
Asimetria	0,68052
Kurtosis soberakina	-0,19251

Leihatila honek menu berri bat du. *Kopiatu* aukerarekin taula hau, RTF (MS Word), LaTeX edo pantailan agertzen den bezala (*Testu arrunta*) inportatu daiteke.

Histograman jasotzen diren banaketen ezaugarri batzuk, estatistiko deskribatzaile hauetan laburbiltzen dira. Batezbestekoa eta mediana lekuzko neurriak dira, desbideratze tipikoa eta aldakuntza koefizientea sakabanatze neurriak dira, eta azkenik, asimetria eta kurtosis soberakina formaren neurriak dira.

Lekuzko neurriek puntu multzoen erdigunearen ideia ematen dute. *Batezbestekoa*, datu guztien batura eta behaketa kopuruaren (laginaren tamaina) zatidura da. Aldagai batek N behaketa baldin baditu $(x_1, x_2, \dots, x_N)$, batezbestekoa edo lehen ordenako lagin momentua,

horrela definitzen da:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_N}{N} = \frac{1}{N} \sum_{i=1}^N x_i \quad (1.1)$$

Batezbestekoa muturreko balioekin oso aldakorra da, hau da, behaketa anomaloek edo ezohikoek eragin handia dute batezbestekoak hartuko duen balioaren gain. Adibidez, azken etxebizitzak prezio *oso handia* izango balu, 1051 mila euro izan beharrean 1350 mila euro adibidez, orduan batezbesteko prezioa ia 6000 eurotan handituko litzateke, 495,84 mila eurotan kokatuz.

Orokorrean, behaketa gutxiren aldakuntzekin (nahiz eta hauek oso handiak izan) asko aldatzen ez diren estatistikoak interesatzen zaizkigu. Propietate hau betetzen duen erdiguneko balioa *mediana* da. Datuak txikitik handira ordenatu ondoren erdiko posizioan gelditzen den balioa mediana da. Adibidean

$$Mediana = 466,6845 \text{ mila euro}$$

Lekuzko neurriak datu multzoaren balio adierazgarriak dira baina dagokien errore-neurri batekin osatu behar izaten dira. Balio bakar honen adierazgarritasuna baloratzeko **sakabanatze edo dispersio neurriak** erabiltzen dira, datuak batezbestekoarekiko hurbil edo urrun dauden adieraziko baitigute. Neurri sinple bat *ibilbidea* da, hau da, datuek hartzen duten balio maximo eta minimoaren arteko diferentzia.

$$\text{Ibilbidea} = \text{Maximoa} - \text{Minimoa}$$

Gure adibidean, prezioen ibilbidea $1110 - 150,25 = 959,75$ mila eurokoa da. Neurri honek, muturreko balioak bakarrik hartzen ditu kontuan. Ondorengo neurriak aldiz, datu guztiekin kalkulatzen dira. Hasteko, desbideratze tipikoa dugu, bariantzaren erro positiboa delarik. Datu multzo baten bariantza, S_x^{*2} , datuek batezbestekoarekiko duten desbideratzeen karra-tuen batezbestekoa da:

$$S_x^{*2} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_N - \bar{x})^2}{N - 1} = \frac{1}{N - 1} \sum_{i=1}^N (x_i - \bar{x})^2 \quad (1.2)$$

Beraz, *desbideratze tipikoa*, S_x^* , horrela definituko genuke:

$$S_x^* = + \sqrt{\frac{1}{N - 1} \sum_{i=1}^N (x_i - \bar{x})^2} \quad (1.3)$$

Bariantza eta desbideratze tipikoa, batezbestekoarekiko sakabanatzearen neurriak dira. Bere balio minimoa zero da eta sakabanatzea minimo hau lortzen denean behaketa guztiek balio berdina hartzen dutelako izaten da. Desbideratze tipikoaren abantaila, bariantzarekin konparatuz, hasierako aldagaiaren neurri unitate berdina izango duela da.

Orokorrean, S_x^* zerotik gero eta hurbilago badago, datuak batezbestekotik hurbilago edo kontzentratuago egongo dira eta beraz, batezbestekoa, behaketa multzoaren adierazgarriagoa izango da. Hala ere, neurri unitateen menpean dagoenez, ez da erraza izaten bi datu multzoen adierazgarritasunak konparatzea. Horregatik, *Aldakuntza Koefizientea* (A.K.) kalkulatzea

komenigarri izaten da. Koefiziente hau sakabanatzearen neurri bat da eta ez dago neurri unitateen menpean. Desbideratze tipikoa eta batezbestekoaren (balio absolutuetan) arteko zatiketa bezala definitzen da. Hau da,

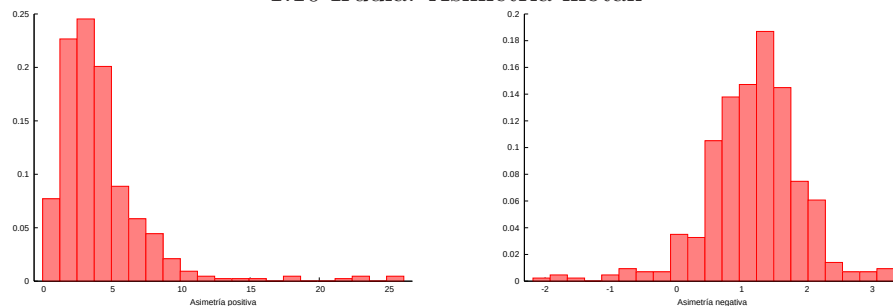
$$A.K. = \frac{S_x^*}{|\bar{x}|} \quad \bar{x} \neq 0 \quad \text{bada.}$$

Gure adibidean, prezioen aldakuntza koefizientea $A.K. = 0,485 < 1$ denez, batezbestekoa datu multzoaren adierazgarri dela esango genuke.

Batezbestekoa eta desbideratze tipikoa estatistiko ezagunenak dira, baina normalean **forma-ren neurriekin** batera azaltzen dira, hauek histogramaren beste ezaugarri batzuk jasotzen baitituzte. Banaketa baten asimetriak, datuak batezbestekoaren inguruan simetrikoki azaltzen diren edo ez neurtzen du. *Asimetria koefizientea* (ASIM) horrela definitzen da:

$$\text{Asimetria koefizientea} = \frac{1}{N} \sum_{i=1}^N \left(\frac{x_i - \bar{x}}{S_x} \right)^3 = \frac{1}{N} \frac{\sum_{i=1}^N (x_i - \bar{x})^3}{S_x^3} \quad (1.4)$$

1.10 Irudia: Asimetria motak



non $S_x = \sqrt{(N-1)/N S_x^*} = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2}$ den. Asimetria koefizientea zero izango da datuak batezbestekoaren inguruan simetrikoki banatzen badira, positiboa eskubiko buztana (batezbestekoa baino handiagoak diren behaketei dagokiona) ezkerrekoa baino luzeagoa bada eta negatiboa bestelako kasuan. Etxebizitzen prezioen adibidean, asimetria positiboa denez, batezbestekoa mediana baino handiagoa irteten da, hau da, $\bar{x} > \text{Mediana}(X)$

Kurtosis koefizienteak banaketaren zorrotasuna jasotzen du eta horrela definitzen da:

$$\text{Kurtosis} = \frac{1}{N} \sum_{i=1}^N \left(\frac{x_i - \bar{x}}{S_x} \right)^4 = \frac{1}{N} \frac{\sum_{i=1}^N (x_i - \bar{x})^4}{S_x^4}$$

Koefiziente honek buztanetako behaketa kopurua neurtzen du batezbestekoaren inguruan direnekin konparatuz. Erreferentzia neurria hiru da, berau banaketa *normal* baten kurtosis koefizientearen balioa baita. Hortaz soberakina honakoa da:

$$\text{Kurtosis soberakina} = \frac{1/N \sum_{i=1}^N (x_i - \bar{x})^4}{S_x^4} - 3 \quad (1.5)$$

Soberakin positibo batek, buztanetan banaketa normal batek baino behaketa gehiago edo pisu gehiago duela esan nahi du eta balio negatibo batek berriz, behaketa gutxiago buztanetan.

Aldagai multzo bat dugunean berriz, aldagai guztien estatistiko deskribatzaileak jasotzen dituen taula lortu daiteke Gretlekin. Hau giteko bi era daude:

1. Aldagaiak aukeratu saguaren ezkerreko botoia eta *Control* tekla batera klikatuz.
2. *Ikusi* $\rightarrow$ *Estatistikoen laburpena* aukerara joan edo aukeratutako aldagaietan saguaren eskubiko botoia klikatuz agertzen den menu laguntzaileko *Estatistiko deskribatzaileak* klikatu.

Horrela, 1 Taulako aldagaien datuekin, ondorengo estatistiko deskribatzaileak lortzen dira:

1.4 Taula: Aldagaien estatistiko nagusiak

Estatistikoen laburpena, 1 - 50 behaketak erabiliz

Aldagaia	BATEZ.	MEDIANA	MIN	MAX
m2	127,34	105,00	55,000	250,00
Errefor	0,62000	1,0000	0,00000	1,0000
Prezioa	489,86	466,68	150,25	1110,0

Aldagaia	D.T.	A.K.	ASIM	KURT.SOB
m2	59,048	0,46370	0,67091	-0,77954
Errefor	0,49031	0,79083	-0,49445	-1,7555
Prezioa	237,42	0,48466	0,68052	-0,19251

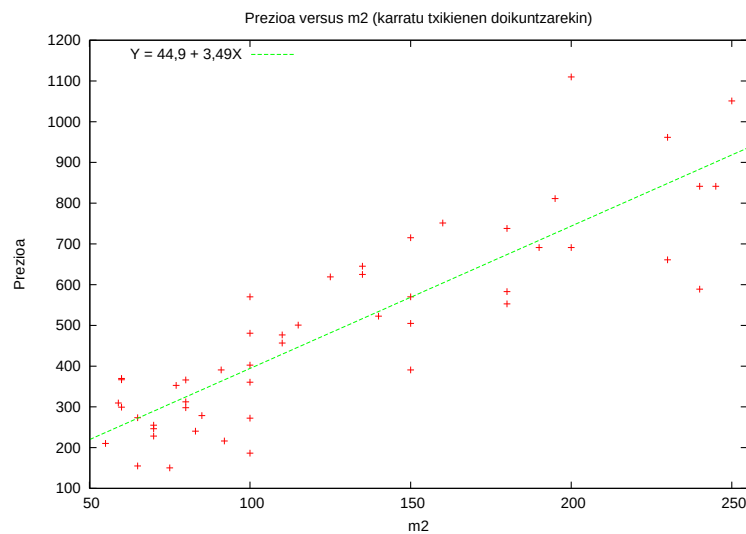
non BATEZ. batezbestekoa den, D.T. desbideratze tipikoa den, MIN aldagaiak hartzen duen balio minimoa, MAX, berriz maximoa eta KURT.SOB kurtosis soberakina den. Eraitza hauek interpretatzerakoan *Errefor* aldagaia kuantitatibo jarraia ez dela kontuan hartu behar da, izatez aldagai koalitatiboa baita zeren 1 edo 0 balioak bakarrik hartzen baititu.

1.5.2 Aldagaien arteko erlazioak

Aztertzen ditugun aldagaieetatik bi kuantitatiboak baldin badira, beraien arteko erlazioa analizatzea komeni da. Orokorrean, bi (edo gehiago) aldagai ikertzerakoan, bien arteko kausalitate erlazioa aztertu daiteke. Etxebizitza baten prezioa bere tamainaren ondorioa dela pentsatu dezakegu, baina ez da kontrakoa ematen. Eragina sortarazten duen aldagaiari (x), azaltzailea edo exogenoa deituko diogu eta eragina jasotzen duenari berriz (y) azaldua edo endogenoa. Bi aldagai hauen arteko erlazioa grafikoen bitartez aztertu daiteke edo zenbakien bitartez, koerlazio koefizientea edo erregresio zuzena kalkulatu. Aldagai multzo baten analisi deskribatzaileko elementu guzti hauek Gretleko *Ikusi* aukeran agertzen dira.

Adierazpen grafikoa. Bi aldagaien arteko erlazioaren ideia bat sakabanutze diagramak *scatterplot*-ak eman diezaguke, planoan (x_i, y_i) , $i = 1, \dots, N$ puntu guztiak adieraziz: aldagai bat

1.11 Irudia: X-Y grafikoa



(x) ardatz horizontalean adierazten da eta beste aldagaia (y) ordea, jatorri ardatzean. Adibidez, Gretl erabiliz bigarren orrialdeko diagrama lortzeko (prezioa vs tamaina), jarraitu beharreko pausuak honakoak dira:

- *Ikusi* → *Grafikoak* → *X-Y grafikoa* eta *definitu grafikoa* lehiatilan ondorengoa markatu:
X-ardatzaren aldagaiak Aukeratu → *m2*
Y-ardatzaren aldagaia Gehitu → *Prezioa*
- Beste aukera bat *Prezioa* eta *m2* aldagaiak aukeratzea da eta saguko eskubiko tekla klikatuz, zehazki *X-Y grafikoa*. Hurrengo kaxan X-ardatzaren aldagaia aukeratzen da, hau da *m2*.

Klikatzean ondorengo grafikoa agertzen da:

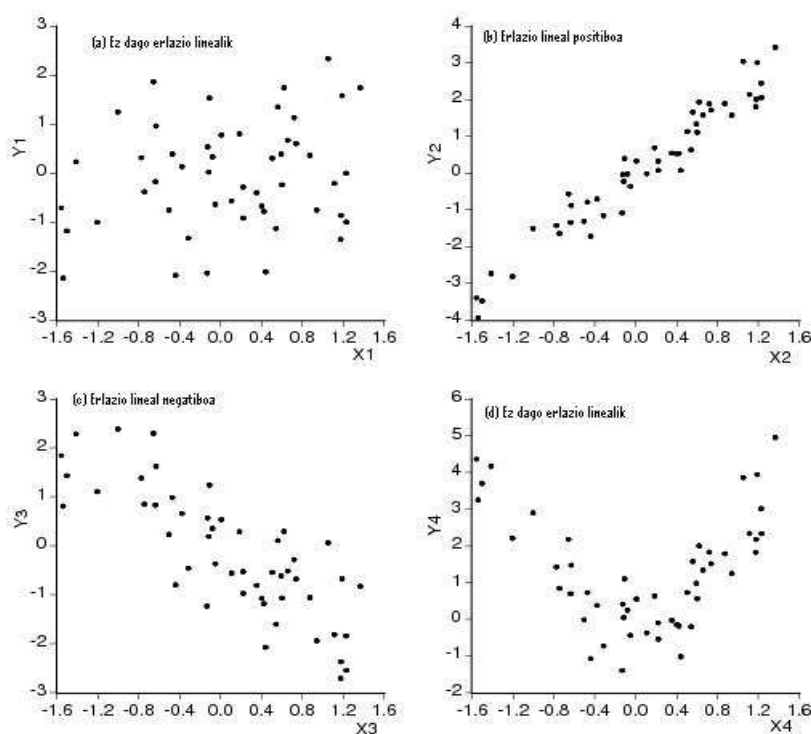
Grafiko honetan, puntu hodeiaz gain, erlazioa laburbiltzen duen zuzena ageri da. Gure adibidean garbi ikusten da prezioa eta tamainaren arteko erlazio lineal zuzena.

Grafikoan klikatuz gero, menu laguntzaile bat agertzen da eta ondorengoa egiteko erabili daiteke:

- Grafikoa esportatu (formato desberdinetara egiteko aukera egonik), *Gorde Windows metafitxategia (EMF) bezala ...*, *PNG, postscript (EPS) ...* eginez.
- *Gorde arbelan* aukeran, grafikoa memorian gordetzen da ondoren beste fitxategi batera esportatzeko.
- Gretleko saioan fitxategia gorde, *Gorde ikono bezala saioan* eginda.
- Fitxategian aldaketak egin daitezke *Editatu* erabiliz. Grafikoaren izenburua aldatzeko *Nagusia* aukeratu. Bestelako aldaketen artean letra mota eta tamaina, zuzena edo

puntuen kolorea, adierazitako aldagaien testu azalgarria edo egindako erregresio zuzena kentzea daude. Ardatzen eskala eta azalpena *X-ardatza* eta *Y-ardatza*-n aldatu daitezke. Datuen adierazpena *Lerroak* aukeran kontrolatzen da (zuzen edo puntu desberdinak) eta baita bere eskala eta aldagaien azalpenak. *Etiketak* aukerarekin, grafikoa testua erantsi daiteke eta *Irteera fitzategia* aukeran, grafikoa gordetzeko formato desberdinak agertzen dira.

1.12 Irudia: Sakabanatze diagrama



Sakabanatze grafikoak, aldagaien artean izan daiteken erlazioa (lineala edo ez) nolakoa den bereizten uzten digu. Bi aldagaien artean **erlazio lineal zuzena edo positiboa** dagoela esango dugu, *x* handitzean *y*-ren batezbesteko balioa ere handitzen denean (1.12 irudiko b grafikoa). Bi aldagaien artean **alderantzizko erlazio lineala edo negatiboa** dagoela esango dugu, *x* handitzean *y*-ren batezbesteko balioa murrizten denean (c grafikoa).

Kobariantza eta koerlazioa. Kobariantza bi aldagaien arteko erlazio linealaren neurri bat da. S_{xy} -gatik adierazten da eta horrela kalkulatzen:

$$S_{xy} = kob(x, y) = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})$$

non $\bar{x}$ eta $\bar{y}$, aldagaien batezbesteko aritmetikoak diren. Kobariantza aldagaien neurri unitateen menpean dagoenez, neurri unitate desberdinetan adierazitako aldagai bikoteen arteko erlazioak ezin dira konparatu. Kasu hauetan, *x* eta *y* aldagaien arteko **koerlazio koefiziente lineala** kalkulatzen da:

$$r_{xy} = koer(x, y) = \frac{S_{xy}}{S_x^* S_y^*} = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^N (y_i - \bar{y})^2}}$$

Koerlazio koefizienteak eta kobariantzak ikur berdina dute: positiboak dira aldagaien arteko erlazioa zuzena bada (1.12 irudiko b grafikoa), negatiboak alderantzizko erlazioa bada (1.12 irudiko c grafikoa) eta zero izango da x eta y independenteak badira (a grafikoa) edo erlazioa izanik, lineala ez bada (d grafikoa). Hartzen duten balioa ez dago aldagaien ordenaren arabera, hau da, $S_{xy} = S_{yx}$ eta $r_{xy} = r_{yx}$, ezta ere neurri unitateen menpean zeren ($-1 \leq r_{xy} \leq 1$) ematen baita. Koerlazio koefizientea balio absolutuan bat bada, aldagaiak linealki zehazki erlazionatzen dira eta datuak lerro baten gainean egongo dira.

Gretlen interesatzen zaizkigun aldagaiak markatu eta $Ikusi \rightarrow Koerlazio\ matrizea$ aukerak taula edo matrizea eskeintzen du, fitxategiko aldagaiak bikoteka duten koerlazio koefizienteak adieraziz. Etxebizitzen datuekin lortutako emaitzak hauek dira:

1.5 Taula: Aldagaien koerlazio matrizea

Koerlazio Koefizienteak, 1 - 50 behaketak erabiliz
%5eko esanguratasuna (alde bikoa) = 0,2787 n = 50 -rentzat

m2	Errefer	Prezioa
1,0000	0,0440	0,8690 m2
	1,0000	0,2983 Errefer
		1,0000 Prezioa

Prezioak beste aldagaiekin duen koerlazio koefizientea lehenengo errenkadan eta hirugarren zutabean agertzen da. Honela prezioa tamaina aldagaien arteko koerlazioa $r_{Prezioa, m2} = 0,869$ da, aldagaien arteko erlazio lineala zuzena eta sakona dela adieraziz. Koefiziente hau aldagai kuantitatiboen artean bakarrik definitzen denez, ez dugu *Erreforma* aldagaiarekin duen koerlazioa interpretatuko.

Bibliografia

Davidson, D. eta J. Mackinnon (2004), *Econometric Theory and Methods*, Oxford University Press, New York.

Greene, W. (2008), *Econometric Analysis*, 6. ed., Prentice-Hall, New Jersey.

Gujarati, D. (1997), *Econometría básica*, 4. ed., McGraw-Hill, México.

Heij, C., de Boer, P., Frances, P., Klok, T. eta H. Van Dijk (2004), *Econometric Methods with Applications in Business and Economics*, Oxford University Press, Oxford.

Peña, D. eta J. Romo (1997), *Introducción a la Estadística para las Ciencias Sociales*, McGraw-Hill, Madrid.

Ramanathan, R. (2002), *Introductory Econometrics with Applications*, 5. ed., South-Western, Ohio.

Stock, J. eta M. Watson (2003), *Introduction to Econometrics*, Addison-Wesley, Boston.

Verbeek, M. (2004), *A Guide to Modern Econometrics*, 2. ed., John Wiley, England.

Wooldridge, J. M. (2003), *Introductory Econometrics. A Modern Approach*, 2. ed., South-Western, Ohio.

2 Gaia

Erregresio linealeko eredu bakuna

Aurkibidea

2.1	Sarrera	26
2.2	Oinarrizko hipotesiak	29
2.3	Karratu Txikiarren Arruntetako metodoa	33
2.4	Doikuntzaren egokitasuna	46
2.5	Esanguratasun analisia eta konfidantza tarteak	47
2.6	Laburpena. Eraitzen aurkezpena	50

2.1 Sarrera

Gai honetan *erregresio linealeko eredu bakuna* zehazten, estimatzen eta analizatzen ikasiko dugu. Helburu honetarako behar den teoria adibide batekin batera eskeiniko da, Ramanathan (2002) liburuko data3-1 datu fitzategia erabiliz.

Eredu bakuna bi aldagai linealki erlazionatzen duen eredua da,

$$Y_i = \alpha + \beta X_i + u_i \quad i = 1, \dots, N \quad (2.1)$$

non,

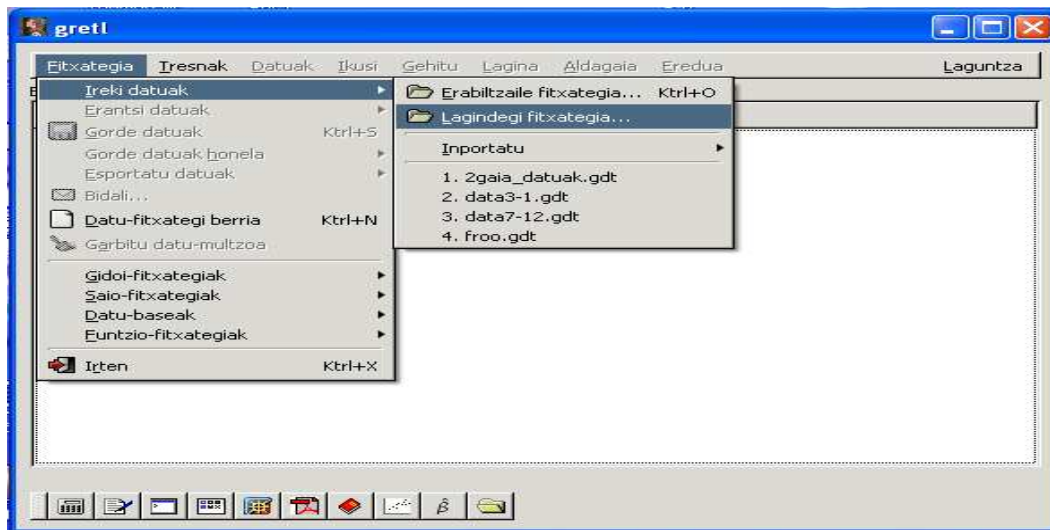
- *Y aldagai azaldua, aldagai dependentea edo endogenoa* den, hau da, azaldu edo analizatu nahi den aldagaia.
- *X aldagai azaltzailea, aldagai independentea edo exogenoa* den.
- α eta β , eredu bakuneko jatorria eta malda, *erregresioaren koefizienteak* diren. *Estimatu beharreko koefizienteen kopurua K bezala* definitzen badugu, orduan eredu bakunean $K = 2$ koefiziente daude estimatzeko.
- u errorea da, zorizko aldagaia, aldagai aleatorioa edo *perturbazioa*.
- Erabilitako i azpiindizeak *behaketa* adierazten du. Orokorrean, i azpiindizea gurutzatutako datuekin lan egiterakoan erabiltzen da eta daukagun laginaren datuak denborazko serieak badira, t azpiindizea erabiliko dugu.
- N *lagin tamaina* da, ikertzen ari garen aldagaien (Y, X) behaketa kopurua. Denborazko serieak ikertzerakoan, T erabiliko dugu lagin tamaina adierazteko.

Ereduan aldagai aleatorio bat (u) sartzearen arrazoiak desberdinak izan daitezke:

- 1 Aurrean ezinezko efektuak, bai egoera ekonomiko berezi batetik edo analizatzen ari garen testuinguruaren ezaugarrietatik eratortzen direnak eta banako edo entitate ekonomikoek dituzten lehentasunak edo gustoetatik sortarazten diren zenbagarri-ezinezko efektuak jasotzeko.
- 2 Interesgarriak zaizkigun aldagaien datuak biltzerakoan sortzen diren neurketa erroreak kontuan hartzeko.
- 3 Zehazpen erroreak biltzeko, aldagairen bat faltan egoterakoan edo ereduko zati sistematikoan agertzen diren erlazio ez linealak kontuan ez hartzerakoan sortzen direnak.

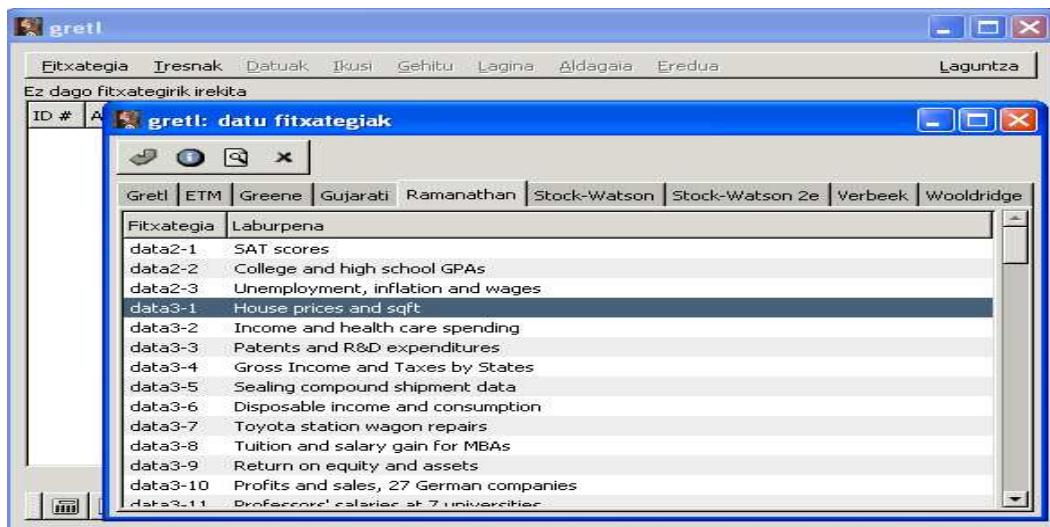
Behin eredu bakunaren zehazpena ikusirik, aurretik esandako datuen analisiarekin hastera goaz. Datu multzo hauek lortzeko: *Fitzategia* $\rightarrow$ *Ireki datuak* $\rightarrow$ *Lagindegi fitzategia...* aukeran.

2.1 Irudia: Datuen bilakaera



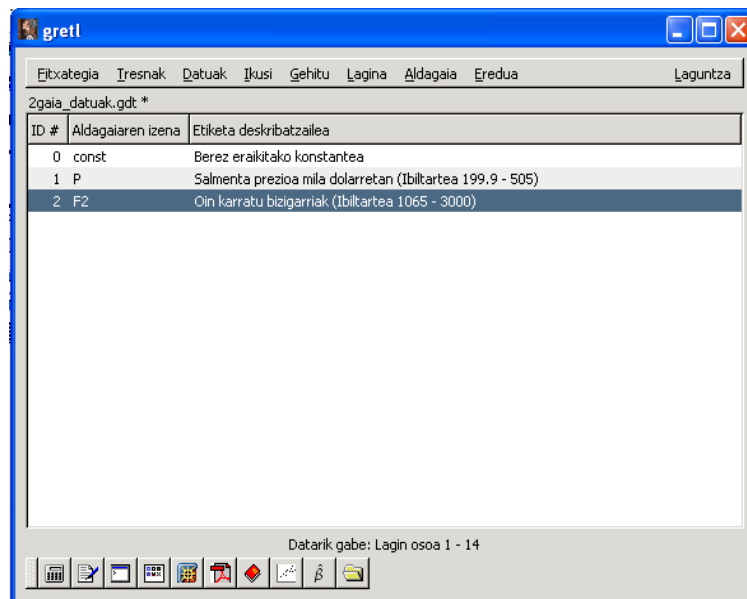
eta Ramanathan karpetaaren barnean *data3-1 House prices and sqft* aukeratu:

2.2 Irudia: Datu-fitxategiaren aukerapena



Segidan aldagaien izenak (“P” eta “F2”), deskribapena (“Salmenta prezioa mila dolarretan (Ibiltartea: 199,9-505)” eta “Oin karratu bizigarriak (Ibiltartea: 1065-3000)”) eta grafikoetan erakutsiko diren izenak (“P” eta “F2”) aldatuko ditugu. Horretarako lehen gaian ikusitako pausuak jarraitu behar dira. Lortzen den emaitza honakoa da:

2.3 Irudia: Datuen erakuspena



Ondoren datuak gorde egingo ditugu Gretl formatuan “2gaia-datuak.gdt” izenarekin. Bi aldagaien balioak ondorengo taulan erakusten dira,

2.1 Taula: “2gaia-datuak.gdt” fitxategiko datuak

i	P_i	F2	i	P	F2
1	199,9	1065	8	365,0	1870
2	228,0	1254	9	295,0	1935
3	235,0	1300	10	290,0	1948
4	285,0	1577	11	385,0	2254
5	239,0	1600	12	505,0	2600
6	293,0	1750	13	425,0	2800
7	285,0	1800	14	415,0	3000

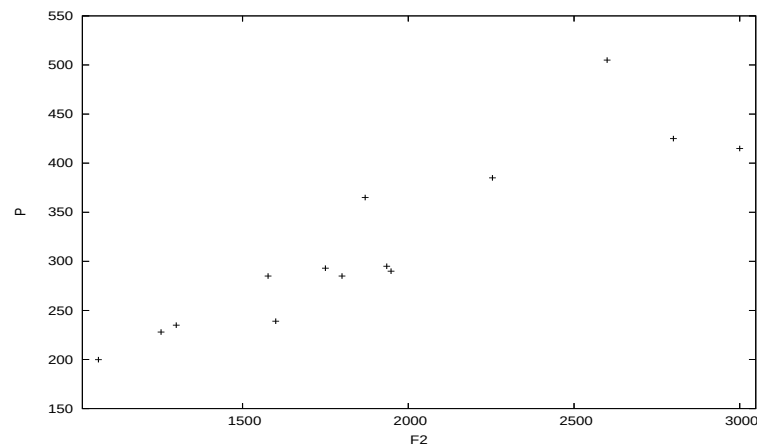
eta aldagai bien sakabanatzea 2.4 irudian aurkezten da, bertan erlazio lineal positiboa ikusten delarik.

Laburbilduz, adibide honetako eredu bakunaren zehazpena honakoa da:

$$P_i = \alpha + \beta F2_i + u_i \quad i = 1, \dots, 14 \quad (2.2)$$

analisi aurrera eramateko $N = 14$ behaketaz osaturiko lagina daukagu eta

- P_i “salmenta prezioa” aldagai dependentearen, endogenoaren edo azalduaren i behaketa da.

2.4 Irudia: (Y, X) ren sakabanatze diagrama

- $F2_i$ “oin karratu bizigarri” aldagai independentearen, exogenoaren edo azaltzailearen i behaketa da.
- Estimatu beharreko koefizienteak α eta β dira. 2.4 irudia begiratzuz, bai jatorria eta baita malda ere, positiboak direlakoan gaude.
- Testuinguru honetako perturbazioak, u_i , azalera (oin karratutan) berdineko etxebizitzaren arteko salmenta prezioen desberdintasunak azaltzen dituen ezaugarriak biltzen ditu: kokalekua, etxearen orientazioa, bistak, argi naturalaren kopurua, etab..

2.2 Oinarrizko hipotesiak

Eredua zehaztu ondoren estimazioaren testuingurua finkatu behar dugu. Horretarako **oinarrizko hipotesi** batzuk betetzen direla suposatuko dugu. Behin estimatzen, doikuntzaren egokitasuna neurtzen, beharrezkoak diren kontrasteak egiten eta aurreanak lortzen ikasten dugunean (kurtsoaren amaieran), oinarrizko hipotesi batzuk erlaxatzen hasi gaitezke. Baina bitartean denak betetzen direla suposatuko dugu. Oinarrizko hipotesi hauek erregresioaren elementu desberdinei buruzkoak dira.

Forma funtzionalari dagokionez

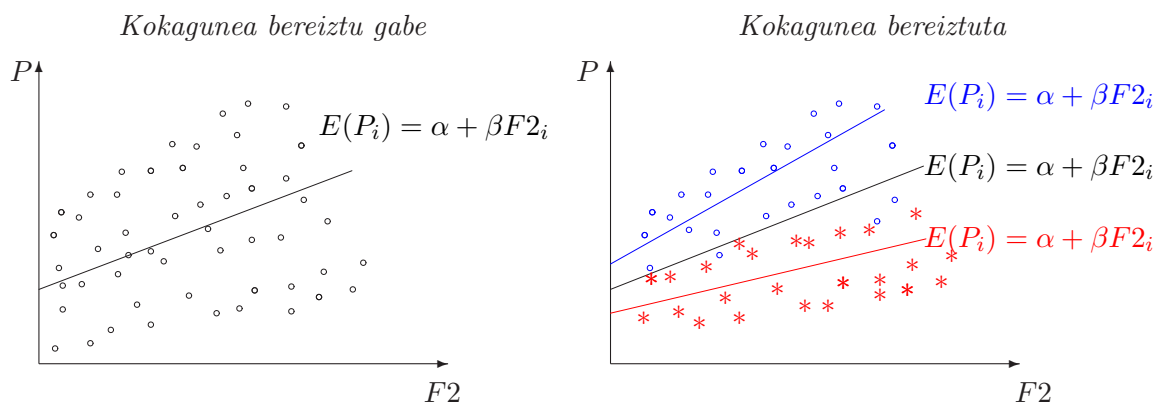
1. *Eredua koefizienteekiko lineala da.* Kurtsoan zehar estimatuko diren ereduak koefizienteekiko linealak dira. Demagun hiri bateko etxebizitzaren batezbesteko prezioa (P) azaldu nahi dugula duten azaleraren ($F2$) menpean (oin karratutan), orduan koefizienteekiko lineala den eredu bat $P_i = \alpha + \beta F2_i + u_i$ litzateke. Bestalde, ereduak ez da zertan aldagai azaltzailearekiko lineala izan behar, hau da, $P_i = \alpha + \beta F2_i^2 + u_i$ ereduak

estimatzea posible litzateke nahiz eta etxebizitzaren azalerak prezioaren duen eragina lineala ez izan, koadratikoa baizik.

Koefizientei dagokienez

2. *Koefizienteak konstante mantentzen dira laginean zehar.* Hasiera batean, aldagai azaltzaileak duen eragina laginean zehar konstante mantentzen dela suposatuko dugu. Aurreko adibidearekin jarraituz, 2.5 irudiko ezkerreko grafikoko puntu hodeitik pasatzen den “erdiko zuzena” estimatzea interesatuko litzaiguke.

2.5 Irudia: Bilboko etxebizitzaren prezioa daukaten azalera bizigarriarekiko



Hala ere, 2.5 irudiko eskubiko grafikoa ikusten den bezala, etxebizitza batzuk hiri zen-trukoak baldin badira (urdinez daudenak) eta besteak hiriaren kanpoaldekoak (gorriz daudenak) badira, orduan gerta daiteke “erdiko zuzen bien” (urdira eta gorria) estimazioa interesatzea horrela etxebizitzaren kokapena bereizten baita. Kasu honetan koefizienteak ez dira konstante mantentzen, desberdinak baitira kokagunearen arabera.

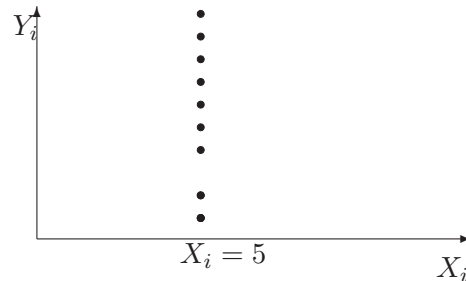
Aldagai azalduari dagokionez

3. *Aldagai azaldua koantitatiboa da.* Aldagai azaldu koalitatiboa daukan eredu bat estimatzeko, kurtso honetan ikasiko ditugun estimazio metodoek ez dute balio.

Aldagai azaltzaileari dagokionez

4. *X aldagai azaltzailearen lagin bariantza (S_X^2) ezin da zero izan eta gainera $N \geq K = 2$ izan behar da.* Hipotesi hau koefizienteak (jatorria eta malda) identifikatzeko beharrezkoa da. Hasteko, estimatu behar diren koefizienteen kopurua behaketen kopurua baino handiago izanez gero, orduan estimazioa aurrera eramateko ez daukagu informazio nahikorik. Bestalde, aldagai azaltzailearen lagin bariantza zero izango balitz, (adibidean, $S_{F2}^2 = 0$), hau da, $F2$ termino konstantearen konbinazio lineala izango balitz (adibidez, $F2 = 5 \times \text{termino konstantea} = 5 \times 1 = 5$), edo beste era batera esanda, etxebizitzek oin karratu berdina izango balute ($F2_i = 5 \forall i$, adibidez), orduan malda identifikatzea ezinezkoa izango litzateke. Egitez, 2.6 irudiko grafikoa ikusten den bezala, $P_i = \alpha + \beta 5 + u_i$ moduko eredu batean ezinezkoa da aldagai azalduaren (P) sakabanatzea azaltzea, aldagai azaltzaileak ez duelako inolako aldakortasunik.

2.6 Irudia: Eredua: $Y_i = \alpha + \beta \times 5 + u_i$, non $S_X^2 = 0$ den

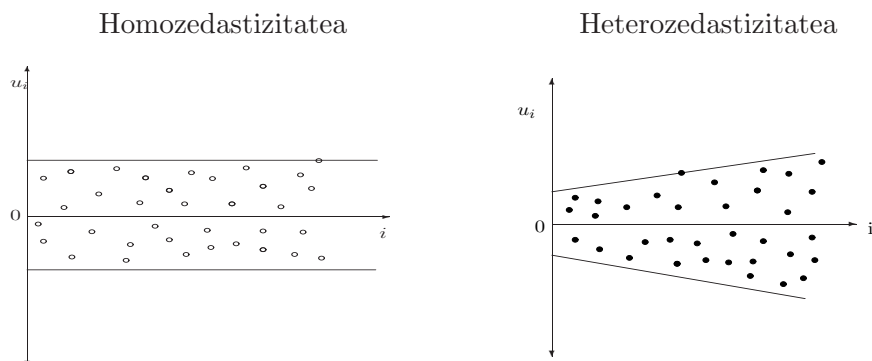


5. *Aldagai azaltzailea ez da aleatorioa* eta ondorioz ez dago errorerekin neurtuta. Gure adibidean etxebizitzaren azalera oin karratutan zehaztasun osoz (inolako milimetroko errorerik gabe) neurtuta dagoela inplikatzeko du.
6. *Eredua ondo zehaztuta dago.* Orokorrean azalduz, ereduak ezin du aldagai nabaririk barneratu gabe utzi eta ezta kontrakoa ere, hau da aldagai ez nabariren bat kontuan hartu. Eredu bakunean, hipotesi hau ezartzerakoan, Y aldagai azaldua azaltzeko behar den aldagai azaltzaile bakarra X aldagaia dela eskatzen ari gara.

Perturbazioari dagokionez

7. *Perturbazioen populazio batezbestekoa zero da*, $E(u_i) = 0$. Aurresan edo estimaezina den errorearen batezbestekoa zero izateak ereduaren zati sistematikoa edo analizatu nahi den batezbesteko portaera $E(Y_i) = \alpha + \beta X_i$ izatea dakar (adibidean, $E(P_i) = \alpha + \beta F_i$).
8. *Perturbazioen populazio bariantza konstantea da.* Aldagai aleatorioaren edo perturbazioaren aldakortasuna laginean zehar konstante mantentzen dela suposatuko dugu 2.7 irudiko ezkerreko grafikoan ikusten den bezala.

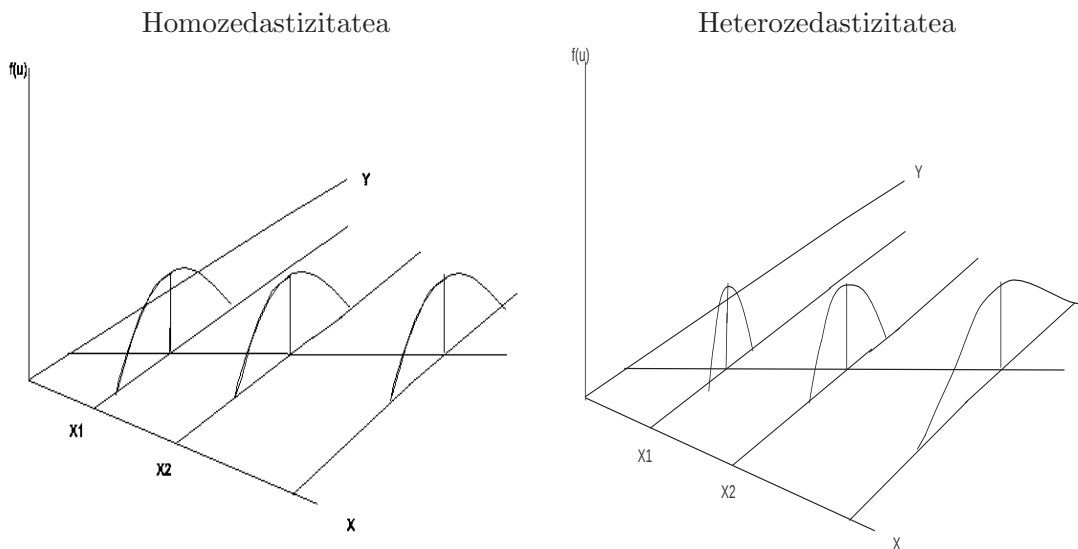
2.7 Irudia: Perturbazioen bariantza laginean zehar



Horrela, 2.8 irudiko ezkerreko grafikoan azaltzen den bezala, aldagai azaltzaileen balioak emanik, aldagai azalduak har ditzakeen balio posibleen tartearen zabalera berdina da (kanpai guztien oinarriaren diametroa berdina da) eta balio bakoitzak irteteko duen probabilitatea X aldagaiak hartzen duen balioarekiko independentea da (kanpai guztien itxurak berdinak dira).

Kontrako kasuan perturbazio heterozedastikoak ditugu, perturbazioen bariantza lagin-nean zehar aldatzen delarik. Horrela, 2.7 irudiko eskubiko grafikoan bariantza gora-korrek dituen perturbazioak marraztu dira.

2.8 Irudia: Perturbazioen bariantza aldagai azaltzailearekiko



Aldagai azaltzailearekiko duen interpretazioa ulertzeko, 2.8 irudiaren eskubiko grafiko-ra begiratu behar dugu. Demagun grafikoko aldagaiak gure adibidekoak direla. Etxebizitzak txikiak direnean, ($F2=$ oin karratu gutxi) prezioek har ditzaketen balio posibleak oso antzekoak izateko probabilitatea handia da eta beraz bariantza txikia da (kanpaiaren diametroa txikia da eta puntaduna). Baina pisuen azalera handitzen doan heinean, prezioek har ditzaketen balio posibleen tartea handiagoa da eta beraz bariantza handiagoa du (kanpaiaren diametroa zabalagoa eta altuera txikiagokoa).

Horrela bada, azalera txikiko etxebizitzaren prezioak nahiko antzekoak dira baina azalera handitzerakoan prezio posibleen tartea ere handitu egiten da eta ondorioz prezio oso desberdinetako etxebizitzak handiak aurkitu ahal dira.

9. *Perturbazioek ez daukate autokoerlazorik.* Momentuz, perturbazio desberdinen arteko koerlazioa zero dela suposatuko dugu ($koer(u_i, u_j) = r_{u_i, u_j} = 0; i \neq j$). Hortaz, beraien arteko kobariantza zero izango da ere, $kob(u_i, u_j) = 0 i \neq j$.
10. *Perturbazioek banaketa Normala jarraitzen du.* Azken hipotesi hau, aurrerago ikusten den bezala, ez da estimazioarentzat beharrezkoa, ezta estimatzailearen propietateak

lortzeko ere, inferentzia (kontrasteak) egitea edo konfidantza tarteak ateratzea interesatzen zaigunean, izango da soilik beharrezkoa.

2.3 Karratu Txikien Arruntetako metodoa

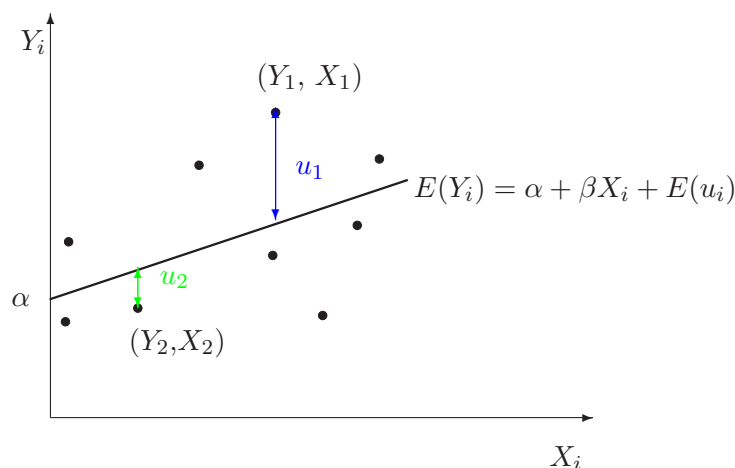
Analizatuko dugun testuinguruaren ezaugarriak (oinarrizko hipotesiak) finkatuz, eredu bakunak zehaztasunez zer adierazten duen ikustera goaz. Hasierako (2.1) ereduak egitura berdineko N ekuazio (berdintzak) biltzen ditu:

$$\begin{aligned} Y_1 &= \alpha + \beta X_1 + u_1 \\ Y_2 &= \alpha + \beta X_2 + u_2 \\ &\vdots \\ Y_i &= \alpha + \beta X_i + u_i \\ &\vdots \\ Y_N &= \alpha + \beta X_N + u_N \end{aligned}$$

Ekuazio hauek begiratzuz, (Y_i, X_i) punto bakoitzatik β maldako zuzen bat pasatzen dela ikusi daiteke. Eta malda berekoak direnez, N zuzenak paraleloak dira. Jatorriari dagokionez, zuzen bakoitzaren jatorria lortzeko, α “batezbesteko jatorriari” zuzen bakoitzari dagokion perturbazioa (u_i) gehitu behar zaio, hau da, zuzen bakoitzaren jatorria $\alpha + u_i$ balioa da.

Ondorengo 2.9 irudian, N zuzen hauetariko bi marrazten dira, lehenengoko biak hain zuzen. Beltzez dagoen lerroa “erdiko zuzenari” dagokio, perturbazio zero izaterakoan lortzen den zuzena.

2.9 Irudia: Erregresio eredu bakuna



Lehen behaketari (Y_1, X_1) dagokion zuzena (lerro urdina) erdikoarekiko paraleloa da, malda berekoa baita, eta berarekiko duen distantzia u_1 balioak neurtzen du. Bigarren behaketari (Y_2, X_2) dagokion lerroak ezaugarri berberak ditu, erdiko zuzenarekiko paraleloa eta erdikotik

u_2 distantziara aurkitzen da. Era berean, zuzen guztiak marraztu ditzakegu, hau da puntu bakoitzarentzat bat.

Erregresio eredu bakun baten **estimazioak** ez du N zuzen guztien estimazioa helburutzat, baizik eta “erdiko zuzenaren” estimazioa. Grafikoki, eredu bakun bat estimatzerakoan datuei hoberen doitzen diren β malda eta α “batezbesteko” jatorria estimatu nahi dira. Teknikoki, ekonometrian oinarrituz, laginetik eratorritako $(Y_1, X_1), (Y_2, X_2), \dots, (Y_N, X_N)$ behaketek ematen duten informaziotik, interesatzen zaigun aldagaiaren (Y_i) batezbesteko portaera ($\alpha + \beta X_i$) zein den kalkulatu nahi dugu.

Ondorioz, eredua estimatzera joan baino lehen, kontzeptu berri batzuk definitu beharrean gaude. Estimatzera interesatzen zaigun “erdiko zuzena” **Populazioaren Erregresio Funtzioa (PEF)** bezala ezagutzen da eta populazio koefiziente ezezagunen (α eta β) menpekoa da. Ereduko zati sistematikoa (estimagarria den zatia) jasotzen du eta aldagai azalduaren batezbesteko portaera neurtzen du:

$$\begin{aligned} E(Y_i) &= E(\alpha + \beta X_i + u_i) = \\ &= \alpha + \beta X_i + \underbrace{E(u_i)}_{=0} = \alpha + \beta X_i. \end{aligned}$$

Ereduko **perturbazioa** (estimaezinezko zatia) aldagai azalduaren benetako balioa eta populazio erregresio funtzioaren arteko diferentzia da:

$$u_i = Y_i - \alpha - \beta X_i.$$

eta ereduoko zati sistematikoak azaldu ezin izan duen guztia biltzen du.

Lagin konkretu batetik lorturiko azken emaitza **Lagin Erregresio Funtzioa (LEF)** bezala ezagutzen da baina praktikan **estimatuak ereduari** buruz hitz egiten da gehienetan. LEF lortzeko erregresio koefizienteak estimatu ($\hat{\alpha}, \hat{\beta}$) behar dira:

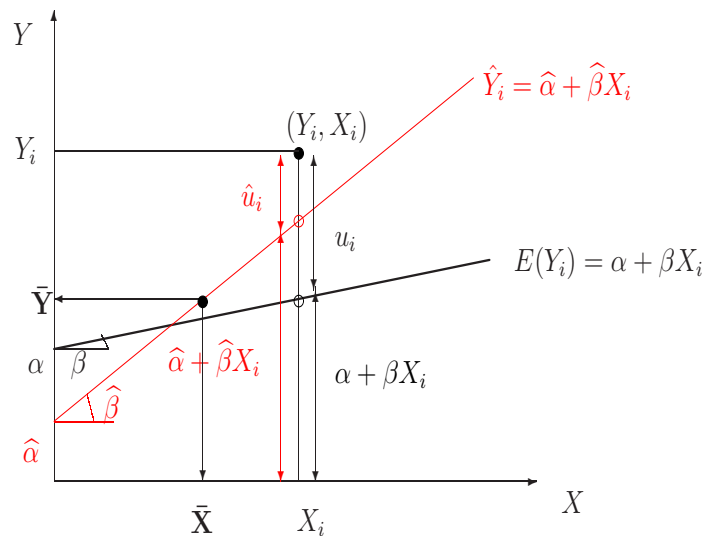
$$\hat{Y}_i = \widehat{E(Y_i)} = \hat{\alpha} + \hat{\beta} X_i.$$

Ereduko **hondarra** berriz, azaldu nahi dugun aldagaia eta lagin erregresio funtzioaren arteko diferentzia da:

$$\begin{aligned} \hat{u}_i &= Y_i - \hat{Y}_i \\ &= Y_i - \hat{\alpha} - \hat{\beta} X_i \\ &= \alpha + \beta X_i + u_i - \hat{\alpha} - \hat{\beta} X_i \\ &= (\alpha - \hat{\alpha}) + (\beta - \hat{\beta}) X_i + u_i \end{aligned} \tag{2.3}$$

Estimazio errore bat da, hau da, eredua estimatzerakoan egiten diren errore guztien batura. Hondarraren barnean bi motako erroreak izango ditugu: elementu estimagarriak (α, β) estimatzerakoan ($\hat{\alpha}, \hat{\beta}$) egiten den erroretik eratortzen dena ($\alpha - \hat{\alpha}, \beta - \hat{\beta}$) eta estimatu ezinezko elementutik (u_i) eratortzen dena. Ondorioz, benetan garrantzitsua da perturbazio eta hondarrak bereiztea eta ez nahastea. Irudikaturiko 2.10 grafikoan populazio erregresio funtzioa

2.10 Irudia: Populazio eta lagin erregresio funtzioak



(beltzez), bere α jatorria eta β maldarekin (populazio koefizienteak) irudikatuta dago. Bertan, aldagai azalduaren edozein i behaketa (Y_i) lortzeko, ereduaren zati sistematikoari $\alpha + \beta X_i$ (PEF gainean kokaturik) perturbazioa (u_i) gehitu egin behar zaiola ($Y_i = \alpha + \beta X_i + u_i$) erraz ikus daiteke.

Lagin erregresio funtzioa eta estimatutako koefizienteak ($\hat{\alpha}$ eta $\hat{\beta}$) gorritz jarrita daude. PEF eta LEF-ren arteko diferentzia koefizienteak estimatzerakoan egiten den erroreengatik ($\hat{\alpha} \neq \alpha$, $\hat{\beta} \neq \beta$) ematen da.

LEF oinarritzat hartuz, aldagai azalduaren edozein i behaketa (Y_i) lortzeko, estimatutako zati sistematikoari ($\hat{\alpha} + \hat{\beta} X_i = \hat{Y}_i$, estimatutako ereduari) dagokion hondarrak hartzen duen balioa ($\hat{u}_i$) gehitu egin behar zaio, horrela $Y_i = \hat{Y}_i + \hat{u}_i$ lortuz. Azkenik, lagin erregresio funtzioa batezbestekoen puntutik ($\bar{Y}, \bar{X}$) pasatzen dela ere ikus daiteke.

Eredu bakuneko **koefizienteen interpretazioari** dagokionez:

- $\alpha = E(Y_i | X_i = 0)$: aldagai azalduaren batezbesteko balioa edo esperotako balioa da aldagai azaltzaileak hartzen duen balioa zero denean.
- $\beta = \frac{\partial E(Y_i)}{\partial X_i} = \frac{\Delta E(Y_i)}{\Delta X_i}$: aldagai azaltzailea unitate batean handitzerakoan, aldagai azalduaren esperotako gehikuntza edo batezbesteko gehikuntza β unitatekoa da.

Gure adibidera bueltatuz, koefizienteen interpretazioak honakoak dira:

$\alpha = E(P_i | F2_i = 0)$: etxebizitzaren batezbesteko prezioa da (mila dolarretan) etxebizitzak duen azalera zero oin karratu denean. Kasu honetan, koefiziente hau zero izatea esperoko

genuke, azalera gabeko etxebizitzari buruz hitz egiteak ez duelako inolako zentzurik. Hala ere, normalean konstantea, eredutik ez da kentzen zeren bestela emaitzen betiko interpreta-zioa galdu egiten baita. Bestalde adibide honetan etxebizitzak duen hasierako batezbesteko prezioa (tramiteak,...) bezala kontsideratu daiteke eta kasu hau emango balitz, orduan posi-tiboa izatea espero genuke.

$\beta = \frac{\partial E(P_i)}{\partial F2_i}$: etxebizitzaren azalera oin karratu batean handitzerakoan espero dugun prezioa-ren batezbesteko gehikuntza β mila unitatekoa da.

Ondoren **Karratu Txikiaren Arruntetako** estimatzailea aterako dugu eta horretarako hondar karratuen batura minimizatu behar da:

$$\min_{\hat{\alpha}, \hat{\beta}} \sum_{i=1}^N \hat{u}_i^2 = \min_{\hat{\alpha}, \hat{\beta}} \sum_{i=1}^N (Y_i - \hat{Y}_i)^2 = \min_{\hat{\alpha}, \hat{\beta}} \sum_{i=1}^N (Y_i - \hat{\alpha} - \hat{\beta}X_i)^2 \quad (2.4)$$

Minimizazio honetatik α eta β koefizienteen estimatzaileen adierazpenak ($\hat{\alpha}$ eta $\hat{\beta}$) lehen ordenako baldintzak askatuz lortzen dira. Hau da, helburu funtzioaren deribatu partzialak berdin zero egiterakoan lortzen diren ekuazioak (**ekuazio normalak**) askatuz:

$$\frac{\partial \sum_{i=1}^N \hat{u}_i^2}{\partial \hat{\alpha}} = -2 \sum (Y_i - \hat{\alpha} - \hat{\beta}X_i) = 0 \Rightarrow \sum \underbrace{(Y_i - \hat{\alpha} - \hat{\beta}X_i)}_{\hat{u}_i} = 0 \quad (2.5)$$

$$\frac{\partial \sum_{i=1}^N \hat{u}_i^2}{\partial \hat{\beta}} = -2 \sum (Y_i - \hat{\alpha} - \hat{\beta}X_i)X_i = 0 \Rightarrow \sum \underbrace{(Y_i - \hat{\alpha} - \hat{\beta}X_i)X_i}_{\hat{u}_i X_i} = 0 \quad (2.6)$$

Ekuazio normal biak askatuz KTAko estimatzaileen adierazpenak lortzen ditugu:

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} \quad (2.7)$$

$$\hat{\beta} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{S_{YX}}{S_X^2} \quad (2.8)$$

Ekuazio normalek, estimatzaileen adierazpenak ateratzeko informazioaz gain, informazio ge-hiago eskaintzen dute. Hasteko, lehen ekuazio normalak (2.5) hondarren batura zero de-la adierazten du, hortaz bere batezbestekoa ere zero izango da: $\bar{\hat{u}} = \frac{1}{N} \sum \hat{u}_i = 0$. Bi-garren ekuazio normalak (2.6), hondarrak ($\hat{u}$) aldagai azaltzailearekiko (X) ortogonalak direla ($\sum \hat{u}_i X_i = 0$) adierazten duenez, beraien arteko lagin kobariantza zero izango da:

$$S_{\hat{u}, X} = N^{-1} \sum_{i=1}^N (X_i - \bar{X})(\hat{u}_i - \bar{\hat{u}}) = N^{-1} \left(\sum_{i=1}^N \hat{u}_i X_i - N \bar{X} \bar{\hat{u}} \right) = 0 \quad (2.9)$$

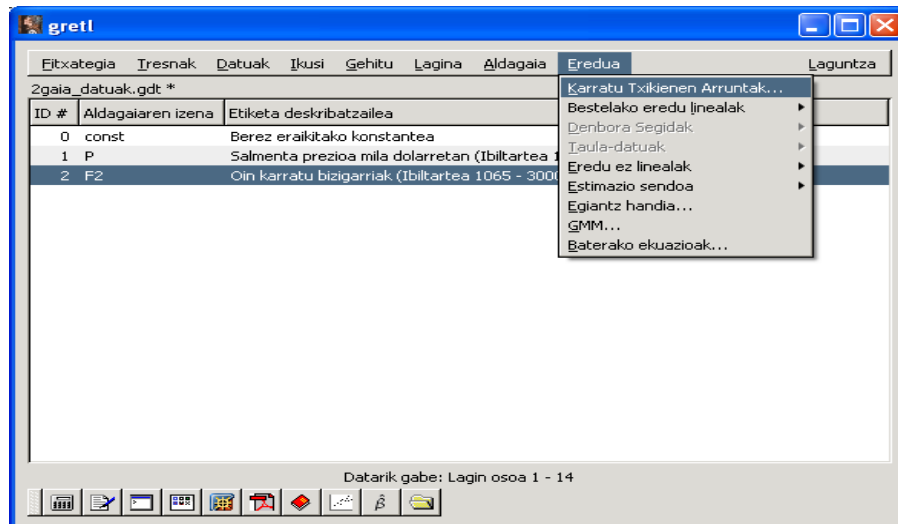
eta ondorioz beraien arteko koerlazio koefizientea ere zero izango da, hau da: $r_{\hat{u}, X} = S_{\hat{u}, X} / S_{\hat{u}} S_X = 0$. Beraz, eredu bat estimatu ondoren azaldu gabe gelditzen den zatian, hots hondarretan ($\hat{u}$), ez dago X aldagaiak azaldu dezakeenik.

Ondoren adibidera itzuliko gara emandako kontzeptu berriak argitzeko asmoz. Estima deza-gun ondorengo eredu bakuna:

$$P_i = \alpha + \beta F2_i + u_i \quad i = 1, \dots, 14 \quad (2.10)$$

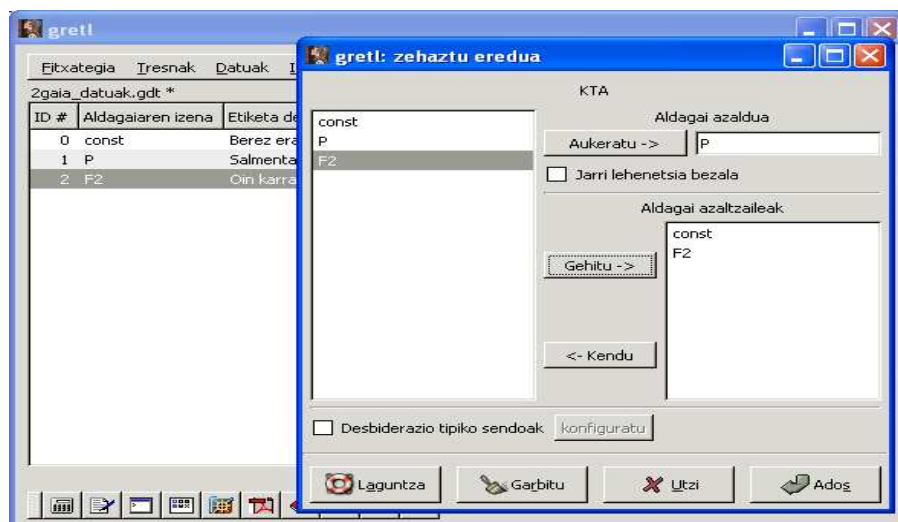
non α eta β koefizienteen interpretazioak jadanik eginda dauzkagun. Eredua karratu txikiaren arruntetako metodoaren bitartez estimatzeko, *Eredua* aukeratu eta gero *Karratu Txikiaren Arruntak...*

2.11 Irudia: Karratu Txikiaren Arruntetako estimatzailea



hemen aldagai azaldu bezala etxebizitzaren prezioa P hartzen dugu eta aldagai azaltzaile bezala etxebizitzaren azalera oin karratutan $F2$:

2.12 Irudia: Ereduaren zehazpena



KTako estimazioaren emaitzak honako taulan agertzen dira:

Eredua 1: KTA estimazioak 14 behaketak erabiliz 1–14

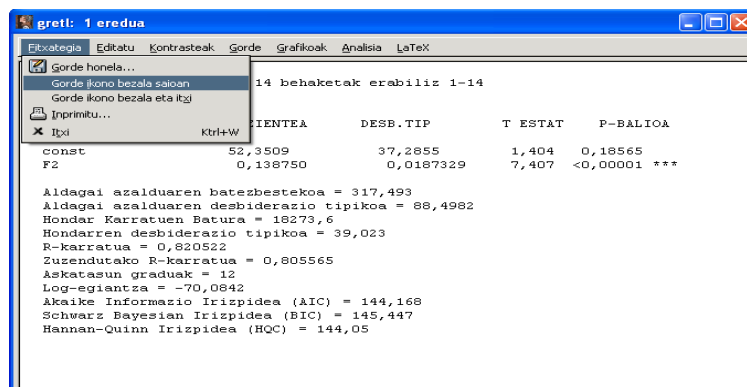
Aldagai azaldua: P

Aldagaia	Koefizientea	Desb. Tipikoa	t -estatistikoa	p-balioa
const	52,3509	37,2855	1,4041	0,1857
F2	0,138750	0,0187329	7,4068	0,0000

Aldagai azalduaren batezbestekoa	317,493
Aldagai azalduaren Desb. Tip.	88,4982
Hondar Karratuen Batura	18273,6
Hondarren desbideratze tipikoa ($\hat{\sigma}$)	39,0230
R^2	0,820522
Zuzendutako $\bar{R}^2$	0,805565
Askatasun graduak	12
Log-egiantza	-70,084
Akaike Informazio Irizpidea	144,168
Schwarz Bayesian Irizpidea	145,447
Hannan–Quinn Irizpidea	144,050

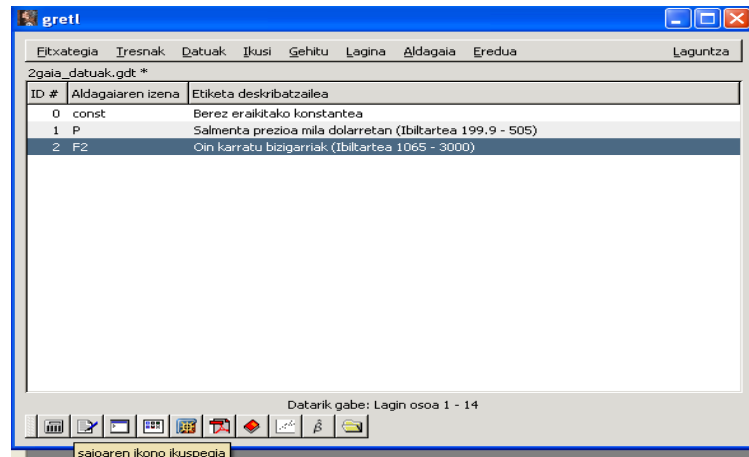
Emaitzak aztertzen hasi baino lehen, eredua ikono bezala gorde egingo dugu gero beranduago berreskuratzeko asmoarekin. Gordetzeko *Fitxategia* aukeratu eta *Gorde ikono bezala saioan* klikatu:

2.13 Irudia: Ikono bezala gordetzen



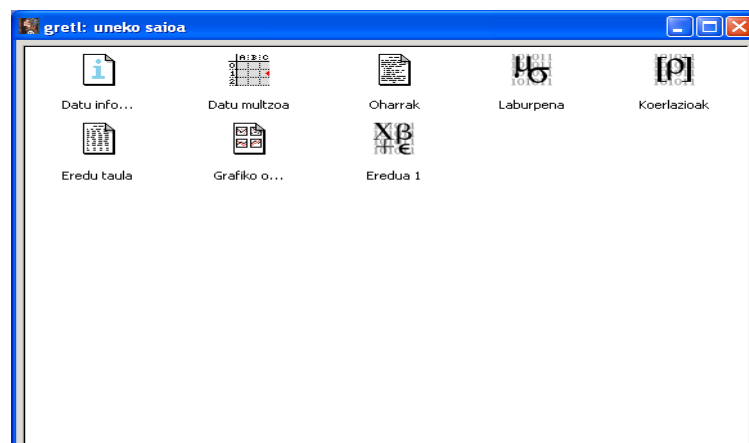
Hau egiterakoan ereduak ikono bezala gordeta gelditzen da **GRET**L programaren barneko **USER** karpetan. Berreskuratzeko ezkerreko behealdean agertzen den ikonoen menuan (ezkerretik hasita laugarren laukia) aukeratu *saioaren ikono ikuspegia*

2.14 Irudia: Estimazio emaitzak ikonoetatik berreskuratzen



eta *Eredua 1* jartzen duen ikonoari klik-bikoitza emanik ereduaren estimazio emaitzak berreskuratzen dira: Beranduago, beste ereduaren bat estimatu eta ikono bezala gordetzen badugu

2.15 Irudia: Ikonoen ikuspegia



Gretlek *Eredua 2* izena jarriko dio.

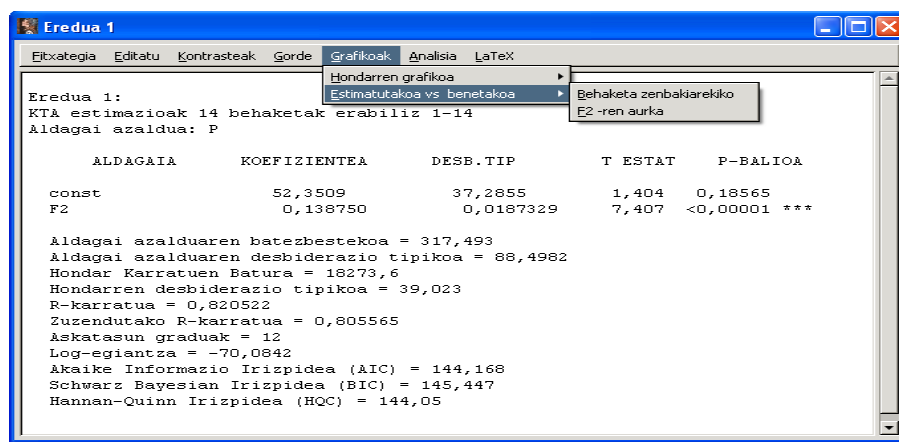
Emaitzen taulara itzuliz lehen zutabean ereduaren barneratutako aldagai azaltzaileen izenak agertzen dira: termino konstantea (const) eta etxebizitzaren azalera oin karratutan ($F2$). Bigarren zutabean, aldagai azaltzaile bakoitzari dagokion koefizientearen KTAko estimazioak dauzkagu. Honela, lagin erregresio funtzioa lortu dezakegu:

$$\hat{P}_i = 52,3509 + 0.138750 F2_i \quad i = 1, \dots, 14. \quad (2.11)$$

Estimatutako koefizienteak interpretatuz $\hat{\alpha} = 52,3509$ azalera gabeko etxebizitzaren batezbesteko prezio **estimatus** da mila dolarretan edo bestela “hasierako batezbesteko prezioa” estimatus. Etxebizitzaren azalera oin karratu batean handitzerakoan espero dugun prezioaren gehikuntza **estimatus** $\hat{\beta} = 0,138750$ mila dolarretakoa da edota $0,138750 \times 1000 = 138,750$ dolarretakoa. Hortaz espero genituen zeinuak atera dira.

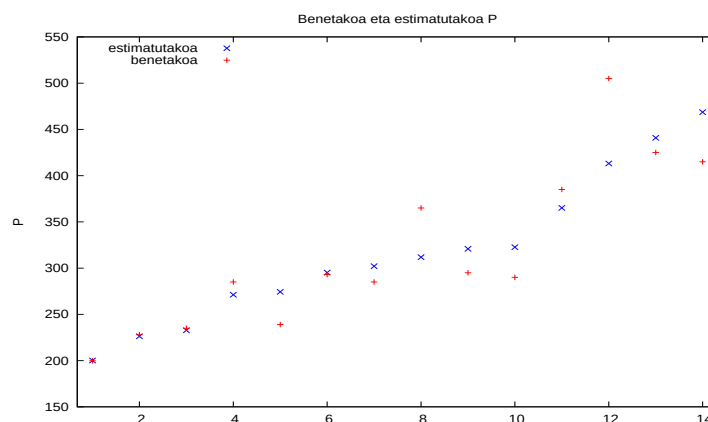
Ondoren interesgarriak diren grafikoak aterako ditugu. Horretarako *Grafikoak* aukeratu eta ondoren $\rightarrow$ *Estimatutakoa vs benetakoa* atalean

2.16 Irudia: Estimatutako vs benetako aldagai azaldua



grafikoa irudikatzeko bi aukera agertuko zaizkigu. Bata *Behaketa zenbakiarekiko* eta bestea *F2 -ren aurka*. Lehenengokoa hartzerakoan irteten zaigun grafikoa honakoa da

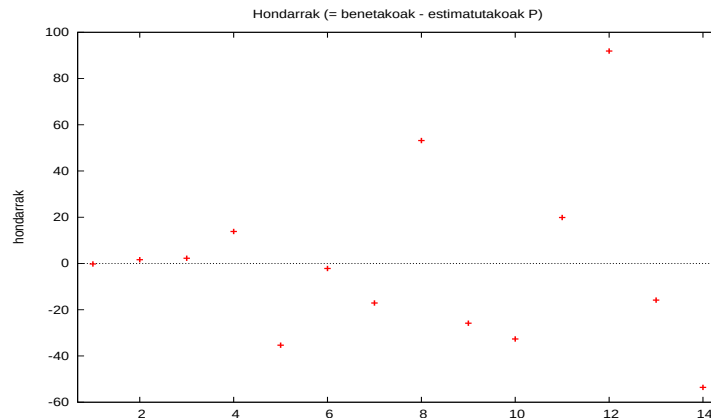
2.17 Irudia: Estimazio emaitzen erakuspena



zeinetan aldagai azalduaren benetako balioak (P_i) eta estimatutakoak ($\hat{P}_i$) agertzen diren. Grafikoa begiratzerakoan badirudi, urruntzen diren hiru-lau puntutan ezik, nahiko ondo hurbildu garela eta estimazioa nahiko ona dela.

Bestalde, *Grafikoak* atalean *Hondarren grafikoa* $\rightarrow$ *Behaketa zenbakiarekiko* aukeratzen badugu lortzen den grafikoa honakoa da. Grafiko honetan hondarrak zero balioaren inguruan

2.18 Irudia: Hondarren grafikoa



banaturik daudela ikus dezakegu eta horrela irten behar da $\hat{u} = 0$ delako. Baina badirudi laginean zehar mugitzen goazen neurrian hondarrak duten sakabanatzea handitzen doala. Eraitza honek arazoren bat dagoela adierazten digu. Posible litzateke homozedastizitatearen oinarritzko hipotesia ez mantentzea gure datuentzat eta hortaz perturbazioak heterozedastizitatekoak izatea, edo erdua ondo zehaztuta dagoelaren oinarritzko hipotesia ez betetzea, sakabanatze gorakorra duen aldagai nabariren baten omisioa izatea.

Estimatutako balioak eta hondarrak grafikoki ikusteaz gain, hartzen dituzten balioak zeintzuk diren jakiteko, *Analisia* aukeran *Erakutsi benetakoak, estimatutakoak, hondarrak* aukeratuz, ateratzen den emaitza honakoa izango da:

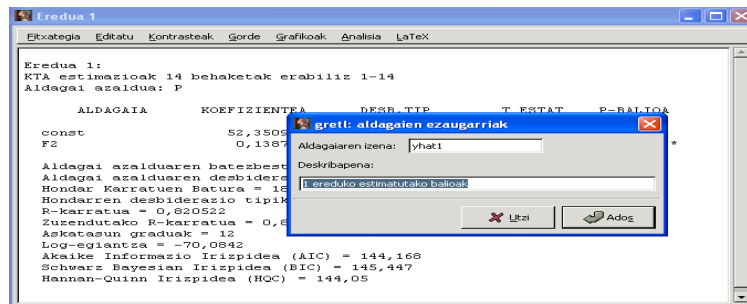
2.19 Irudia: Estimazio emaitzen erakuspena

gretl: erakutsi datuak			
Ereduaren estimazio ibiltartea: 1 - 14			
Hondarren desbiderazio tipikoa = 39,023			
Behaketa	Pestimatutakoa	hondarrak	
1	199,9	200,1	-0,2
2	228,0	226,3	1,7
3	235,0	232,7	2,3
4	285,0	271,2	13,8
5	239,0	274,4	-35,4
6	293,0	295,2	-2,2
7	285,0	302,1	-17,1
8	365,0	311,8	53,2
9	295,0	320,8	-25,8
10	290,0	322,6	-32,6
11	385,0	365,1	19,9
12	505,0	413,1	91,9
13	425,0	440,9	-15,9
14	415,0	468,6	-53,6

Bigarren zutabeko balioak aldagai azalduarenak dira. Hirugarren zutabekoak berriz, estimatutako balioak dira eta hauek ateratzeko $\hat{P}_i = 52,3509 + 0.138750 F2_i$ erabili da. Azken zutabeko balioak ateratzeko berriz, $\hat{u}_i = P_i - \hat{P}_i$ erabili da, hau da, bigarren eta hirugarren zutabe bien arteko kenketa.

Balio hauek gorde nahi izanez gero *Gorde* barneko aukeren artean *Estimatutako balioak* hartu eta honako leihatila aterako zaigu: non aldagai azaldu estimatuarentzat “yhat1” izena

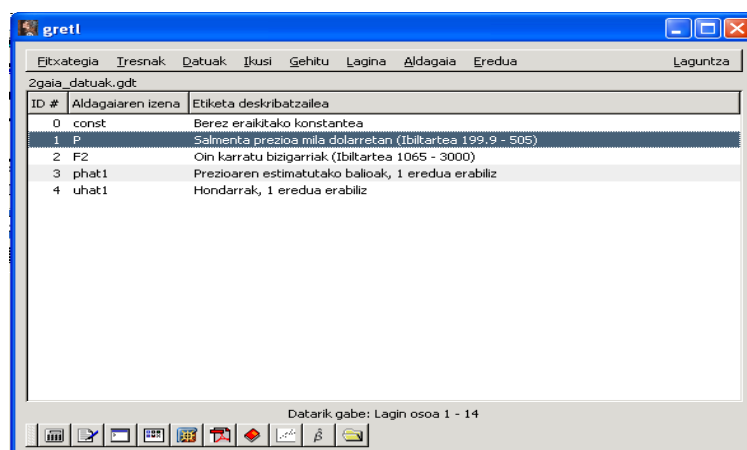
2.20 Irudia: Estimatutako balioak gordetzen



eta deskribapenean “1 ereduko estimatutako balioak” berez agertzen diren. Dena den izen eta deskribapen hauek aldatu ditzakegu. Adibide honetako aldagai azaldua prezioa (P) denez, izen bezala **phat1** jarriko dugu eta deskribapenean **Prezioaren estimatutako balioak**, 1 eredu erabiliz. Horrela “1 eredu” deskribapenean mantentzean, ikono bezala gordeta dauden estimatutako balioak zein eredutik lortu diren jakitea errazagoa da.

Aurreko pausuak hondarrentzat errepikatzerakoan, irtetzen den lehatilan berriz “uhat1” agertzen da hondarren izentzat eta deskribapen bezala “1 ereduaren hondarrak”. Hortaz, azken hau aldatuko dugu **Hondarrak**, 1 eredu erabiliz jarritz. Behin gordeta daudenean, P aldagai azaldua eta $F2$ aldagai azaltzailearekin batera agertuko dira.

2.21 Irudia: Gordetako datuen erakuspena



Ondoren bestelako informazioa aztertuko dugu, dauzkagun aldagai guztien estatistikoak hain zuzen. Horretarako, P , $F2$, phat1 eta uhat1 aldagaiak aukeratu (azpimarratu) ondoren, *Estatistikoen laburpena* klikatzean honako taula irtetzen da:

2.2 Taula: Aldagai guztien estatistiko nagusiak

Estatistikoen laburpena, 1 - 14 behaketak erabiliz

Aldagaia	BATEZ.	MEDIANA	MIN	MAX
P	317,493	291,500	199,900	505,000
F2	1910,93	1835,00	1065,00	3000,00
phat1	317,493	306,958	200,120	468,602
uhat1	0,000000	-1,1919	-53,601	91,8983
Aldagaia	D.T.	A.K.	ASIM	KURT. SOB
P	88,4982	0,278741	0,653457	-0,529833
F2	577,757	0,302344	0,485258	-0,672125
phat1	80,1640	0,252491	0,485258	-0,672125
uhat1	37,4921	6,15597e+15	1,02687	0,817927

Emaitzak aztertzerakoan:

1. Lehen ekuazio normalatik (2.5) ondorioztatzen den bezala eta lehen zutabean ikus daitekeenez, honarren, (“uhat1” bezala izendaturik) batezbesteko aritmetikoa (gorriz) zero da: $\hat{u} = 0$.
2. Aldagai azalduaren P eta estimatutako aldagai azalduaren $\hat{P}$ (“phat1”) batezbesteko aritmetikoak (laranjaz) berdinak dira: $\bar{P} = \hat{\bar{P}}$.
3. Estimututako aldagai azalduaren ($\hat{P}$) formaren ezaugarriak, asimetria (urdin ilunez) eta kurtosis (urdin argiz), aldagai azaltzailetik ($F2$) eratortzen ditu. Honen arrazoia honakoa da: estimatutako aldagai azalduaren balioak ($\hat{P}_i$), aldagai azaltzailearen balioei ($F2_i$) eskala aldaketa bat ($\times \hat{\beta}$) eta jatorri aldaketa bat ($+\hat{\alpha}$) eginez lortzen dira, eta estimatutako koefizienteen ($\hat{\beta}$ eta $\hat{\alpha}$) balioek ez dute formako estatistikoentzat inolako informaziorik eskaintzen.

Dauzkagun aldagai guztien koerlazio matrizea ateratzeko aldagai guztiak sailkatu ondoren, *Ikusi* $\rightarrow$ *Koerlazio matrizea* klikatu eta irtetzen den matrizea ondorengoa da:

2.3 Taula: Aldagai guztien koerlazio matrizea

Koerlazio Koefizienteak, 1 - 14 behaketak erabiliz
 %5eko esanguratasuna (alde biko) = 0,5324 n = 14 -rentzat

P	F2	phat1	uhat1	
1,0000	0,9058	0,9058	0,4236	P
	1,0000	1,0000	-0,0000	F2
		1,0000	-0,0000	phat1
			1,0000	uhat1

Emaitzak aztertuz:

1. Bigarren ekuazio normaletik (2.6) eratorritako emaitzak adierazten duen bezala, hondarren (“uhat”) eta aldagai azaltzailearen (F2) arteko koerlazioa (urdin ilunez) zero da ($r_{\hat{u},F2} = 0$) zeren hondarrak eta aldagai azaltzailea ortogonalak baitira eta hondarren batezbesteko aritmetikoa zero baita.
2. Aldagai azalduaren (P) koerlazioa estimatutako aldagai azalduarekiko ($\hat{P}$) eta aldagai azaltzailearekiko (F2) (orlegi ilunez) berdina da: $r_{P,\hat{P}} = r_{P,F2}$ (zeren $\hat{P} = \hat{\beta}F2$ baita).
3. Estimututako aldagai azalduaren ($\hat{P}$) eta aldagai azaltzailearen (F2) arteko koerlazioa (laranjez) bat da: $r_{\hat{P},F2} = 1$ (zeren $\hat{P} = \hat{\beta}F2$ baita).
4. Estimututako aldagai azalduaren ($\hat{P}$) eta hondarren ($\hat{u}$) arteko koerlazioa (gorriz) zero da ($r_{\hat{u},\hat{P}} = 0$) zeren $r_{\hat{u},F2} = 0$ baita.

Teoriara itzuliz, oinarritzko hipotesien menpean lortutako KTAko estimatzaileen propietateak zeintzuk diren jakitea komeni da. Alde batetik, KTAko estimatzaileak **perturbazioekiko linealak** dira, hau da, bai $\hat{\alpha}$ eta $\hat{\beta}$ ere, perturbazioen ($u_1, \dots, u_N$) konbinazio lineal bat bezala idatzi daitezke Bestaldetik KTAko estimatzaile hauek alboragabeak direnez,

$$E(\hat{\alpha}) = \alpha \quad E(\hat{\beta}) = \beta$$

batezbestekoaren inguruan (zentraturik) banatzen dira. Azkenik, estimazioaren zehaztasunari buruz, Gauss-Markoven teoremaren arabera, estimatzaile lineal (perturbazioekiko) eta alboragabe guztien artetik KTAko estimatzaileak **bariantza txikienekoak** dira. Eredu bakunean estimatzaileen populazio edo benetako bariantzen adierazpenak honakoak dira:

$$\text{bar}(\hat{\alpha}) = \sigma^2 \left(\frac{\sum_{i=1}^N X_i^2}{N \sum_{i=1}^N (X_i - \bar{X})^2} \right) \quad (2.12)$$

$$\text{bar}(\hat{\beta}) = \sigma^2 \left(\frac{1}{\sum_{i=1}^N (X_i - \bar{X})^2} \right). \quad (2.13)$$

Bariantza biak perturbazioen populazioko bariantzaren ($\text{bar}(u_i) = \sigma^2$) menpekoak dira eta baita aldagai azaltzailearen (X) dispertsioaren menpekoak (S_X^2). Zenbat eta dispertsio

$(\sum_{i=1}^N (X_i - \bar{X})^2)$ handiagoa izan, aldagai azaltzaileak duen informazioa gero eta handiagoa izaten da. Ondorioz bariantza txikiagoak lortzen dira. Estimatzailen arteko kobariantzari dagokionez,

$$kob(\hat{\alpha}, \hat{\beta}) = \sigma^2 \left(\frac{-\bar{X}}{\sum_{i=1}^N (X_i - \bar{X})^2} \right) \quad (2.14)$$

ez da zero izango baldin eta aldagai azaltzailearen batezbesteko aritmetikoa desberdin zero bada eta gainera bere zeinua batezbestekoaren aurkakoa izango da.

Populazio momentu hauen estimazioak lortzeko, lehendabiziko pausua perturbazioaren bariantza estimatzea da. Erabiliko dugun estimatzaile alboragabea honako da

$$\widehat{bar}(u_i) = \hat{\sigma}^2 = \frac{1}{N-K} \sum_{i=1}^N \hat{u}_i^2$$

non $\sum_{i=1}^N \hat{u}_i^2$ **hondar karratuen batura** den eta $N-K$ dauzkagun askatasun graduak. Honela izanik, KTAko estimatzaileen bariantzen eta kobariantzaren estimatzaile alboragabeak hurrengoak dira:

$$\widehat{bar}(\hat{\alpha}) = \frac{\sum_i \hat{u}_i^2}{N-K} \left(\frac{\sum_{i=1}^N X_i^2}{N \sum_{i=1}^N (X_i - \bar{X})^2} \right) \quad (2.15)$$

$$\widehat{bar}(\hat{\beta}) = \frac{\sum_i \hat{u}_i^2}{N-K} \left(\frac{1}{\sum_{i=1}^N (X_i - \bar{X})^2} \right) \quad (2.16)$$

$$\widehat{kob}(\hat{\alpha}, \hat{\beta}) = \frac{\sum_i \hat{u}_i^2}{N-K} \left(\frac{-\bar{X}}{\sum_{i=1}^N (X_i - \bar{X})^2} \right) \quad (2.17)$$

Adibideko emaitzetara itzuliz:

Eredua 1: KTA estimazioak 14 behaketak erabiliz 1-14
Aldagai azaldua: P

Aldagaia	Koefizientea	Desb. Tipikoa	t-estatistikoa	p-balioa
const	52,3509	37,2855	1,4041	0,1857
F2	0,138750	0,0187329	7,4068	0,0000

Aldagai azalduaren batezbestekoa	317,493
Aldagai azalduaren Desb. Tip.	88,4982
Hondar Karratuen Batura	18273,6
Hondarren desbideratze tipikoa ($\hat{\sigma}$)	39,0230
R^2	0,820522
Zuzendutako $\bar{R}^2$	0,805565
Askatasun graduak	12
Log-egiantza	-70,084
Akaike Informazio Irizpidea	144,168
Schwarz Bayesian Irizpidea	145,447
Hannan-Quinn Irizpidea	144,050

2.4 Taula: KTAko estimatzailearen bariantza eta kobariantza matrizea

Erregresio koefizienteen kobariantza matrizea

const	F2	
1390,21	-0,670583	const
	3,50920e - 04	F2

“Desb. Tipikoa” deituriko zutabean, koefizienteen estimatzaileen estimatutako desbideratzeak (aurreko (2.15) eta (2.16) bariantza estimatuen erro positiboa) agertzen dira. Beraien arteko kobariantza estimatua zein den jakiteko bariantza eta kobariantza matrize estimatua kalkulatu behar da. Horretarako: *Analisia* $\rightarrow$ *Koefizienteen kobariantza matrizea* aukeratu eta ateratzen den emaitza honakoa da:

estimatzen ari garen matrizea ondorengoa dela jakinik

$$\widehat{Bar} \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \begin{pmatrix} \widehat{bar}(\hat{\alpha}) & \widehat{kob}(\hat{\alpha}, \hat{\beta}) \\ \widehat{kob}(\hat{\alpha}, \hat{\beta}) & \widehat{bar}(\hat{\beta}) \end{pmatrix} = \begin{pmatrix} 1390,21 & -0,670583 \\ 0,000350920 & \end{pmatrix}$$

diagonal nagusian bariantza estimatuak agertzen dira eta diagonal nagusitik kanpo dagoen elementua estimatutako kobariantzari dagokio. Estimatutako momentu hauek kalkulatzera-koan erabili den perturbazioaren bariantza estimatua $\hat{\sigma}^2 = \frac{\sum_i^N \hat{u}_i^2}{N-K}$ ez da zuzenean ematen, bere erro karratua baizik, hau da, perturbazioaren desbideratze estimatua. Honela, bariantza desbideratzea karratura eginez kalkulatu daiteke $\hat{\sigma}^2 = 39,0230^2 = 1522,8$ edo bestela eman-dako formula aplikatuz $\hat{\sigma}^2 = 18273,6/12 = 1522,8$ zeren hondar karratuen batura (18273,6) eta askatasun graduak ($N - K = 14 - 2 = 12$) emaitzetan agertzen baitira.

Zoritxarrez, nahiz eta estimaturiko bariantzak estimazioaren zehaztasuna neurtu, ateratako emaitzak onak diren edo ez bereizteko gai ez gara. Arrazoiak: bariantzak nahi ditugun bezain txikiak egitea posible da aldagai bien (dependentea eta independentea) eskala aldatuz. Ondorioz, estimatutako balioak eta benetakoak hurbil dauden edo ez jakiteko, ontasun neurriren bat behar dugu doikuntzaren egokitasuna aztertzeke.

2.4 Doikuntzaren egokitasuna

Eredu bakunaren doikuntzaren ontasuna neurtzeko erabiliko dugun neurria mugatze koefizientea izango da (R-karratua). Koefiziente honen adierazpena honakoa da:

$$R^2 = r_{X,Y}^2 = 1 - \frac{\sum (Y_i - \hat{Y}_i)^2}{\sum (Y_i - \bar{Y})^2} = 1 - \frac{\sum \hat{u}_i^2}{\sum (Y_i - \bar{Y})^2}. \quad (2.18)$$

Koefiziente honek X aldagai azaltzaile gabeko eredu batetik

$$Y_i = \alpha + u_i$$

X aldagaiak barneratzen duen eredu batera pasatzean

$$Y_i = \alpha + \beta X_i + u_i$$

lortzen den irabazia neurtzen du. Hortaz R^2 , aldagai azaltzailearen (X) bariantzak, aldagai azalduaren (Y) bariantzaren zenbateko portzentaia azaltzen duen neurtzen du era lineal batean. Adibideko mugatze koefizientea analizatzean $R^2 = 0,820522$, etxebizitzaren azalaren ($F2$) bariantzarekin salmenta prezioaren bariantzaren %82,0522a azaltzen dela era lineal batean esango genuke.

2.5 Esanguratasun analisia eta konfidantza tarteak

Ereduko X aldagai azaltzailea, aldagai azaldua azaltzeko nabaria den baieztatzeko kontraste bat egin dezakegu. Horretarako, hipotesi hutsean aldagai azaltzaileari laguntzen dion koefizientea berdin zero jarriko genuke eta aurkakoa desberdin zero dela:

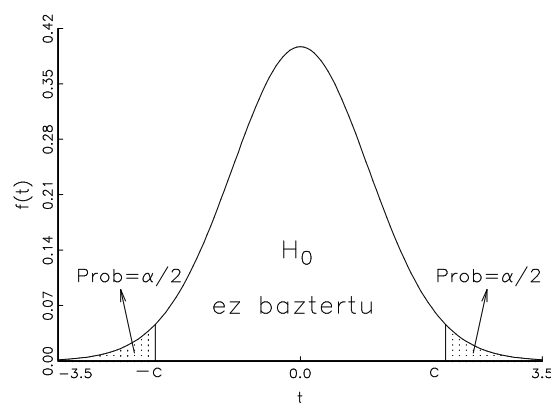
$$\begin{cases} H_0 : \beta = 0 & (X \text{ aldagai azaltzailea ez da nabaria}) \\ H_a : \beta \neq 0 & (X \text{ aldagai azaltzailea nabaria da}) \end{cases}$$

Kontraste honetarako estatistikoa honakoa litzateke:

$$t_{est} = \frac{\hat{\beta} - \beta}{\widehat{des(\hat{\beta})}} \stackrel{H_0}{\sim} t_{(N-K)}$$

eta erabaki araua hurrengoa izango da:

2.22 Irudia: Erabaki araua



kalkulatu behar den estatistikoaren balioa $[-c = -t_{(N-K)\alpha/2}, c = t_{(N-K)\alpha/2}]$ balioen artean badago, konfidantza tartean aurkitzen denez, ez dugu hipotesi hutsa baztertuko erabilitako esangura-mailarentzat eta konklusio bezala, X aldagai azaltzailea Y aldagai azaldua azaltzeko aldagai nabari bat **ez** dela esango dugu. Bestalde, kalkulaturako estatistikoaren balioa $-c = -t_{(N-K)\alpha/2}$ balioa baino txikiagoa bada, edota $c = t_{(N-K)\alpha/2}$ balioa baino handiagoa, orduan eskualde kritikoan erortzen denez, hipotesi hutsa baztertu egingo dugu erabilitako esangura-mailarekin eta konklusio bezala, X aldagai azaltzailea, aldagai azaldua azaltzeko aldagai nabaria dela baieztatuko dugu.

Adibidera itzuliz, ikus dezagun ea etxebizitzaren azalera aldagai nabaria den etxebizitzaren salmenta prezioa zehazteko:

$$\begin{cases} H_0 : \beta = 0 \\ H_a : \beta \neq 0 \end{cases} \quad t_{est} = \frac{\hat{\beta} - \beta}{\widehat{des}(\hat{\beta})} \stackrel{H_0}{\sim} t_{(N-K)}$$

Kalkulatu behar den estatistikoaren balioa Gretlek emandako estimazio emaitzetan agertzen denez, egin behar den gauza bakarra tauletako balioarekin ($t_{(14-2)0,05/2} = 2,179$) konparatzea da. Kasu honetan $7,4068 > 2,179 = t_{(14-2)0,05/2}$ ematen denez, estatistikoaren balioa eskualde kritikoan erortzen da eta ondorioz, hipotesi hutsa baztertu egingo dugu %5eko esangura-mailarekin. Konklusio bezala, etxebizitzaren azalera etxebizitzaren salmenta prezioa azaltzeko aldagai azaltzaile nabari bat dela esango genuke.

Behin $\beta \neq 0$ dela eta beraz $F2$ aldagai nabari bat dela jakinik, koefizienteak har dezakeen balioa jakitea interesagarria izan daiteke. Adibidez:

- Etxebizitzaren azalera oin karratu batean handitzerakoan, posible litzateke etxebizitzaren batezbesteko salmenta prezioa 100 dolarretan gehitzea? Erantzuteko ondorengo kontrastea burutuko dugu:

$$\begin{cases} H_0 : \beta = 0, 1 \\ H_a : \beta \neq 0, 1 \end{cases} \quad t_{est} = \frac{\hat{\beta} - \beta}{\widehat{des}(\hat{\beta})} \stackrel{H_0}{\sim} t_{(N-K)}$$

Kontuan izan behar da lehen erabilitako estatistikoaren balioak (estimazio taulan agertzen dena) banakako esanguratasuna kontrastatzeko soilik balio duela. Beraz, beste kontraste hau egin nahi izanez gero, estatistikoaren balioa berriro kalkulatu behar da $t_{est} = \frac{0,138750-0,1}{0,0187329} = 2,068$, eta tauletako balioa izanik ($t_{(14-2)0,05/2} = 2,179$), lortzen den emaitza $-t_{(14-2)0,05/2} = -2,179 < 2,068 < 2,179 = t_{(14-2)0,05/2}$ da. Estatistikoaren balioa konfidantza tartean erortzen denez hipotesi hutsa ez da baztertzeko %5eko esangura-mailarekin eta ondorioz etxebizitza baten azalera oin karratu batean handitzerakoan, bere batezbesteko prezioan 100 dolarreko gehikuntza dakarrela esan dezakegu.

- Eta azaleraren oin karratu bateko gehikuntza berdinarekin, posiblea litzateke etxebizitzaren batezbesteko salmenta prezioa 150 dolarretan gehitzea? Ikus dezagun:

$$\begin{cases} H_0 : \beta = 0, 15 \\ H_a : \beta \neq 0, 15 \end{cases} \quad t_{est} = \frac{\hat{\beta} - \beta}{\widehat{des}(\hat{\beta})} \stackrel{H_0}{\sim} t_{(N-K)}.$$

Estatistikoaren balioa kasu honetan $t_{est} = \frac{0,138750-0,15}{0,0187329} = -0,6005$ konfidantza tartean erortzen denez ($-2,179 < -0,6005 < 2,179$) hipotesi hutsa ez da baztertzeko %5eko esangura-mailarekin. Orduan, etxebizitza baten azalera oin karratu batean handitzerakoan, bere batezbesteko prezioan 150 dolarreko gehikuntza dakarrela baieztatu dezakegu.

Zer gertatzen ari da? Puntuzko estimazioak β koefizientearen balio posible bat ematen digu, baina errealitatean balio posibleen multzo bat daukagu. Balio posible hauek lortzeko koefizientearen konfidantza tartea atera behar dugu. Goiko kontrasteak aztertuz, erraz ikus

daiteke, estatistikoaren balioa $(-2, 179; 2, 179)$ tartearen barnean erortzea posible egiten duten balio guztiak, koefizientearen balio posible guztiak izango direla.

Teorikoki koefiziente baten konfidantza tartea ateratzeko, Student- t banaketaren taulako $(-t_{(N-K)\alpha/2} ; t_{(N-K)\alpha/2})$ bi balioen artean egon ahal diren estatistikoaren balio posible guztiak kalkulatu lortzen da:

$$Pr \left[-t_{(N-K)\alpha/2} < \frac{\hat{\beta} - \beta}{\widehat{des}(\hat{\beta})} < t_{(N-K)\alpha/2} \right] = 1 - \alpha \quad (2.19)$$

askatuz:

$$Pr \left[\hat{\beta} - t_{(N-K)\alpha/2} \widehat{des}(\hat{\beta}) < \beta < \hat{\beta} + t_{(N-K)\alpha/2} \widehat{des}(\hat{\beta}) \right] = 1 - \alpha \quad (2.20)$$

$\%(1 - \alpha)$ -ri dagokion konfidantza tartea ateratzen da. Tarte hau puntuzko estimazioan zentratu dago eta zentru honetatik urruntzen den kantitatea jakiteko $t_{(N-K)\alpha/2}$ aldiz estimatzailearen desbideratze tipiko estimatua $(\widehat{des}(\hat{\beta}))$ kalkulatu behar dugu. Hemendik aurrera, konfidantza tarte bat idazteko erabiliko dugun notazioa hurrengoa izango da:

$$KT(\beta)_{1-\alpha} = \left[\hat{\beta} \pm t_{(T-K)\alpha/2} \widehat{des}(\hat{\beta}) \right].$$

Bestalde, α koefizienteari dagokion konfidantza tartea era berdinean lortzen da:

$$KT(\alpha)_{1-\alpha} = \left[\hat{\alpha} \pm t_{(T-K)\alpha/2} \widehat{des}(\hat{\alpha}) \right].$$

Adibidearekin bukatzeko, koefizienteen konfidantza tarteak aterako ditugu, horretarako, *Ana-lisia* $\rightarrow$ *Koefizienteen konfidantza tarteak* aukeratu eta ateratzen dena honakoa da:

2.5 Taula: Tartezko estimazioa

$t(12, .025) = 2,179$

ALDAGAIA	KOEFIZIENTEA	%95eko KONFIDANTZA TARTEA
const	52,3509	$(-28,8872, \setminus 133,589)$
F2	0,138750	$(0,0979349, \setminus 0,179566)$

Lortzen den emaitzan, bigarren zutabeko balioak puntuzko estimazioak dira, $\hat{\alpha} = 52,3509$ eta $\hat{\beta} = 0,138750$ eta hirugarren zutabean, %95eko konfidantza tartearen behe eta goi kotak agertzen dira, hau da:

$$KT(\alpha)_{0,95} = [-28,887 ; 133,587]$$

$$KT(\beta)_{0,95} = [0,0979349 ; 0,179566]$$

Beraz, etxebizitzaren azaleran oin karratu bateko gehikuntza batek, bere batezbesteko salmenta prezioa 97,9349 eta 179,566 dolar bitarteko gehikuntza ekarriko duela esan dezakegu %95eko konfidantzarekin.

2.6 Laburpena. Emaizten aurkezpena

Eredu baten estimazio emaitzak laburbildu daitezke, lagineko erregresio funtzioa eta emaitzak ebaluatzeko baliogarriak diren estatistiko multzo bat eskeiniz. Ohituraz, emaitzak bi erataraz auztezen dira:

Lehen era:

$$\begin{array}{c} \hat{P} \\ (\widehat{des}) \end{array} = 52,3509 + 0,138750 F^2$$

$$\begin{array}{ccc} (37,285) & (0,018733) & \\ N = 14 & R^2 = 0,82 & \hat{\sigma} = 39,023 \end{array}$$

non koefiziente estimatu bakoitzaren azpian desbideratze tipiko estimatua agertzen den.

Bigarren era:

$$\begin{array}{c} \hat{P} \\ (t_{est}) \end{array} = 52,3509 + 0,138750 F^2$$

$$\begin{array}{ccc} (1,404) & (7,407) & \\ Askatasun graduak = 12 & R^2 = 0,82 & \hat{\sigma} = 39,023 \end{array}$$

non koefiziente estimatu bakoitzaren azpian t-estatistikoa agertzen den.

Bibliografia

Ramanathan, R. (2002), *Introductory Econometrics with Applications*, 5. ed., South-Western, Ohio.

3 Gaia

Erregresio linealeko eredu orokorra

Aurkibidea

3.1	Sarrera	52
3.2	Karratu Txikiarren Arruntetako estimazioa	53
3.2.1	Koefizienteak	57
3.2.2	Desbideratze tipikoak eta konfidantza tarteak	59
3.2.3	Banakako eta baterako esanguratasunak	61
	Banakako esanguratasuna	61
	Baterako esanguratasuna	62
3.3	Doikuntzaren ontasuna eta ereduaren sailkapena	63

3.1 Sarrera

Gai honetan erregresio linealeko eredu orokorra aztertuko dugu. Ereduan konstante bat izateaz gain beste aldagai azaltzaile bat baino gehiago izango dugu, hau da, $K > 2$ izango da:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \dots + \beta_K X_{Ki} + u_i \quad i = 1, 2, \dots, N$$

Aurretik ikusitako oinarritzko hipotesi guztiak betetzen direla suposatuko dugu. Hala ere, erregresio orokorra izaterakoan hipotesi batzuk orokortu egin behar dira:

1. Aldagai azaltzaileak $X_{2t}, \dots, X_{Kt}$ ez dira estokastikoak.
2. Aldagai azaltzaileak linealki independenteak dira.
3. Koefiziente guztiak ($\beta_j, \quad j = 1, \dots, K$) estimatzea posible izateko, behaketa kopuru nahiko izan behar dugu: $K < T$

Eredu orokorraren koefizienteen interpretazioa egiterakoan, termino konstantearena lehen bezala izango da, baina maldak interpretatzerakoan kontuan izan behar da aldagai azaltzaile gehiago daudela. Honela, $X_{ji} \quad j = 1, \dots, K$ unitate batean handitzean aldagai azalduaren (Y_i) batezbesteko gehikuntza β_j unitatekoa da, **beste aldagai azaltzaile guztiak konstante mantenduz**.

Gai honetan erabiliko dugun adibidearen datuak Ramanathaneko (2002) *data4-1 Prices of single-family homes* fitxategian daude. Izatez aurreko gaian erabilitako datu berberak dira baina aldagai azaltzaile gehiago erantsiz. Halaber, orain etxebizitzaren prezioa analizatu nahi izango dugu bere tamaina, gela eta komun kopuruaren menpean. Hemendik aurrera erabiliko ditugun aldagaien izenak, fitxategi honek ematen dituenak izango dira. Hau da:

A Eredua

$$PRICE_i = \beta_1 + \beta_2 SQFT_i + \beta_3 BEDRMS_i + \beta_4 BATHS_i + u_i \quad i = 1, \dots, 14 \quad (3.1)$$

PRICE: Etxebizitzaren salmenta prezioa mila dolarretan (Ibiltartea 199,9 - 505)
 SQFT: Etxebizitzaren azalera oin karratutan (Ibiltartea 1065 - 3000)
 BEDRMS: Gela kopurua (Ibiltartea 3 - 4)
 BATHS: Komun kopurua (Ibiltartea 1,75 - 3)

Eredu hau matrizialki idatzi daiteke:

$$\underset{(14 \times 1)}{Y} = \underset{(14 \times 4)}{X} \underset{(4 \times 1)}{\beta} + \underset{(14 \times 1)}{u} \quad (3.2)$$

non matrize bakoitzaren itxura honakoa den:

$$\underbrace{\begin{bmatrix} PRICE_1 \\ PRICE_2 \\ \vdots \\ PRICE_i \\ \vdots \\ PRICE_{14} \end{bmatrix}}_Y = \underbrace{\begin{bmatrix} 1 & SQFT_1 & BEDRMS_1 & BATHS_1 \\ 1 & SQFT_2 & BEDRMS_2 & BATHS_2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & SQFT_i & BEDRMS_i & BATHS_i \\ \vdots & \vdots & \vdots & \vdots \\ 1 & SQFT_{14} & BEDRMS_{14} & BATHS_{14} \end{bmatrix}}_X \underbrace{\begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix}}_{\beta} + \underbrace{\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_i \\ \vdots \\ u_{14} \end{bmatrix}}_u$$

Koefizienteen interpretazioa:

- β_1 : etxebizitzaren batezbesteko salmenta prezioa, azalera zero denean eta gelarik eta komunik ez duenean.
- β_2 : etxebizitzak oin karratu bat gehiago izaterakoan, gela eta komun kopurua mantenduz, salmenta prezioaren batezbesteko gehikuntza β_2 mila dolarrekoa da.
- β_3 : etxebizitzak gela bat gehiago izaterakoan, azalera eta komun kopurua mantenduz, salmenta prezioaren batezbesteko gehikuntza β_3 mila dolarrekoa da.
- β_4 : etxebizitzak komun bat gehiago izaterakoan, azalera eta gela kopurua mantenduz, salmenta prezioaren batezbesteko gehikuntza β_4 mila dolarrekoa da.

Eredu orokorra analizatzean, aldagai azaltzaile bakoitzak duen “**efektu gehigarria**”, beste aldagai azaltzaile guztien efektuak kontrolatuz, aztertzea posible da.

3.2 Karratu Txikiaren Arruntetako estimazioa

Karratu Txikiaren Arrunten (KTA) bitartez eredua estimatzerakoan minimizatu behar den helburu funtzioa Hondar Karratuen Batura (HKB) izaten jarraitzen du. Adibidean lau aldagai azaltzaile ditugunez ($K = 4$), demagun aldagai azaldua $Y_i \equiv PRICE_i$ dela eta aldagai azaltzaileak, $X_{2i} \equiv SQFT_i$, $X_{3i} \equiv BEDRMS_i$ eta $X_{4i} \equiv BATHS_i$ izendatzen ditugula. Horrela, Karratu Txikiaren Arruntetako estimatzaileak lortzeko HKB minimizatzen dugu jarraian adierazten den moduan:

$$\min_{\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4} \sum_{i=1}^{N=14} \hat{u}_i^2 \equiv \min_{\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4} \sum_{i=1}^N (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i} - \hat{\beta}_4 X_{4i})^2$$

Hemendik, lehen ordenako baldintzak deribatuz ($\frac{\partial HKB}{\partial \hat{\beta}_j} = 0 \quad \forall j$ eginez), lau ekuazio normal lortzen dira:

$$\begin{aligned} \sum Y_i &= N\hat{\beta}_1 + \hat{\beta}_2 \sum X_{2i} + \hat{\beta}_3 \sum X_{3i} + \hat{\beta}_4 \sum X_{4i} \\ \sum Y_i X_{2i} &= \hat{\beta}_1 \sum X_{2i} + \hat{\beta}_2 \sum X_{2i}^2 + \hat{\beta}_3 \sum X_{3i} X_{2i} + \hat{\beta}_4 \sum X_{4i} X_{2i} \\ \sum Y_i X_{3i} &= \hat{\beta}_1 \sum X_{3i} + \hat{\beta}_2 \sum X_{2i} X_{3i} + \hat{\beta}_3 \sum X_{3i}^2 + \hat{\beta}_4 \sum X_{4i} X_{3i} \\ \sum Y_i X_{4i} &= \hat{\beta}_1 \sum X_{4i} + \hat{\beta}_2 \sum X_{2i} X_{4i} + \hat{\beta}_3 \sum X_{3i} X_{4i} + \hat{\beta}_4 \sum X_{4i}^2 \end{aligned}$$

Eredu bakunean bezala, lehen ekuazio normalak, termino konstantetik eratortzen dena alegia, hondarren batura zero dela adierazten du. Besteek aldiz, hondarrak eta aldagai azaltzaileak ortogonalak direla adierazten dute.

Ekuazio normal hauek matrizialki jartzen baditugu,

$$X'Y = (X'X)\hat{\beta}$$

eta hemendik $\hat{\beta}$ askatuz, Karratu Txikiaren Arruntetako estimatzailea lortzen dugu:

$$\hat{\beta}_{KTA} = (X'X)^{-1}X'Y.$$

Gure adibideko eredua

$$\mathbf{A} \text{ Eredua} \quad PRICE_i = \beta_1 + \beta_2 SQFT_i + \beta_3 BEDRMS_i + \beta_4 BATHS_i + u_i$$

KTA bitartez estimatzean lortzen diren emaitzak honakoak dira:

A Eredua KTA estimazioak 14 behaketak erabiliz 1–14

Aldagai azaldua: price

Aldagaia	Koefizientea	Desb. Tipikoa	<i>t</i> -estatistikoa	p-balioa
const	129,062	88,3033	1,4616	0,1746
sqft	0,154800	0,0319404	4,8465	0,0007
bedrms	−21,587	27,0293	−0,7987	0,4430
baths	−12,192	43,2500	−0,2819	0,7838

Aldagai azalduaren batezbestekoa	317,493
Aldagai azalduaren Desb. Tip.	88,4982
Hondar Karratuen Batura	16700,1
Hondarren desbideratze tipikoa ($\hat{\sigma}$)	40,8657
R^2	0,835976
Zuzendutako $\bar{R}^2$	0,786769
$F(3, 10)$	16,9889
p-balioa $F()$	0,000298587
Log-egiantza	−69,453
Akaike Informazio Irizpidea	146,908
Schwarz Bayesian Irizpidea	149,464
Hannan–Quinn Irizpidea	146,671

Ondorengo azpiataletan Gretl programarekin lorturiko emaitza hauek interpretatuko ditugu. Baina honekin hasi baino lehen, jakin ezazue eredu ekonometrikoak estimatzeko software asko daudela eta ohituraz (bai liburuetan, bai artikuluetan) eredu orokorreko emaitzen aurkezpena honakoa izaten dela:

$$\widehat{PRICE}_i = 129,062 + 0,154800 SQFT_i - 21,5875 BEDRMS_i - 12,1928 BATHS_i$$

(1,462) (4,847) (−0,799) (−0,282)

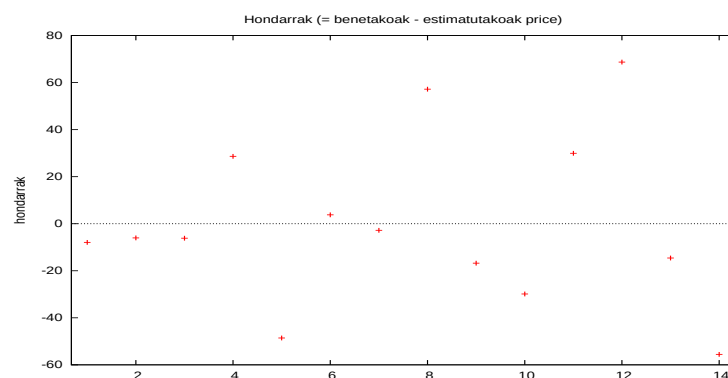
$$T = 14 \quad R^2 = 0,8359 \quad \bar{R}^2 = 0,7868 \quad F(3, 10) = 16,989$$

(parentesi artean *t*-estatistikoak)

Emaitza hauek aurkezterakoan, parentesi artean desbideratze tipikoak (bariantza estimatuaren erro karratu positiboa), t-estatistikoak (banakako esanguratasun kontrastea burutzeko estatistikoaren balioa) edota p-balioak (Gretl emaitzetako azken zutabean agertzen dira eta aurrerago ikusiko dugu zer adierazten duten) jarri daitezke, askotan hirurak agertzen direlarik.

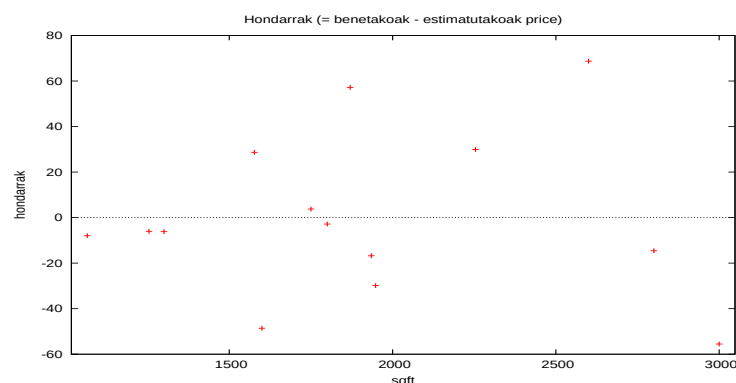
Estimazio emaitzen grafiko batzuk. Estimazio emaitzen leihatilan grafiko desberdinak ateratzeko aukera dago. Hondarren kasuan, hauek laginean zehar edo aldagai azaltzaile batekiko irudikatzeko aukera dago. Adibidez, hondarrak laginean zehar irudikatuz lortzen den grafikoa honakoa da:

3.1 Irudia: Hondarrak laginean zehar



hemen hondarrak zero inguruan mugitzen direla ikusi dezakegu. Egitez, horrela izan behar da zeren hondarren batezbesteko aritmetikoa zero baita. Behaketen dispersioari dagokionez, azken behaketen dispersioa hasierakoena baino handiagoa dela ikusten da. Hondarrak *sqft* aldagai azaltzailearekiko irudikatzerakoan lortzen den grafikoa ondorengoa da:

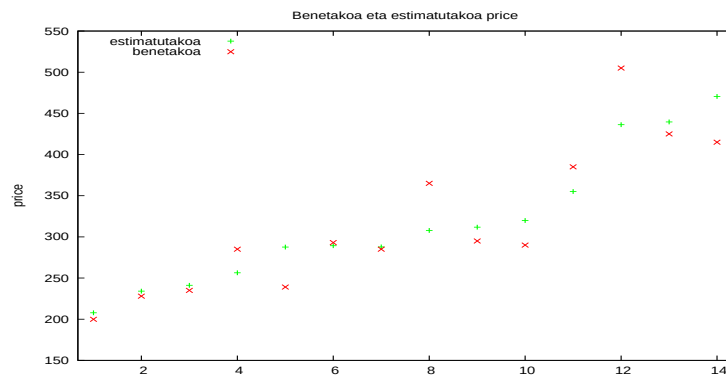
3.2 Irudia: Hondarrak *sqft* aldagai azaltzailearekiko



eta kasu honetan, zero batezbestekoaren inguruan mugitzen direlaz gain, *sqft* aldagaiaren balioa handitzerakoan hondarren dispersioa ere handitzen doala nabaritzen da. Hortaz, ereduaren zehazpena zuzena delaren suposiziopean, badirudi perturbazioaren bariantza konstantea delaren oinarritzko hipotesia ez dela betetzen eta perturbazioaren bariantza *sqft* aldagaiaren menpekkoa dela.

Ondorengo grafikoan benetako eta estimatutako aldagai azalduaren balioak aurkezten dira laginean zehar:

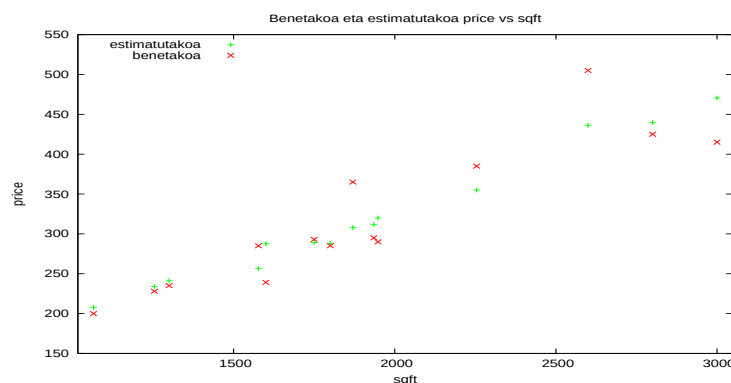
3.3 Irudia: Benetako eta estimatutako prezioak laginean zehar



honela ereduaren doikuntza ikusteko aukera izanik. Lagineko azken behaketen doikuntza txarragoa dela nabaritzen da.

Azkenik, benetako eta estimatutako aldagai azalduaren balioak aurkezten dira *sqft* aldagai azaltzailearen aurka:

3.4 Irudia: Benetako eta estimatutako prezioak vs sqft



Ereduaren doikuntza tamaina txikiko etxebizitzentzat hobeagoa dela ikusi daiteke eta 2000 oin karratu baino gehiago dituzten etxebizitzentzat berriz, txarragoa dela.

3.2.1 Koefizienteak

Emaitza honetako bigarren zutabeko koefizienteen estimazioak hartuz, **A ereduari** dagokion lagineko erregresio funtzioa honakoa da:

$$\widehat{PRICE}_i = 129,062 + 0,1548 SQFT_i - 21,588 BEDRMS_i - 12,193 BATHS_i$$

Zeinuak aztertzerakoan, hasiera batean SQFT aldagaiaren koefizientearen zeinu positiboarekin ados gaude, zenbat eta etxebizitza handiagoa izan garestia izatea normala baita. BEDRMS eta BATHS aldagaien koefizienteen zeinu negatiboekin aldiz, ados ez gaudela pentsa dezakegu, gela edo komun gehiago izateak salmenta prezioa igo beharko baitluke. Hala ere, erregresio orokorrean kontuz ibili behar gara, koefiziente hauek “efektu gehigarria” neur-tzen baitute beste aldagaiak **konstante mantenduz**. Ondorioz, etxebizitzaren azalera eta komun kopurua konstante mantenduz, gela bat gehiago izateak gela guztiak txikiagoak izan behar direla inplikatzeko du. Azkenean, etxebizitzaren azalera zati gehiagotan banatu behar denez, etxebizitzaren kalitatea jaitsi egiten da, bere batezbesteko salmenta prezioa jaitziz. Adibidean, gela bat gehiago izateak, azalera eta komun kopuru berdinarekin, etxebizitzaren batezbesteko salmenta prezioa 21,588 mila dolarretan jaisten du. Bestalde, komun bat gehiago izateak, azalera eta gela kopurua konstante mantenduz, etxebizitzaren batezbesteko salmenta prezioa 12,193 mila dolarretan gutxierazten du.

Zer gertatuko litzateke BEDRMS-ren koefizientearen zeinuarekin SQFT eta BATHS aldagaiak ereduaren ez baditugu barneratzen? Zeinu negatiboa mantenduko litzateke? Erantzuna ezezkoa da. Gela kopurua bakarrik sartzen badugu (konstanteaz gain) bere koefizientearen zeinua positiboa izango da, orokorrean gela bat gehiago izateak batezbesteko salmenta prezioa gehitu egingo du, azalera aldagaia ez denez barneratu, azalera konstante mantentzen dela ez baita suposatuko behar. Hau da, zehaztutako ereduaren ondorengoa izanik, dagokion KTA estimazioaren emaitzetan ikusi daiteke zeinu aldaketa.

B Eredua $PRICE_i = \lambda_1 + \lambda_2 BEDRMS_i + u_i \quad i = 1, \dots, N$

B Eredua: KTA estimazioak 14 behaketak erabiliz 1–14

Aldagai azaldua: price

Aldagaia	Koefizientea	Desb. Tipikoa	<i>t</i> -estatistikoa	p-balioa
const	112,853	179,123	0,6300	0,5405
bedrms	56,1756	48,7511	1,1523	0,2716
Hondar Karratuen Batura			91671,7	
Hondarren desbideratze tipikoa ($\hat{\sigma}$)			87,4031	
R^2			0,0996248	
Zuzendutako $\bar{R}^2$			0,0245935	
Akaike Informazio Irizpidea			166,747	
Schwarz Bayesian Irizpidea			168,025	

Berdina gertatzen da BATH aldagaia bakarrik barneratuko bagenu, hau da:

$$\mathbf{C} \text{ Eredua} \quad PRICE_i = \theta_1 + \theta_2 BATHS_i + u_i \quad i = 1, \dots, N$$

C Eredua: KTA estimazioak 14 behaketak erabiliz 1–14

Aldagai azaldua: price

Aldagaia	Koefizientea	Desb. Tipikoa	t -estatistikoa	p-balioa
const	4,50655	101,869	0,0442	0,9654
baths	132,782	42,5153	3,1232	0,0088

Hondar Karratuen Batura	56163,1
Hondarren desbideratze tipikoa ($\hat{\sigma}$)	68,4124
R^2	0,448381
Zuzendutako $\bar{R}^2$	0,402413
Akaike Informazio Irizpidea	159,888
Schwarz Bayesian Irizpidea	161,166

Eta zer gertatuko litzateke ereduan BEDRMS eta BATHS bakarrik barneratzen baditugu?

$$\mathbf{D} \text{ Eredua} \quad PRICE_i = \delta_1 + \delta_2 BEDRMS_i + \delta_3 BATHS_i + u_i \quad i = 1, \dots, N$$

D Eredua: KTA estimazioak 14 behaketak erabiliz 1–14

Aldagai azaldua: price

Aldagaia	Koefizientea	Desb. Tipikoa	t -estatistikoa	p-balioa
const	27,2633	149,652	0,1822	0,8588
bedrms	-10,137	46,9811	-0,2158	0,8331
baths	138,795	52,3450	2,6515	0,0225

Hondar Karratuen Batura	55926,4
Hondarren desbideratze tipikoa ($\hat{\sigma}$)	71,3037
R^2	0,450706
Zuzendutako $\bar{R}^2$	0,350834
$F(2, 11)$	4,51285
Akaike Informazio Irizpidea	161,829
Schwarz Bayesian Irizpidea	163,746

Kasu honetan, BEDRMS aldagaiaren koefizientearen estimazioaren zeinua negatiboa da eta BATHS aldagaiarena berriz, positiboa. Gela kopurua handitzen bada, komun kopurua mantenduz, etxebizitzaren prezio estimatua murrizten dela pentsatzea logikoa badirudi ere, komun kopurua handitzean, gela kopurua mantenduz, prezio estimatua handitu egiten dela harritzekoa da. Zergatik ematen dira zeinu aldaketa hauek eredu bat edo bestea estimatzera-koan? Zein da arazoa? SQFT aldagai azaltzaile nabaria kanpoan uztean, ondo egiten ari al gara? Fidagarriak izango ote dira ateratako emaitzak? Gai honetan eta ondorengo gaietan, kontrasteetan eta beste zenbait irizpideetan oinarrituz, galdera hauei erantzuten saiatuko gara.

3.1 Taula: KTAko estimatzaileen bariantza eta kobariantza matrizea

Erregresio koefizienteen kobariantza matrizea

const	sqft	bedrms	baths	
7797,47	0,670891	-1677,13	-1209,37	const
	0,00102019	-0,0754606	-0,995066	sqft
		730,585	-356,4	bedrms
			1870,56	baths

3.2.2 Desbideratze tipikoak eta konfidantza tarteak

Aurreko gaietan aipatu den bezala, puntuzko estimazioak lagin konkretu batekin ateratako balioak direnez, estimazio errore bat daukagu. Badakigu lagin desberdinak hartzen badiugu, koefizienteen puntuzko estimazio desberdinak aterako ditugula. Horregatik, batezbestekoz ondo egitea nahi dugu, hau da, estimazio posible guztien erdiko balioa benetako balioa izatea eta bere ingurutik hurbil egotea. Estatistika ikuspuntutik, alboragabea izatea $(E(\hat{\beta}_{KTA}) = \beta)$ eta bariantza minimodunekoa izatea nahi dugu. Eredu orokorreko koefiziente guztien populazio bariantzak eta kobariantzak ateratzeko ondorengo adierazpena dugu:

$$\text{Bar}(\hat{\beta}_{KTA}) = \sigma^2(X'X)^{-1}.$$

Honela bada eta azken oinarritzko hipotesian adierazten den perturbazioen Normaltasunean oinarrituz, estimatzailearen banaketa atera dezakegu:

$$\hat{\beta}_{KTA} \sim N(\beta, \sigma^2(X'X)^{-1}). \quad (3.3)$$

Praktikan, bariantza eta kobariantza matrizea estimatzeko, perturbazioen bariantza (σ^2) estimatu behar dugu. Azken hau estimatzeko erabiliko dugun estimatzaile alboragabea aurreko gaian ikusitakoa izango da:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^N \hat{u}_i^2}{N - K}$$

non hondarrak aldagai azaldua eta bere estimazioaren arteko diferentziak diren: $\hat{u}_i = Y_i - \hat{Y}_i$. Eredu orokorrean gaudenez eta lagin erregresio funtzioa $\hat{Y}_i = \hat{\beta}_1 + \dots + \hat{\beta}_K X_{Ki}$ denez, hondar hauek $\hat{u}_i = Y_i - \hat{\beta}_1 - \dots - \hat{\beta}_K X_{Ki}$ bezala kalkulatzeko dira edota matrizialki, $\hat{u} = Y - X\hat{\beta}_{KTA}$ eginez. Hortaz, KTA estimatzailearen bariantza eta kobariantza matrizea estimatzeko erabiliko dugun estimatzaile alboragabea honakoa da:

$$\widehat{\text{Bar}}(\hat{\beta}_{KTA}) = \hat{\sigma}^2(X'X)^{-1}.$$

Gogora ezazue diagonal nagusian koefiziente bakoitzaren KTA estimatzailearen bariantza estimatuak agertzen direla.

Adibidearekin jarraituz, **A Ereduko** KTAko estimatzailearen bariantza eta kobariantza matrizea estimatzerakoan lortzen dugun emaitza ondorengo da:

3.2 Taula: Koefizienteen tartezko estimazioa

$$t(10, .025) = 2,228$$

ALDAGAIA	KOEFIZIENTEA	%95eko KONFIDANTZA TARTEA
const	129,062	(-67,6903, 325,814)
sqft	0,154800	(0,0836321, 0,225968)
bedrms	-21,5875	(-81,8126, 38,6376)
baths	-12,1928	(-108,560, 84,1742)

Honela, bigarren koefizientearen estimatzailearen bariantza estimatua $\widehat{var}(\hat{\beta}_2) = 0,00102019$ da eta bere desbideratze tipikoa $\widehat{des}(\hat{\beta}_2) = \sqrt{\widehat{var}(\hat{\beta}_2)} = \sqrt{(0,00102019)} = 0,0319404$ da. Diagonal nagusitik kanpo dauden elementuak kobariantza estimatuak dira. Honela, $\hat{\beta}_2$ eta $\hat{\beta}_4$ estimatzaileen arteko kobariantza estimatua $\widehat{cov}(\hat{\beta}_2, \hat{\beta}_4) = -0,995066$ da.

Aurreko gaian ikusi den bezala, koefizienteen tartezko estimazio bat lortzea posible da ere. Horretarako bi aukera izango ditugu. Lehenengoa, ondorengo formula aplikatuz norberak kalkulatzeko da:

$$KT(\beta_i)_{1-\alpha} = \left(\hat{\beta}_i \pm t_{(N-K)\alpha/2} \widehat{des}(\hat{\beta}_i) \right).$$

Adibidez, atera ditugun estimazioetan oinarrituz eta tauletako balioa $t_{(14-4)0,05/2} = 2,228$ dela jakinik, β_2 koefizientearen %95eko konfidantza tartea ondorengo da:

$$KT(\beta_2)_{1-0,05} = (0,154800 \pm 2,228 \times 0,0319404) = (0,083 \quad ; \quad 0,225)$$

Bigarren aukera Gretl programaren bitartez kalkulatzeko da, aurreko gaian ikusi den bezala, eredu KTA bitartez estimatu ondoren lortzen den lehiatilan, *Analisisa* $\rightarrow$ *Koefizienteen konfidantza tarteak* aukera erabiliz. Irtetzen den emaitza ondorengo da:

Konfidantza tarte bakoitzean koefizientearen balio posible guztiak agertzen dira, beti ere $1 - \alpha = 0,95$ konfidantzarekin. Beste era batera esanda, benetako balioa (β_j) tarte horren barnean aurkituko da 0,95eko probabilitatearekin. Tartea zenbat eta estuagoa izan, hau da, koefiziente horren estimatzailearen bariantza estimatua zenbat eta txikiagoa izan, hobeto edo zehatzago estimatzen ariko gara.

Azkenik, gogoratu ere konfidantza tarteek kontrasteak burutzeko balio dutela, honela $H_0 : \beta_j = c$ kontrastatu nahi badugu $H_a : \beta_j \neq c$ hipotesiaren aurka, orduan $c \in KT(\beta_j)_{1-\alpha}$ ematen bada, ez dugu hipotesi hutsa baztertuko α esangura-maila batentzat.

3.2.3 Banakako eta baterako esanguratasunak

Banakako esanguratasuna

Eredu orokorrean aldagai azaltzaile baten banakako esanguratasuna kontrastatzerakoan, kontuan izan behar dugu, “aldagai azaltzaile horrek eskaintzen duen efektu gehigarriaren” nabaritasuna kontrastatzen ari garela. Banakako esanguratasuna edo nabaritasuna kontrastatzeko hipotesi hutsa eta aurkakoa ondorengoak dira:

$$\begin{aligned} H_0: \beta_j &= 0 \\ H_a: \beta_j &\neq 0 \end{aligned}$$

Eta erabiliko dugun kontrasterako estatistikoa berriz:

$$t_{est\ j} = \frac{\hat{\beta}_j}{\widehat{des}(\hat{\beta}_j)} \stackrel{H_0}{\sim} t_{(N-K)}$$

Hipotesi hutsa ez dugu baztertzen α esangura-mailarekin baldin eta estatistikoaren balioa konfidantza tartean erortzen bada: $t_{est\ j} \in (-t_{(N-K)\alpha/2}, t_{(N-K)\alpha/2})$ edota $|t_{est\ j}| < t_{(N-K)\alpha/2}$ bada. Bestelako kasuan hipotesi hutsa baztertu egingo dugu. Hipotesi hutsa ez bada baztertzen, orduan koefiziente hori estatistikoki zero dela esango dugu eta dagokion aldagai azaltzailea nabaria ez dela baieztatuko dugu, beti ere α esangura-mailarekin.

Gure adibidera bueltatuz, **A Ereduko** aldagaien banakako esanguratasunak aztertuko ditugu. Estimazioaren emaitzetako azken aurreko zutabean (**t-estatistikoa**) agertzen diren estatistikoetan oinarrituz, banakako esanguratasun kontrastea egiterakoan, hipotesi hutsa baztertu egingo dugu baldin eta zutabeko balioak (balio absolutuan) tauletako balioa $t_{(N-K)\alpha/2} = t_{(14-4)\alpha/2} = 2,228$ baino handiagoak badira. Honela, adibide honetan, $H_0: \beta_2 = 0$ hipotesi hutsa baztertuko dugu $|4,847| > 2,228$ delako $\alpha = 0,05$ eko esangura-mailarekin eta ondorioz, SQFT (etxebizitzaren azalera) aldagaia nabaria dela esango dugu. Gelditzen diren beste bietan berriz, $H_0: \beta_3 = 0$ eta $H_0: \beta_4 = 0$ hipotesi hutsak ez ditugu baztertuko $|-0,799| < 2,228$ eta $|-0,282| < 2,228$ baitira $\alpha = 0,05$ eko esangura-mailarekin, hau da ez BEDRMS (logelen kopurua) ezta BATHS ere (komun kopurua), ez dira banaka nabariak.

Bestalde, Gretlek emandako emaitzetan oinarrituz, banakako esanguratasunak kontrastatzeko beste era bat dago: *p-balioa* erabiliz. Definizioz, aztertzen ari garen laginean oinarrituz, *p-balioak* hipotesi hutsa baztertzeko behar den esangura-mailarik txikiena ematen digu. Izatez, alde biko kontrasteetan *p-balioak*, ezkerrean uzten den azaleraren bikoitza ematen du:

$$p\text{-balioa} = 2 P(t_j > t_{est} | H_0).$$

Honela, zenbat eta *p-balioa* handiagoa izan, hipotesi hutsaren kontrako ebidentzia estatistikoa txikiagoa izango da. Zein da kontrasterako erabaki araua *p-balioan* oinarrituz? Ateratako *p-balioa* aukeratu dugun α baino txikiagoa izanez gero hipotesi hutsa baztertu egingo dugu. Kontrakoa ematen bada ez da hipotesi hutsa baztertzen. Azken finean, erabaki arau hau erabiltzea edo orainartekoa erabiltzea, berdina da.

Adibidera itzuliz eta aurreko banakako esanguratasunak $\alpha = 0,05$ eko esangura-mailarekin egin direnez, emaitzetan agertzen diren *p-balioak* esangura-maila honekin konparatuko ditugu. Erraz ikusten den bezala, $\alpha = 0,05$ balioa baino txikiagoa duen *p-balio* bakarra SQFT aldagaiari dagokiona da. Hortaz, $H_0: \beta_2 = 0$ hipotesi hutsa baztertuko dugu eta SQFT

aldagai nabari bat dela esango dugu. Beste p-balio biak $\alpha = 0,05$ balioa baino handiagoak direnez, dagozkien hipotesi hutsak ez dira baztertzen, hortaz, BEDRMS eta BATHS aldagai azaltzaileak ez dira banaka nabariak. Ikusten den bezala, ondorio berdinetara heltzen gara.

Banakako kontrasteekin bukatzeko, estimazio lehiatilako emaitzetako azken zutabeko “izarrek” zer adierazten duten azaltzera goaz. Kalkulatutako p-balioak izar bakarra duenean, hipotesi hutsa $\alpha = 0$, leko esangura-mailarekin baztertzen dela adierazten du. Izar bi agertzerakoan, hipotesi hutsa $\alpha = 0,05$ eko esangura-mailarekin baztertzen dela esan nahi du eta hiru izar baldin badaude, hipotesi hutsa $\alpha = 0,01$ eko esangura-mailarentzat baztertzea posiblea dela adierazten du. SQFT aldagaiari dagokion p-balioak hiru izar dituenek, $H_0 : \beta_2 = 0$ hipotesi hutsa $\alpha = 0,01$ eko esangura-mailarekin baztertzea posiblea dela adierazten du. Egia esan, emaitza hau bagenekien zeren eta ematen duen p-balioa $\alpha = 0,01$ esangura-maila baino txikiagoa baita. Azkenik, hipotesi hutsa $\alpha = 0,01$ esangura-mailarekin baztertzen badugu, orduan argi dago esangura-maila handiagoentzat ere ($\alpha = 0,05$ edota $\alpha = 0,1$) hipotesi hutsa baztertuko dela. Azken hau konprobatzeko, kasu bakoitzean estatistikoaren balio absolutua ($|t_{est\ j}|$) tauletako balioekin konparatuko dugu:

$$|t_{est\ 2}| = 4,847 > 3,169 = t_{(10)0,001} \quad H_0 : \beta_2 = 0 \text{ baztertu } \alpha = 0,01 \text{ rentzat}$$

$$|t_{est\ 2}| = 4,847 > 2,228 = t_{(10)0,025} \quad H_0 : \beta_2 = 0 \text{ baztertu } \alpha = 0,05 \text{ rentzat}$$

$$|t_{est\ 2}| = 4,847 > 1,812 = t_{(10)0,05} \quad H_0 : \beta_2 = 0 \text{ baztertu } \alpha = 0,1 \text{ rentzat}$$

BEDRMS eta BATHS aldagaientzat berriz, egiaztapen berdina eginez, hipotesi hutsak ez dira baztertzen edozein esangura-mailarentzat (0,1; 0,05 eta 0,01).

Hortaz, banakako kontrasteen konklusioa (erabiltzen den erabaki araua edozein izanik) SQFT aldagaia banaka nabaria dela da eta beste aldagai biak, BEDRMS eta BATHS, banaka ez dira nabariak. Horrela bada, eremuan behin SQFT aldagaia izanik, BEDRMS eta BATHS aldagaiek eskeintzen duten “informazio gehigarriak” ez du merezi. Izan ere, SQFT aldagaiak bi hauen informazioa barneraturik du, normalean logela edo komun bat gehiago izaterakoan, azalera handitu egiten baita.

Baterako esanguratasuna

Eredu bakunean, termino konstanteaz gain, beste aldagai azaltzaile bakar bat zegoenez, baterako esanguratasunak ez zuen inolako zentzurik. Baina eredu orokorrean, eremuan barneratutako aldagai azaltzaileak aldagai azaldua azaltzen duten jakin nahi izango dugu. Horretarako, aldagai azaltzaileen baterako esanguratasuna kontrastatuko dugu, behar dugun hipotesi hutsa eta aurkakoa ondorengoak izanik:

$$H_0: \beta_2 = \beta_3 = \beta_4 = \dots = \beta_K = 0$$

$$H_a: \text{berdintzaren bat ez da ematen}$$

non termino konstanteari dagokion koefizientea **ez** den barneratzen. Kontrasterako estatistikoak aldiz ondorengoak da:

$$F = \frac{R^2/(K-1)}{(1-R^2)/(N-K)} = \frac{R^2}{(1-R^2)} \frac{(N-K)}{(K-1)} \stackrel{H_0}{\sim} \mathcal{F}_{K-1, N-K}. \quad (3.4)$$

Hipotesi hutsa baztertu egingo dugu, baldin eta $F > \mathcal{F}_{(K-1, N-K)\alpha}$ bada eta aldagai azaltzaileak batera nabariak direla esango dugu, α esangura-mailarekin, hau da, guztien artean aldagai azalduaren bariantza azaltzeko informazioa daukate. Bestela, $F \leq \mathcal{F}_{(K-1, N-K)\alpha}$ bada, hipotesi hutsa ez dugu baztertzen α esangura-mailarekin eta aldagaiak nabariak ez direla esango dugu. Azken kasu honetan, zehazpen errore baten aurrean gaude, ereduak ez da ona, ereduko aldagai azaltzaile guztiak kontuan izanik ez baita aldagai dependentearen bariantza azaltzen. Horrela, kontraste honen bitartez ereduaren doikuntzaren ontasuna kontrastatzen da.

Gure adibidera itzuliz, aldagaien baterako esanguratasuna kontrastatzeko hipotesiak

$$H_0: \beta_2 = \beta_3 = \beta_4 = 0$$

$$H_a: \text{berdintzaren bat ez da ematen}$$

dira, estatistikoa kalkulatzeko mugatze koefizientea $R^2 = 0,835976$ da, estimatu beharreko koefizienteen kopurua $K = 4$ da eta behaketa kopurua $N = 14$ da. Beraz, baterako esanguratasun kontrastea burutzeko estatistikoaren balioa honakoa litzateke:

$$F = \frac{0,835976}{1 - 0,835976} \frac{(14 - 4)}{(4 - 1)} = 16,9889.$$

Dena den, estatistikoaren balio hau, Gretleko KTAko estimazio emaitzetan agertzen da:

$$F\text{-estatistikoa } (3, 10) = 16,9889 \text{ (p-balioa} = 0,000299)$$

eta alboan, dagokion p-balioa agertzen da. Horrela, kontrastea burutzeko bi aukera ditugu. Lehena, $F = 16,9889 > \mathcal{F}_{(3,10)0,05} = 3,71$ enez hipotesi hutsa baztertu genuke $\alpha = 0,05$ eko esangura-mailarekin eta ondorioz aldagai azaltzaileak batera nabariak direla esango genuke. Bigarren aukera p-balioa erabiltzean datza: ateratzen den p-balioa $\alpha = 0,05$ balioa baino txikiagoa denez, hipotesi hutsa baztertu egingo dugu $\alpha = 0,05$ eko esangura-mailarekin. Izatez, $\alpha = 0,01$ balioa baino txikiagoa denez, hipotesi hutsa baztertu egingo da $\alpha = 0,1$, $\alpha = 0,05$ eta $\alpha = 0,01$ esangura-mailarekin.

3.3 Doikuntzaren ontasuna eta ereduaren sailkapena

Aurreko gaitan, ereduaren termino konstante bat egonik, mugatze koefizientearen adierazpena eta esanahia ikusita dauzkagu:

$$R^2 = \frac{\sum_t (\hat{Y}_t - \bar{Y})^2}{\sum_t (Y_t - \bar{Y})^2} = 1 - \frac{\sum_t \hat{u}_t^2}{\sum_t (Y_t - \bar{Y})^2}$$

Doikuntza neurri hau aldagaien neurri unitateekiko independentea da $0 \leq R^2 \leq 1$ betetzen delako. Aldagai azaltzaileen bariantzarekin aldagai azalduaren bariantzaren zein portzentai azaldu ahal den neurtzen du era lineal batean.

Orokorrean, koefiziente hau zenbat eta handiagoa izan, hau da, batetik zenbat eta hurbilago egon, hobe izango da. Baina *batzuetan* batetik oso hurbil egoteak, lagin arazoren bat egon daitekeela adierazten du (kasu hau hurrengo gai batean aztertuko da). Normalean, denbora

serietako datuak aztertzean lortzen den mugatze koefizientea, gurutzatutako datuekin lortzen dena baino handiagoa izaten da, datuek dituzten ezaugarrietan oinarritzen baita. Denbora serietan entitate edo banako ekonomiko baten datuak dauzkagu denboran zehar, hau da, banakoari dagozkion aldagaien garapena analizatzen ari gara. Baina gurutzatutako datuetan, denbora finko mantentzen da eta banakoak aldatzen dira. Beraz, banako desberdinen aldagai berdinak analizatzen ari gara. Banako guztiak oso antzekoak izanez gero, orduan R^2 altua lortzea posiblea litzateke, baina banakoen ezaugarriak desberdintzen doazen neurrian, R^2 jaisten joango da nahiz eta ereduak ondo zehaztuta egon.

Ereduen konparaketa eta sailkapena R^2 -ren funtzioan egin nahi izanez gero, aldagai azaldua derrigorrez **berdina** izan beharko da, ez bakarrik behaketa kopuru berdina baizik eta behaketa berdina. Nahiz eta horrela izan, mugatze koefizienteak arazo bat aurkeztu du. Ereduan aldagai azaltzaileak barneratzen goazen neurrian, mugatze koefizientearen balioa handitzen doa. Hasiera batean, “handitzearekin” ados egon gaitezke baldin eta barneratzen diren aldagaiak nabariak badira, hau da, aldagai azaltzaile berri bakoitzak eskeintzen duen informazio gehigarriak merezi badu. Arazoa, aldagai ez nabariak barneratzerakoan mugatze koefizientea handitzearekin agertzen da. Arazo hau konpontzeko, zuzendutako mugatze koefizientea erabili daiteke:

$$\bar{R}^2 = 1 - \frac{\sum \hat{u}_t^2 / (N - K)}{\sum (Y_t - \bar{Y})^2 / (N - 1)} = 1 - \frac{(N - 1)}{(N - K)} (1 - R^2) \quad -\infty < \bar{R}^2 \leq R^2.$$

Doikuntza neurri honek aldagai ez nabarien barnerapena kontuan hartzen du. Zuzendutako mugatze koefizientearen balio maximoa mugatze koefizientearen baliotik oso hurbil egongo da, ereduko aldagai azaltzaile guztiak banaka nabariak direla adieraziz. Baina bestalde, balio negatiboak ere har ditzake, kasu horretan ereduen zehazpena okerra dela adierazten duelarik, hau da, aldagai azaltzaileak nabariak ez direnean.

Egitez, ereduan zenbat eta aldagai gehiago barneratu ($K \uparrow$) hainbat eta koefiziente gehiago estimatu behar dira, askatasun graduak murriztuz ($(N - K) \downarrow$) eta ondorioz, askatasun gradu gutxiago izaterakoan, estimazioaren zehaztasuna txikitu egiten da. Beraz, aldagai ez nabariak barneratzerakoan, askatasun graduak galtzen ari gara inolako irabazirik gabe, ez baitute aldagai azaldua azaltzeko informaziorik.

Zuzendutako mugatze koefizienteak, ereduan aldagai berri bat barneratzerakoan lortuko den irabazia (“informazio gehigarria”) eta agertzen den galera (“askatasun graduen murrizketa”) neurtu eta baloratu egiten du. *Orokorrean* irabazia handiagoa denean zuzendutako mugatze koefizientea handitu egingo da eta galera handiagoa denean jaitsi egingo da. *Hala ere*, zuzendutako mugatze koefizienteak ere badu bere muga: aldagai azaltzaile bati dagokion t-estatistikoaren balio absolutua bat baino handiagoa denean ($|\hat{\beta}_j / \widehat{des}(\hat{\beta}_j)| > 1$), aldagai hori ereduan mantentzean, nahiz eta banakako esanguratasun kontrastean ez nabaria dela ($|\hat{\beta}_j / \widehat{des}(\hat{\beta}_j)| > 1 < t_{(N-K)\alpha/2}$) irten, zuzendutako mugatze koefizientearen balioa handitu egiten da beti.

Gretlek ematen dituen KTAko estimazio emaitzetan ereduak konparatzeko beste neurri batzuk agertzen dira: AIC (Akaike Informazio Irizpidea) eta BIC (Bayes Informazio Irizpidea). Neurri hauen helburua, zuzendutako mugatze koefizientearena ($\bar{R}^2$) bezalakoa da, ereduan aldagai berri bat barneratzerakoan irabazten den informazio gehigarria eta galtzen diren askatasun graduak ebaluatu, aldagai berria barneratzeak merezi duen ala ez erabakitzeko. Neurri

3.3 Taula: Estimatu eta konparatuko diren zehazpen desberdinak

A Eredua	$PRICE_i = \beta_1 + \beta_2 SQFT_i + \beta_3 BEDRMS_i + \beta_4 BATHS_i + u_i$
B Eredua	$PRICE_i = \lambda_1 + \lambda_2 BEDRMS_i + u_i$
C Eredua	$PRICE_i = \theta_1 + \theta_2 BATHS_i + u_i$
D Eredua	$PRICE_i = \delta_1 + \delta_2 BEDRMS_i + \delta_3 BATHS_i + u_i$
E Eredua	$PRICE_i = \alpha_1 + \alpha_2 SQFT_i + u_i$
F Eredua	$PRICE_i = \gamma_1 + \gamma_2 SQFT_i + \gamma_3 BEDRMS_i + u_i$
G Eredua	$PRICE_i = \mu_1 + \mu_2 SQFT_i + \mu_3 BATHS_i + u_i$

hauek hondar karratuen batura ($\sum \hat{u}_i^2$, non $\hat{u}_i = Y_i - \hat{Y}_i$ estimatzerakoan egiten den errorea den) penalizatzen dute askatasun graduen ($N - K$) galeragatik. Honela, aldagai gehiago izaterakoan Hondar Karratuen Batura txikitu egingo da baina askatasun graduak galtzeagatik erabilitako penalizazioa handitu egingo da. Neurri hauek “errore neurriak” direnez, AIC edota BIC txikiena duen eredua aukeratuko genuke.

Gure adibidera bueltatuz, etxebizitzaren salmenta prezioa zehazteko eredu desberdinak proposatuko ditugu eta beraien artetik bat aukeratuko, estimazio emaitzek ematen duten informazio guztia ebaluatuz eta aztertuz. Aztertuko diren zehazpen desberdinak honakoak dira:

Eredu hauen arteko diferentziak barneratutako aldagai azaltzaileetan datza, aldagai azaldua eta lagina berdinak izanik. Aldagai azaltzaile bati dagokion koefizientea desberdina da eredu-tik eredura, aldagai horren eragina aldagai azalduan desberdina delako eta beraz, koefiziente estimatuaren balioa ez da zertan berdina izan behar.

Ereduak konparatuz, **A Eredua** orokorra da zeren eta besteek barneraturiko aldagai azaltzaile guztiak baititu. Honela, eredu hau oinarritzat har dezakegu, besteak berekiko konparatuz azpi-eredu bat izango balira bezala. Izatez, **A Eredutik B Eredua** lortzeko, SQFT eta BATHS aldagaiak kendu beharko lirateke eta helburu hau lortzeko, aldagai hauei dagozkien koefizienteak zero direla *inposatu* behar da. Hau da, **A Ereduan** $\beta_2 = 0$ eta $\beta_4 = 0$ murrizketak inposatuz **B Eredua** lortuko genuke. Antzera egin beharko genuke gainontzeko ereduak lortzeko.

Hala ere, nola jakin zein den eredu egokia? Eredu guztietatik zeinekin geldituko garen erabakitzea edota *inposatzen* ari garen murrizketak egiazkoak diren aztertzea gauza bera da. Eredu egokiena aukeratzeko esanguratasun kontrasteetan eta Akaike Informazio eta Bayes Informazio Irizpideetan oinarrituko gara.

Zehaztutako ereduaren emaitzak konparatzeko **E**, **F** eta **G Ereduen** estimazio emaitzak ondoren aurkezten dira:

E Eredua: KTA estimazioak 14 behaketak erabiliz 1–14

Aldagai azaldua: price

Aldagaia	Koefizientea	Desb. Tipikoa	t -estatistikoa	p-balioa
const	52,3509	37,2855	1,4041	0,1857
sqft	0,138750	0,0187329	7,4068	0,0000

Hondar Karratuen Batura	18273,6
Hondarren desbideratze tipikoa ($\hat{\sigma}$)	39,0230
R^2	0,820522
Zuzendutako $\bar{R}^2$	0,805565
Akaike Informazio Irizpidea	144,168
Schwarz Bayesian Irizpidea	145,447

F Eredua: KTA estimazioak 14 behaketak erabiliz 1–14

Aldagai azaldua: price

Aldagaia	Koefizientea	Desb. Tipikoa	t -estatistikoa	p-balioa
const	121,179	80,1778	1,5114	0,1589
sqft	0,148314	0,0212080	6,9933	0,0000
bedrms	−23,910	24,6419	−0,9703	0,3527

Hondar Karratuen Batura	16832,8
Hondarren desbideratze tipikoa ($\hat{\sigma}$)	39,1185
R^2	0,834673
Zuzendutako $\bar{R}^2$	0,804613
Akaike Informazio Irizpidea	145,019
Schwarz Bayesian Irizpidea	146,936

G Eredua: KTA estimazioak 14 behaketak erabiliz 1–14

Aldagai azaldua: price

Aldagaia	Koefizientea	Desb. Tipikoa	t -estatistikoa	p-balioa
const	79,5053	61,7859	1,2868	0,2246
sqft	0,152570	0,0312901	4,8760	0,0005
baths	−22,723	40,5073	−0,5610	0,5861

Hondar Karratuen Batura	17765,3
Hondarren desbideratze tipikoa ($\hat{\sigma}$)	40,1874
R^2	0,825514
Zuzendutako $\bar{R}^2$	0,793789
Akaike Informazio Irizpidea	145,774
Schwarz Bayesian Irizpidea	147,691

Azterketa errazagoa izan dadin, erregresio hauetan lortutako zenbait emaitza 3.4 taulan laburbiltzen dira. Eredu guzti hauen arteko konparaketa egiteko esanguratasun kontrasteekin

3.4 Taula: Zehazpen desberdinen estimazio emaitzen laburpena.

Eredua	Aldagaia	Esanguratsua ($\alpha = 0,05$)	R^2	$\bar{R}^2$	AIC	BIC
A EREDUA	SQFT BEDRMS BATHS	Bai Ez Ez	0,8359	0,7867	146,908	149,464
B EREDUA	BEDRMS	Ez	0,09962	0,02459	166,747	168,025
C EREDUA	BATHS	Bai	0,4483	0,4024	159,888	161,166
D EREDUA	BEDRMS BATHS	Ez Bai	0,4507	0,3508	161,829	163,746
E EREDUA	SQFT	Bai	0,820522	0,8055	144,168	145,447
F EREDUA	SQFT BATHS	Bai Ez	0,8346	0,8046	145,019	146,936
G EREDUA	SQFT BEDRMS	Bai Ez	0,8255	0,7937	145,774	147,691

hasiko gara. Taulan eredu bakoitzean barneratutako aldagaien esanguratasuna jarri da, dago-kion t-estatistikoa Student-t banaketako koantilarekin konparatu ondoren. Ikusi dezakegunez, BEDRMS aldagaia barneratu den eredu guztietan, ez esanguratsua irteten da baina BATHS berriz, SQFT aldagaia ere barneratzen denean ez esanguratsua da eta bakarrik edota BEDRMS aldagaiarekin barneratzean esanguratsua da. Badirudi, behin ereduan SQFT aldagai azaltzailea egonik, beste aldagaiek, BEDRMS eta BATHS, ez dutela informazio gehigarrik eransten eta beraz, ez duela merezi hauek sartzeak, ez baitira banaka esanguratsuak. Izan ere, logela eta komun gehiago izaterakoan ohikoena etxebizitzak duen azalera gehitzea da. Horrela bada, BEDRMS eta BATHS aldagaiek duten informazioa SQFT aldagaien barnean agertzen da nolabait.

Horrela izanik eta ereduren bat aukeratu nahi izanez gero, zalantza **A** eta **E Ereduen** artean izango genuke. Egin dezagun bi eredu hauen arteko azterketa sakonago bat. **A Ereduekin** hasiko gara, bera baita aldagai azaltzaile guztiak barneratzen duen eredu. Bertako mugatze koefizientea interpretatuz, SQFT, BEDRMS eta BATHS aldagai azaltzaileen bariantzarekin PRICE aldagai azalduaren bariantzaren %83,5a azaltzea lortzen da era lineal batean. Ikusi daitekeenez, guztietatik **A Eredukoa** da mugatze koefizienterik handiena. Eraitza hau espero genuen? Bai. Teorikoki arazo honen existentziaz bagenekien, ereduan zenbat eta aldagai azaltzaile gehiago barneratu, aldagai hauek nabariak izan ala ez, R^2 handitu egiten da eta HKB txikitu.

Horregatik, zuzendutako mugatze koefizienteak begiratuko ditugu eta gure adibidean, alde-rantzizko joera duela ondorioztatuko dugu: **E Eredutik A Eredura** pasatzean $\bar{R}^2$ jaisten doa. Izatez barneratutako aldagai berrien banakako esanguratasunen t-estatistikoen balio absolutuak bat balioa baino txikiagoak direnez, eredura eransten duten informazio gehigarriarekin lortzen diren irabaziek ez dute askatasun graduaren galera konpentsatzen. Horrela bada, zuzendutako mugatze koefizientearen oinarrituz, eredu hauetatik etxebizitzaren salmenta prezioa azaltzeko eredurik hoberena **E Eredua** izango litzateke.

AIC eta BIC irizpideetan oinarrituz, **E Eredutik A Eredura** pasatzean neurri hauen balioak handitzen doaz. Neurri hauek “errore neurriak” direla kontuan hartuz, BEDRMS eta BATHS aldagaiak barneratzerakoan errorearen neurri hauen gehikuntzak, eredu okerrago zehazten ari garela adierazten dute. Hortaz, irizpide hauek erabiliz, konklusio berdinerara heldu gara: **E eredu**a hoberen zehaztuta dagoen eredu da.

Atal honekin bukatzeko, nahiz eta erabilitako adibidean estimazio eraitza guztien konparaketek ondorio berdinerara zuzendu, aipatu beharra dago errealitatean ez dela beti horrela izaten. Horrelako kasu baten aurrean gaudenean, normalean bi susmo izaten ditugu: perturbazioari buruzko oinarritzko hipotesiren bat ez dela betetzen edota aldagai azaltzailearen bat aleatorioa eta perturbazioarekin erlazionatuta dagoela.

Bibliografia

Ramanathan, R. (2002), *Introductory Econometrics with Applications*, 5. ed., South-Western, Ohio.

4 Gaia

Murrizketa linealen kontrasteak eta aurreanak

Aurkibidea

4.1	Sarrera	72
4.2	Murrizketa linealen kontraste orokorra	72
4.3	Murrizketei baldintzaturiko Karratu Txikiaren Arruntetako estimatzailea	74
4.4	Murrizketa linealen kontrasteak Gretl erabiliz	76
4.5	Puntuzko eta tartezko aurreanak	80
4.6	Puntuzko eta tartezko aurreanak Gretl erabiliz	81

4.1 Sarrera

Gai honetan, koefizienteen edozein murrizketa linealen kontrastea egiteko estatistiko orokor bat aztertuko dugu. Aurreko gaietan azaldutako esanguratasun kontrasteak izaten direnez ekonometrian gehien erabiltzen direnak, erregresioak egiten dituen edozein programa informatikok automatikoki kontraste horiek egiteko informazioa ematen du. Hala ere, badira beste kontraste interesgarri batzuk ere, ekonomilariak portaera ekonomikoari buruzko teoriak garatu eta ebaluatzen baitituzte eta hipotesien kontrasteak, teoria hauek ebaluatzeko prozedurak dira.

Atal honetan ikusiko dugun bezala, estatistiko bakar batekin koefizienteen murrizketa linealen multzo bat kontrasta daiteke. Jakin, aurretik ikusitako banakako esanguratasun kontrasterako estatistikoa eta baterako esanguratasun kontrastekoa, jarraian lortuko dugun estatistikotik eratorriak izan daitezkeela. Aldi berean, Gretl erabiliz kontraste hauek nola egin aztertuko dugu.

4.2 Murrizketa linealen kontraste orokorra

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_K X_{Ki} + u_i \quad i = 1, 2, \dots, N$$

Aurreko erregresio linealeko eredu orokorrean, koefizienteen edozein q murrizketa lineal ($R\beta = r$ adierazpen matriziala izanik) kontrastatu nahi izanez gero: ezagunak diren konstanteen matrizea R ($q \times K$ ordenakoa) eta konstante ezagunen bektorea r ($q \times 1$ ordenakoa) definitu behar ditugu kontrastatu nahi diren murrizketen arabera. Hipotesi hutsan murrizketak matrizialki ($R\beta = r$) jarri behar direnez, jarraian adibide batzuen bitartez adierazten da nola osatu R eta r :

1. X_2 aldagai azaltzailearen banakako esanguratasun kontraste batean:

$$H_0 : \beta_2 = 0 \quad \text{vs} \quad H_a : \beta_2 \neq 0$$

$$R\beta = r \Rightarrow \underbrace{\begin{bmatrix} 0 & 1 & 0 & \dots & 0 \end{bmatrix}}_R \underbrace{\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix}}_{\beta} = \underbrace{(0)}_r$$

Hipotesi hutsa $H_0 : \beta_2 = 8$ bada, aldatu beharreko gauza bakarra $r = (8)$ izango litzateke.

2. Baterako esanguratasun kontraste batean: $H_0 : \beta_2 = \beta_3 = \dots = \beta_K = 0$

$$R\beta = r \Rightarrow \underbrace{\begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & & & \ddots & \\ 0 & 0 & 0 & & 1 \end{bmatrix}}_R \underbrace{\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix}}_{\beta} = \underbrace{\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}}_r$$

3. Koefizienteen konbinazio linealen murrizketak:

(a) Murrizketa bat: $H_0 : \beta_2 + \beta_3 = 7$ vs $H_a : \beta_2 + \beta_3 \neq 7$

$$R\beta = r \Rightarrow \underbrace{\begin{bmatrix} 0 & 1 & 1 & 0 & \cdots & 0 \end{bmatrix}}_R \underbrace{\begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_K \end{bmatrix}}_{\beta} \underbrace{\begin{bmatrix} 7 \end{bmatrix}}_r$$

(b) Murrizketa bat baino gehiago:

$$H_0 : \begin{cases} 5\beta_2 + 4\beta_3 = 5 \\ \beta_1 - 4\beta_4 = 6 \end{cases} \quad \text{vs} \quad H_a : \begin{cases} 5\beta_2 + 4\beta_3 \neq 5 \\ \beta_1 - 4\beta_4 \neq 6 \end{cases} \quad \text{edota}$$

$$R\beta = r \Rightarrow \underbrace{\begin{bmatrix} 0 & 5 & 4 & 0 & \cdots & 0 \\ 1 & 0 & 0 & -4 & \cdots & 0 \end{bmatrix}}_R \underbrace{\begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_K \end{bmatrix}}_{\beta} \underbrace{\begin{bmatrix} 5 \\ 6 \end{bmatrix}}_r$$

Teoriara bueltatuz, ikusi dugunez, oinarritzko hipotesien menpean KTA estimatzailearen banaketa $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X'X)^{-1})$ da. Adierazitako murrizketa linealen kontrasteak egiteko, ordea, $R\hat{\beta}$ -ren banaketan oinarrituko gara:

$$\underbrace{R\hat{\beta}}_{(q \times 1)} \sim \mathcal{N}\left(\underbrace{R\beta}_{(q \times 1)}, \underbrace{\sigma^2 R(X'X)^{-1} R'}_{(q \times q)} \right) \quad (4.1)$$

Estatistiko honekin eta $\hat{\sigma}^2 = \frac{\hat{u}'\hat{u}}{N-K}$ estimatzailearekin, Wald-en F estatistiko bezala ezagutzen den hurrengo estatistikoa osatuko genuke:

$$F = \frac{\left(R\hat{\beta} - r\right)' \left(R(X'X)^{-1} R'\right)^{-1} \left(R\hat{\beta} - r\right) / q}{\hat{\sigma}^2} \stackrel{H_0}{\sim} \mathcal{F}_{(q, N-K)} \quad (4.2)$$

H_0 ez baldin bada egiazkoa, $(R\hat{\beta} - r)$ diferentzia handia izango denez, F estatistikoak balio altuak hartuko ditu eta H_0 baztertuko genuke α esangura-mailarekin. Hau da, hipotesi hutsa baztertuko dugu baldin eta $F > \mathcal{F}_{(q, N-K)\alpha}$ bada eta ez dugu baztertuko bestelako kasuan.

Bada ordea estatistiko honen baliokidea den beste estatistiko bat: murriztu gabeko ereduko hondar karratuen batura eta murriztutako ereduko hondar karratuen baturan oinarritzen den estatistikoa. Ikus dezagun beste adierazpen hau lortzeko burutu behar diren pausuak.

4.3 Murrizketei baldintzaturiko Karratu Txikiaren Arruntetako estimatzailea

Erregresio linealeko eredu orokorra izanik (matrizialki $Y = X\beta + u$), ereduko koefizienteek buruzko informazioa eskuragarri baldin bada, oso garrantzitsua da informazio berri hori egiazkoa den edo ez kontrastatzea zeren β ezezaguna denez, ez dakigu ziurtasunez $R\beta = r$ ematen den. Kontrastearen ondorioz, hau da, hipotsi hutsa ez bada baztertzen, informazioa egiazkoa dela ondorioztatuko dugu α esanguratasun mailarekin eta murrizketa kontuan hartuta lortuko genukeen estimatzaile berria (Murriztutako Karratu Txikiaren Arruntetako estimatzailea ($\hat{\beta}_{MKT}$ edo $\hat{\beta}_M$)), KTA estimatzailea baino efizienteagoa izango da, bariantza txikiagoa izango baitu.

Ereduko koefizienteak murrizketa hauek kontuan hartuta estimatu nahi izanez gero ($\hat{\beta}_M$), murrizketa hauei baldintzatutako Hondar Karratuen Batura minimo egin beharko genuke. Minimizazio prozeduraren ondorioz eta zenbait eragiketa eginik, honako emaitza lortuko genuke:

$$\hat{\beta}_M = \hat{\beta}_{KTA} + (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(r - R\hat{\beta}_{KTA}) \quad (4.3)$$

non aurreko atalean ikusi dugun bezala, koefizienteen edozein murrizketa multzo $R\beta = r$ bezala adierazten dugun eta R eta r matrizeak ezagunak diren, murrizketen arabera osatuak izanik.

KTA estimatzailearekin bezala, posible da KTM estimatzailearekin Hondar Karratuen Batura edota mugatze koefizientea lortzea:

$$HKB_M = \hat{u}'_M \hat{u}_M = (Y - X\hat{\beta}_M)'(Y - X\hat{\beta}_M) \quad \text{eta} \quad R_M^2 = 1 - (\hat{u}'_M \hat{u}_M / \sum (Y_t - \bar{Y})^2)$$

Dena den, bide hau jarraituz estimazioak lortzeko egin beharreko eragiketa kopurua handia da eta eskuz egin beharko genituzke. Bada ordea, Gretl erabiliz $\hat{\beta}_M$ lortzeko eta bide batez, HKB_M eta R_M^2 lortzeko beste metodo alternatibo bat: $R\beta = r$ murrizketak eredu ordezkatuz, **murriztutako eredu**a lortuz ($Y_* = X_*\beta_* + u_*$) eta bertako koefizienteak KTA bitartez estimatuz ($\hat{\beta}_* = (X'_*X_*)^{-1}X'_*Y_*$). Horrela, hasierako ereduko (**murriztu gabeko eredu**a) koefiziente batzuk bakarrik estimatuko genituzkeenez, gainontzekoak lortzeko murrizketetara bueltatuko ginateke.

Aukera hau adibide batekin ikusiko dugu. Izan bedi ondorengo erregresio eredu:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \quad i = 1, 2, \dots, N$$

non $\beta_2 = 5$ dela dakigun. Informazio hau eredu ordezkatuz, murriztutako eredu lortuko dugu. Aipatu bezala, eredu honetan Hondar Karratuen Batura minimo eginez, bertako koefizienteak estimatu ahal izango ditugu, hauek murrizketak betetzen dituztelarik.

Eredu honetan ordea, ez daude murriztu gabeko ereduko koefiziente guztiak eta murrizketetara zuzendu beharko gara guztiak lortzeko. Ikus dezagun:

$$\text{Hasierako eredua: } Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

$$\begin{aligned} \text{Murriztutako eredua: } Y_i &= \beta_1 + \beta_2 X_{2i} + 5X_{3i} + u_i^* \\ Y_i - 5X_{3i} &= \beta_1 + \beta_2 X_{2i} + u_i^* \\ Y_i^* &= \beta_1 + \beta_2 X_{2i} + u_i^* \end{aligned}$$

non $Y_i^* = Y_i - 5X_{3i}$ den. Hondar Karratuen Batura minimo eginik, β_1 eta β_2 estimatu ahal izango ditugu $\hat{\beta}_1^M$ eta $\hat{\beta}_2^M$ lortuz. Gainontzekoak murrizketetara bueltatuz lortuko genituzke.

Horrela, murriztutako estimatzailea $\hat{\beta}_M = \begin{bmatrix} \hat{\beta}_1^M & \hat{\beta}_2^M & 5 \end{bmatrix}'$ izango da.

Eredu berdinarekin jarraituz, beste adibide bat ondorengo litzateke: $\beta_1 + 3\beta_2 = 4$ murrizketari loturikoa edo baliokidea dena, $\beta_1 = 4 - 3\beta_2$.

$$\text{Hasierako eredua: } Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

$$\begin{aligned} \text{Murriztutako eredua: } Y_i &= \beta_1 + (4 - 3\beta_2)X_{2i} + \beta_3 X_{3i} + u_i^* \\ \underbrace{Y_i - 4X_{2i}}_{Y_i^*} &= \beta_1 + \beta_3 \underbrace{(X_{3i} - 3X_{2i})}_{X_i^*} + u_i^* \end{aligned}$$

Azken ere du honetan $\hat{\beta}_1^M$ eta $\hat{\beta}_3^M$ estimatzen dira eta ondoren $\hat{\beta}_2^M = 4 - 3\hat{\beta}_3^M$ lortzen da.

Murriztutako eredua estimatzerakoan, erregresio horretan lortuko genukeen Hondar Karratuen Batura ($HK B_* = \hat{u}'_M \hat{u}_M = Y'_* Y_* - \hat{\beta}'_* X'_* Y_*$) eta lehen lortutako $HK B_M$ berdina dira; hau da, hasierako ereduan murriztutako estimatzailearekin lortutako HKB eta murriztutako eredukoak berdina dira. Bestalde, murriztutako ereduko aldagai azaldua eta murriztu gabekoa berdina bada, murriztutako ereduko mugatze koefizientea (R_*^2) eta lehenago definitu duguna KTM estimatzailearekin (R_M^2), berdina dira. Guzti hau jakinik eta Gretl erabiliz, eragiketak asko errazten dira, baina lehenik, murriztutako ereduko aldagai berrien behaketak kalkulatu beharko genituzke eta aurreko gaitan ikusitakoa aplikatu: aldagai berriak definitu eta murriztutako ereduan KTA aplikatu.

Esan bezala, aipatutako HKB hauek erabiliz, posible da ere koefizienteen murrizketa linealen kontrastea burutzea, hau da, posible da lehen adierazitako estatistikoaren beste adierazpide baliokide bat lortzea: halaber, hipotesi hutsagatik murriztutako ereduko HKBn ($\hat{u}'_M \hat{u}_M$) eta murriztu gabeko ereduko HKBn ($\hat{u}' \hat{u}$), non $\hat{u}_M = Y - X \hat{\beta}_M$ eta $\hat{u} = Y - X \hat{\beta}_{KTA}$ diren, oinarrituriko estatistikoa ondorengo da:

$$F = \frac{(\hat{u}'_M \hat{u}_M - \hat{u}' \hat{u})/q}{\hat{u}' \hat{u}/(N-K)} \stackrel{H_0}{\sim} \mathcal{F}_{(q, N-K)} \quad (4.4)$$

non $\hat{u}'_M \hat{u}_M = (Y - X \hat{\beta}_M)'(Y - X \hat{\beta}_M)$ edota murriztutako ereduan ($Y_* = X_* \beta_* + u_*$), $\hat{u}'_* \hat{u}_* = Y'_* Y_* - \hat{\beta}'_* X'_* Y_*$. Lehen bezala, hipotesi hutsa baztertuko dugu baldin eta $F > \mathcal{F}_{(q, N-K)\alpha}$ bada eta ez dugu baztertuko bestelako kasuan.

Honen baliokidea den beste estatistiko bat mugatze koefizienteetan oinarritzen dena da. Al-

de batetik, murriztu gabeko eredua KTA bitartez estimatuz lortuko genukeena ($R^2 = R_{MG}^2$) eta bestalde, KTM bitartez lortuko genukeena (R_M^2) (azken hau murriztutako ereduko mugatze koefizientearen (R_*^2) berdina da, murrizketa kontuan hartuta aldagai azaldua ez bada aldatzen, hau da $Y = Y_*$ denean, $R_M^2 = R_*^2$ izango da). Honakoa da aipaturiko estatistikoa:

$$F = \frac{(R_{MG}^2 - R_M^2)/q}{(1 - R_{MG}^2)/(N - K)} \stackrel{H_0}{\sim} \mathcal{F}_{(q, N-K)} \quad (4.5)$$

Kontrastea aurreko estatistikoarekin bezala burutzen delarik.

Azken estatistiko hau aprobatxatuz, aldagai azaltzaileen baterako esanguratasun kontrastea aztertu genuenean proposatzen genuen estatistikoa gogoratuko dugu. Erregresio lineal eredu orokor batean izango genukeen hipotesi hutsa

$$H_0 : \beta_2 = \beta_3 = \dots = \beta_K = 0$$

izango litzateke eta honi atxikitutako murrizturiko eredua $Y_t = \beta_1 + u_t$ izango da non $R_*^2 = 0 = R_M^2$ izango den, horrelako kontraste baterako proposatzen genuen estatistikoa justifikatuz:

$$F = \frac{R^2/(K - 1)}{(1 - R^2)/(N - K)} \stackrel{H_0}{\sim} \mathcal{F}_{(K-1, N-K)} \quad (4.6)$$

4.4 Murrizketa linealen kontrasteak Gretl erabiliz

Murrizketa linealen kontrasteak formalki nola egingo genituzkeen azaldu dugu baina zenbait programa informatikok, besteak beste Gretl, estatistiko hauen balioa lortzeko aukera eskeintzen du (Wald-en testa), murrizketen kontrasteak egiteko aukera asko erreztuz.

Erregresio lineal eredu orokor batean eta Gretl programa erabiliz, koefizienteak nola adierazten ditugun eta programak nola adierazten dituen kontuan hartu beharra dago. Gretl programan, guk β_1 bezala definitu dugun koefizientea (hau da termino konstantea) $b1$ da, gure β_2 , $b2$ izango da, β_3 berriz $b3$ eta antzera ereduko koefiziente guztiekin. Hau da:

$$\text{Eredu orokorra: } Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_K X_{Ki} + u_i$$

$$\text{Gretl programan: } Y = b1 + b2 * X_2 + b3 * X_3 + \dots + bK * X_K + u$$

Horrela, koefizienteen edozein murrizketa ($R\beta = r$) adierazterakoan, kontuan izan beharko dugu deitura desberdintasunak eta biderkaketak nola adierazten diren. Adibidez, $\beta_1 + 5\beta_3 = 4$ eta $\beta_2 + 4,5\beta_4 = 2$ murrizketak, Gretlekin $b1 + 5 * b3 = 4$ eta $b2 - 4,5 * b4 = 2$ idatzi behar ditugu. Murrizketa bakoitza ekuazio bakar batean idatzi behar dugu, berdintzaren ezker aldean parametroen konbinazio lineala adieraziz eta eskubian dagokion zenbakia.

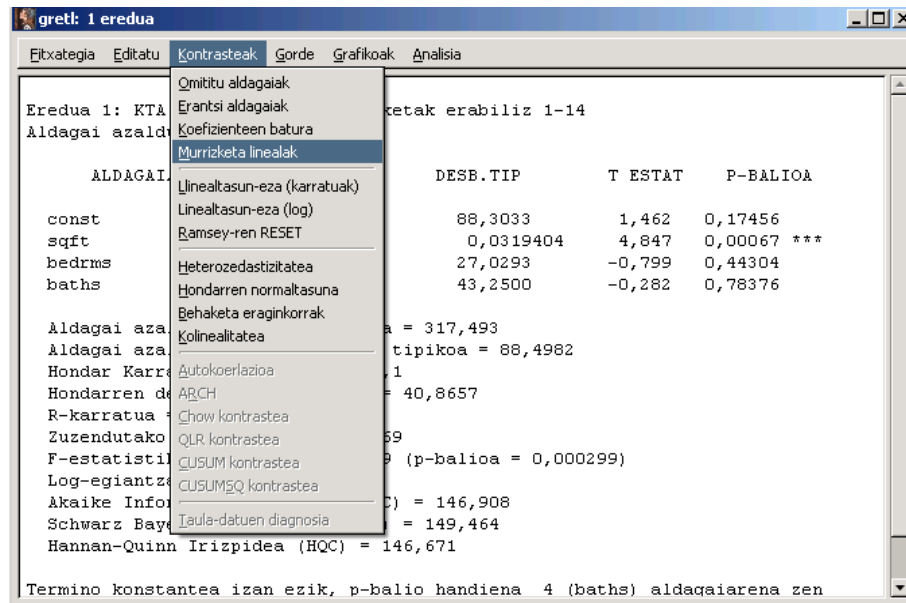
Etxebizitzaren prezioen adibideko ereduan¹

$$PRICE_i = \beta_1 + \beta_2 SQFT_i + \beta_3 BEDRMS_i + \beta_4 BATHS_i + u_i$$

ondorengo irudian azaltzen den bezala, eredua KTA bitartez estimatu ondoren lortzen den lehiatilan kontrasteak egin daitezke: *Kontrasteak* $\rightarrow$ *Murrizketa linealak*.

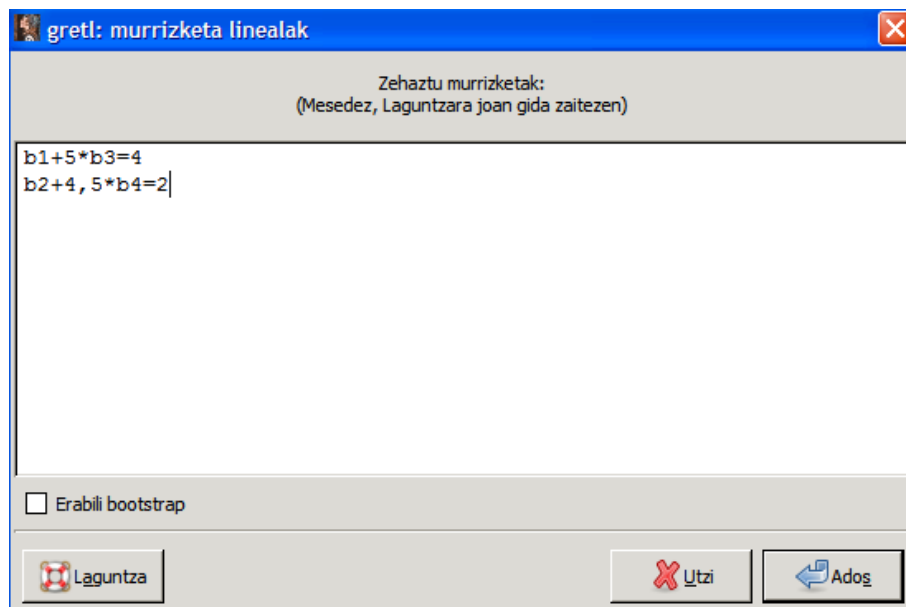
¹Ramanathaneko (2002) data4-1.gdt datu-fitxategia.

4.1 Irudia: Murrizketa linealen kontrasteak



Ondoren interesatzen zaizkigun murrizketak adieraziko ditugu:

4.2 Irudia: Murrizketa linealen zehazpena



Ados klikatzean, murrizketak kontuan hartuz lortzen diren murriztutako estimazio emaitzak agertzen dira. Hau da Murriztutako Karratu Txikiaren estimatzailea. Bertan, Wald-en F

4.3 Irudia: Murrizketa linealen kontrastearen emaitza

gretl: murrizketa linealak

Murrizketa multzoa

1: $b[\text{const}] + 5 \cdot b[\text{bedrms}] = 4$
 2: $b[\text{sqft}] + 4,5 \cdot b[\text{baths}] = 2$

Kontrasterako estatistikoa: $F(2, 10) = 0,0434871$, p-balioarekin = 0,957625

Murriztutako estimazioa:

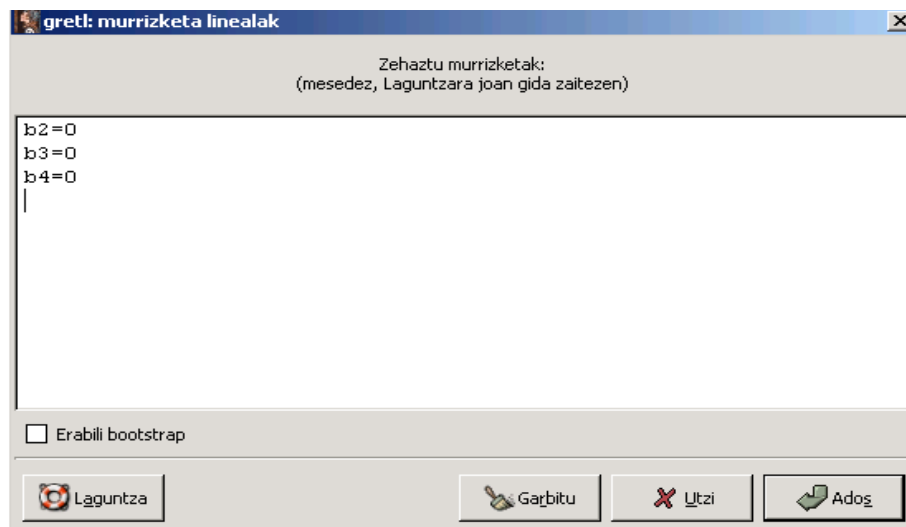
ALDAGAIA	KOEFIZIENTEA	DESB.TIP	T ESTAT	P-BALIOA
const	119,214	68,6318	1,737	0,10796
sqft	0,147222	0,00992816	14,829	<0,00001 ***
bedrms	-23,0428	13,7264	-1,679	0,11903
baths	0,411728	0,00220626	186,618	<0,00001 ***

Hondarren desbideratze tipikoa = 37,467

estatistikoaren balioa ere agertzen da ($F=0,04389$) eta dagokion p-balioa, kontrastea burutu ahal izateko.

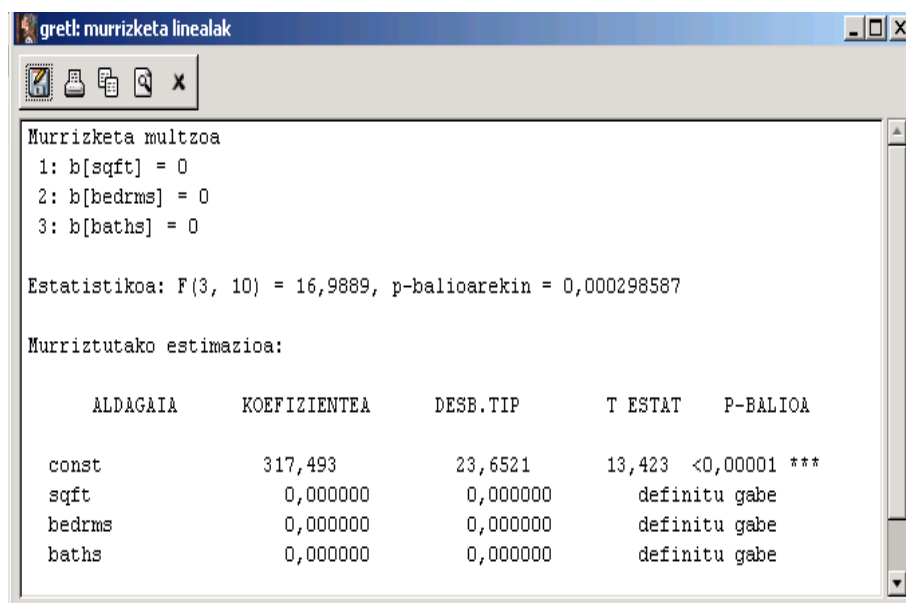
Horrela koefizienteen edozein murrizketa linealen kontrasteak burutu ditzakegunez, jada egin ditugun banakako eta baterako esanguratasun kontrasteak ere egin ditzakegu. Baterako esanguratasun kontrasteen adibidez, horrela adieraziko genuke:

4.4 Irudia: Baterako esanguratasunaren murrizketen zehazpena



Ondorengo emaitza lortuz:

4.5 Irudia: Baterako esanguratasun kontrastearen emaitza



Bertan ikusten dugun bezala, oraingoan proposatzen dugun Wald-en F estatistikoaren balioa 16,9889 da, hirugarren gaian ikusi genuen bezala. Horregatik, kontrastea era berean egin dezakegu, estatistikoaren balioarekin batera, alboan dagokion p-balioa agertzen baita. Dakusagunez, $F = 16,9889 > \mathcal{F}_{(3,10)0,05} = 3,71$ da eta hipotesi hutsa baztertu egiten da $\alpha = 0,05$ eko esangura-mailarekin, ondorioz aldagai azaltzaileak batera nabariak direlarik. Bide batez, p-balioa erabiliz, $\alpha = 0,01$ balioa baino txikiagoa denez, hipotesi hutsa baztertu egingo dugu $\alpha = 0,01$ eko esangura-mailarekin eta baita $\alpha = 0,1$ eta $\alpha = 0,05$ ekin.

4.5 Puntuzko eta tartezko auresanak

Ekonometriaren helburu nagusia ereduaren estimazio on baten lorpena dela pentsatzen bada ere, sarritan auresan zehatzak lortzea ere oso garrantzitsua da. Eredua zuzenki zehaztu ondoren, parametroak estimatu ditugu eta kontrasteak eginez, ereduari oniritzia eman diezaiokegu edo ez. Eraitza ezezkoa bada, eredia berriro zehaztu beharko dugu eta pausu guztiak berregin. Eredua onartuz gero, auresanak egiteko erabili dezakegu edota “zein izango litzatekeen aldagai azalduaren balioa aldagai azaltzaileek balio konkretu bat hartzen badute” motako galderei erantzun ahal izango diegu.

Suposa dezagun ondorengo eredia T behaketekin estimatu dela:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \dots + \beta_K X_{Ki} + u_i \quad i = 1, \dots, N$$

hurrengo lagineko erregresio funtzioa lortuz:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \dots + \hat{\beta}_K X_{Ki}$$

Orduan, aldagai azaltzaileen behaketa berri bat emanik,

$$X'_p = [1 \quad X_{2p} \quad \dots \quad X_{Kp}] \quad p \notin \{1, 2, \dots, N\}$$

KTAn bitartez estimatutako eredia erabili daiteke aldagai azalduak izango duen balioa auresateko (puntuzko auresana), hau da, balio hauek lagineko erregresio funtzioan ordezkatzuko genituzke, $\hat{Y}_p$ lortuz.

$$\hat{Y}_p = X'_p \hat{\beta}_{KTA}$$

Edo beste era batera:

$$\hat{Y}_p = \hat{\beta}_1 + \hat{\beta}_2 X_{2p} + \dots + \hat{\beta}_K X_{Kp}.$$

Hala ere, aldagai azalduaren p momentuko benetako balioa jakiterakoan, errore bat egin dugula antzemango dugu (auresanaren errorea: $e_p = Y_p - \hat{Y}_p$), aldagai azaltzaileen balioetan errorea dagoelako, β koefizienteen estimatzaileak erabili ditugulako, Y_p u_p -ren menpean dagoelako (hau da behaketa horri dagokion perturbazioaren menpe), etab. Horregatik, komenigarria izaten da tartezko auresan bat egitea, nolabait auresanaren zehaztasun neurri bat kontuan hartzen baitu.

Tartezko auresana lortzeko, auresanaren erroreak banaketan oinarrituko gara, u_p eta $\hat{\beta}$ aldagai aleatorio normalak badira, auresanaren errorea ere normala baita:

$$e_p \sim \mathcal{N}(0, \sigma^2 (1 + X_p' (X'X)^{-1} X_p)).$$

Orokorrean σ^2 ezezaguna izaten denez, bere estimatzaile alboragabea erabiliz, ondorengo banaketa lortuko genuke:

$$\frac{e_p}{\hat{\sigma} \sqrt{1 + X_p' (X'X)^{-1} X_p}} \sim t_{(N-K)}.$$

Eta hemendik, puntuzko aurreanaren inguruan, aldagai azalduak p momentuan hartuko duen balioaren aurreanaren tarte bat lortuko genuke $1 - \alpha$ konfidantza mailarekin, ondorengo adierazpena lortuz:

$$KT(Y_p)_{1-\alpha} = \left(\hat{Y}_p - t_{(N-K)\frac{\alpha}{2}} \hat{\sigma}_{e_p}, \hat{Y}_p + t_{(N-K)\frac{\alpha}{2}} \hat{\sigma}_{e_p} \right)$$

non $\hat{\sigma}_{e_p}^2 = \hat{\sigma}^2 (1 + X_p' (X'X)^{-1} X_p)$ eta $X_p' = [1 \quad X_{2p} \quad \cdots \quad X_{Kp}]$ diren.

4.6 Puntuzko eta tartezko aurreanak Gretl erabiliz

Aurreko atalean aurreanak nola lortzen diren aztertu dugu baina eskuz egin nahi izanez gero, egin beharreko eragiketak asko dira. Gretlek aurreanak lortzeko aukera eskeintzen duenez, etxebizitzaren prezioen ariketa erabiliko dugu Gretlen jarraitu beharreko pausuak zein diren jakiteko. Demagun honako eredua

$$PRICE_i = \beta_1 + \beta_2 SQFT_i + \beta_3 BEDRMS_i + \beta_4 BATHS_i + u_i \quad (4.7)$$

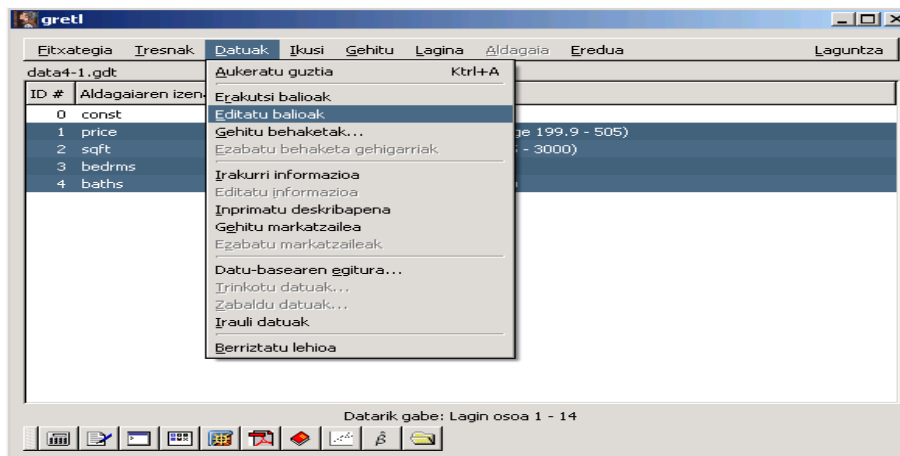
dugula eta beste etxebizitza baten ondorengo informazioa: $SQFT = 3200$, $BEDRMS = 5$ eta $BATHS = 3$ izanik, bere prezioa aurrean nahi dugula. Horretarako (4.7) eredua erabiliz ezaugarri hauetako etxebizitzaren prezioaren aurrean bat egin dezakegu eta bide batez, aztertu ea eskatzen diguten prezioa ($PRICE = 500$) arrazoizkoa den edo ez.

Gretlen bitartez aurrean hau egiteko lehen pausua datu berriak datu-basean barneratzea da: *Datuak* $\rightarrow$ *Aukeratu guztia* egin ondoren, *Datuak* $\rightarrow$ *Editatu balioak* aukeran

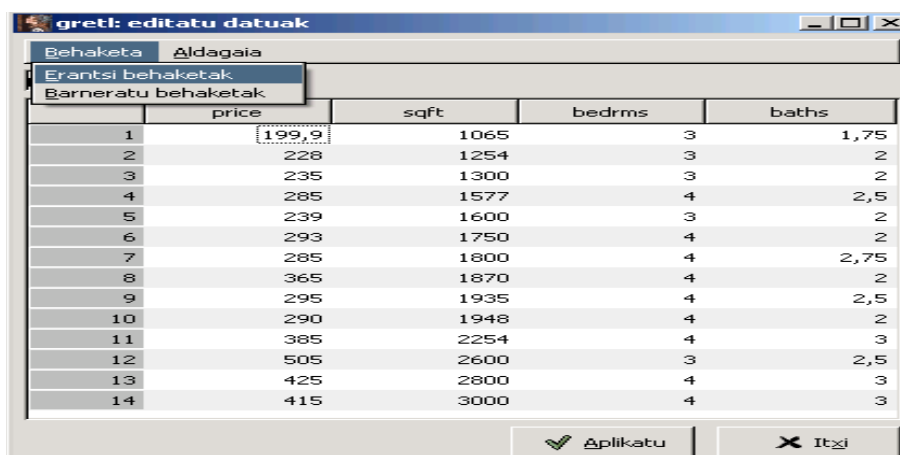
Behaketa $\rightarrow$ *Erantsi behaketak* klikatu

eta behaketa berria barneratuko dugu. Hau egiteko, barneratuko ditugun behaketa kopurua (adibidean behaketa bat) zein den adierazi behar dugu:

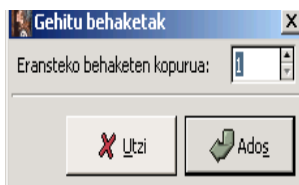
4.6 Irudia: Balioak editatzea



4.7 Irudia: Behaketak erantsi



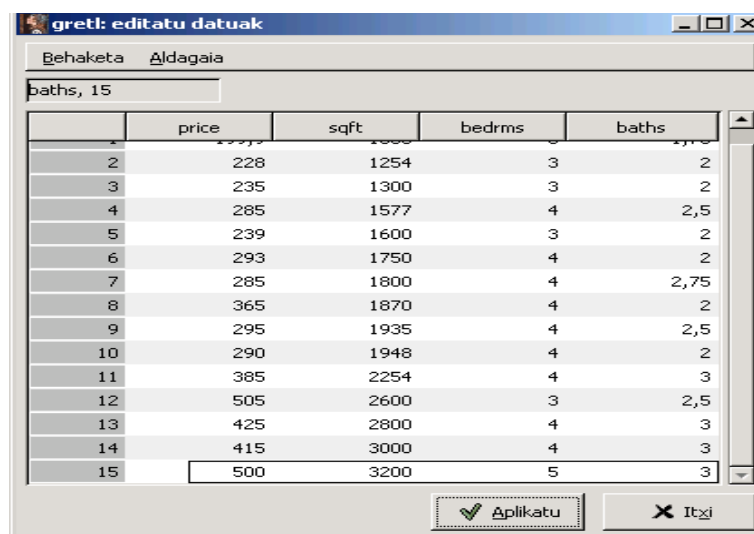
4.8 Irudia: Eransteko behaketen kopurua



eta bakoitzari dagokion lekuan ezagutzen ditugun balioak barneratzen ditudu azken errenkada osatuz. Jarraian, *Aplikatu* klikatuz.

Ondoren, kontu handia izan behar dugu, gure eredua estimatzerakoan, behaketa berri hau ez baitugu kontuan hartu behar (lagina 14 behaketez osaturik zegoen). Horregatik, *Lagina* →

4.9 Irudia: Datu berrien ikuspena



Behaketa: Aldagaia

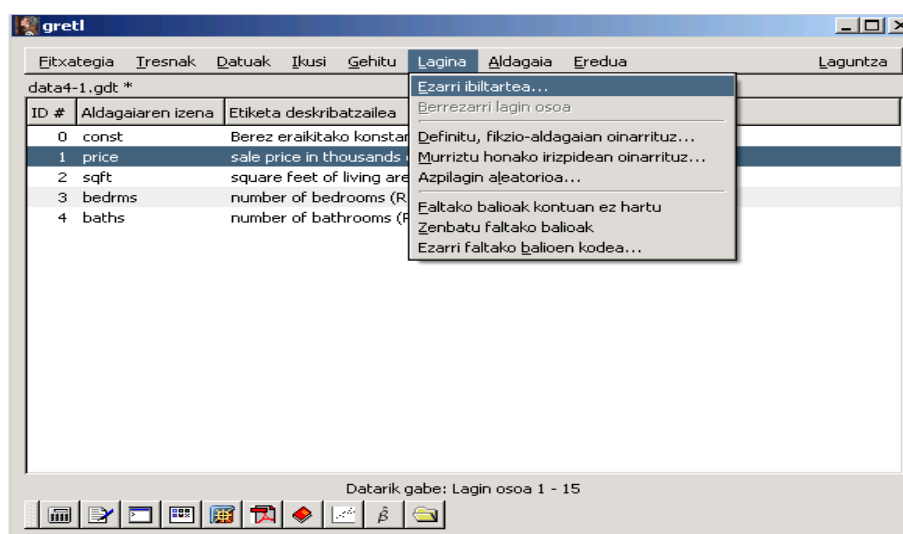
baths, 15

	price	sqft	bedrms	baths
2	228	1254	3	2
3	235	1300	3	2
4	285	1577	4	2,5
5	239	1600	3	2
6	293	1750	4	2
7	285	1800	4	2,75
8	365	1870	4	2
9	295	1935	4	2,5
10	290	1948	4	2
11	385	2254	4	3
12	505	2600	3	2,5
13	425	2800	4	3
14	415	3000	4	3
15	500	3200	5	3

Aplikatu Itxi

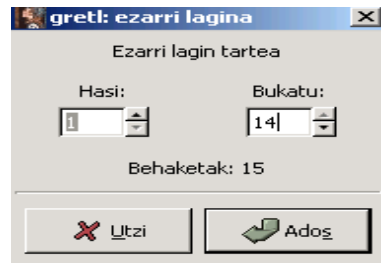
Ezarri ibiltartea aukeran

4.10 Irudia: Ibiltartearen ezarpena



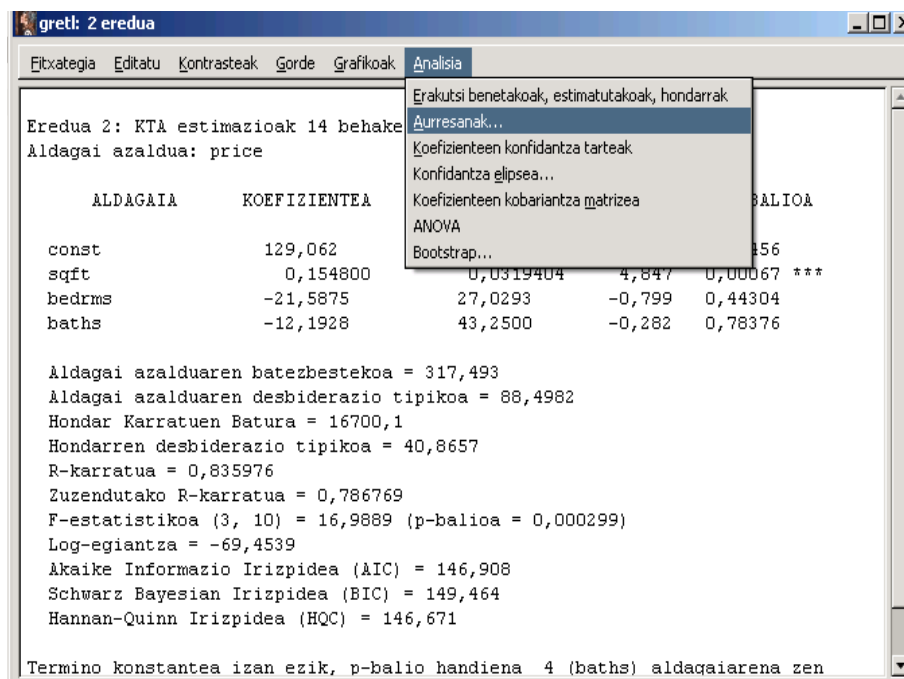
lagina nola osatzen den adieraziko dugu, gure kasuan 14 behaketez:

4.11 Irudia: Lagin tartearen ezarpena



Ados klikatu ondoren, aurreko gaietan ikusi bezala, eredua KTA bitartez estimatuko dugu eta emaitzen lehiatilan *Analisia* → *Aurreanak* klikatuko.

4.12 Irudia: Aurreanak



Lehiatila berrian 15. behaketa sartuko dugu, ondoren *Ados* klikatuz.

4.13 Irudia: Aurresan tartearen ezarpena

gretl: aurresanak

Hasi Bukatu

Aurresan tartea: 15 15

☐ aurreikuspen automatikoa
☐ aurresan dinamikoa
☒ aurresan estatikoa
☐ atzerakada bateko aurresanak

Marrazteko aurreikusitako behaketa kopurua 6

Lehiatila berrian, interesatzen zaizkigun balioak azken errenkadan daude. Hau da, ezaugarri horietako etxebizitza baten prezioaren puntuzko aurresana 479,91 da edota %95eko konfidantza mailarekin 356,49 eta 603,32 bitartean egongo denez, 500 prezioa arrazoizkoa dela deritzogu.

4.14 Irudia: Puntuzko eta tartezko aurresanen balioak

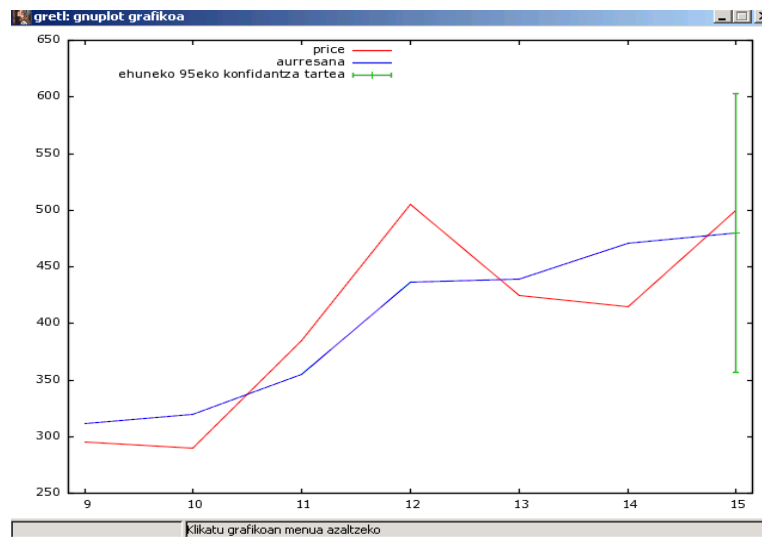
gretl: aurresanak

%95eko konfidantza tarteentzat, $t(10, .025) = 2,228$

Obs	price	aurresana	desb. tipikoa	%95eko konfidantza tartea
9	295,00	311,77		
10	290,00	319,88		
11	385,00	355,05		
12	505,00	436,30		
13	425,00	439,57		
14	415,00	470,53		
15	500,00	479,91	55,390	356,49 - 603,32

Aurreko emaitzarekin batera beste irudi bat agertzen da. Bertan orain arte lortutakoa laburbiltzen da, halaber, interesatzen zaigun aldagaiaren balioaren garapena (gorriz), aurreanak (urdinez) eta %95eko konfidantza mailako konfidantza tarte (berdez). Askotan lagungarria izaten da emaitzak horrela ikustea, zenbakiz osatutako taula batean baino hobeto nabari direlako aztergai dugun aldagaiaren ezaugarri nagusienak.

4.15 Irudia: Aurreanen grafikoa



Bibliografia

Ramanathan, R. (2002), *Introductory Econometrics with Applications*, 5. ed., South-Western, Ohio.

5 Gaia

Zehazpen errorea

Aurkibidea

5.1	Sarrera	88
5.2	Aldagai nabarien omisioaren ondorioak	88
5.3	Aldagai ez nabarien barnerapenaren ondorioak	94

5.1 Sarrera

Erregresio linealeko eredu orokor bat lehendabiziz zehazterakoan erabaki askoren menpean dago:

- Aldagai azalduaren aukeraketa.
- Aldagai azaltzaileen aukeraketa.
- Aldagaien neurketa.
- Erlazioaren forma funtzionala. Egonkortasuna.
- Perturbazioaren ezaugarriak.

Aurreko gaitan eredu zehazterakoan oinarritzko hipotesi batzuk betetzen zirela suposatzen genuen. Errealitatean hipotesi hauetatik batzuk ez betetzearen arrazoia, okerreko erabakiak hartzea edota aztertzen ari garen aldagaien edo datuen ezaugarri bereziengatik ematen da. Egite honek, erabiltzen ari garen estimatzailearen propietateengain eragin negatiboak ditu, baita inferentzia ere eta ondorioz, kontraste hauetatik ateratako emaitzak eta konklusioak ez dira fidagarriak, oker daude.

Gehienetan, eredu baten ebaluazioaren fidagarritasuna lehen zehazpen honen menpean dago. Horrela bada, azterketa grafikoak eta zehazpen kontrasteak ezinbesteko tresnak izaten dira oinarritzko hipotesi guztiak betetzen direla ziurtatzeko. Tresna hauetariko batzuk gai honetan ikusiko dira eta beste batzuk hurrengoetan. Gai honetan zehazki, aldagai azaltzaileen aukeraketan okerreko erabakiak hartzeak dakartzan ondorioak aztertuko ditugu.

Zehazpen errore honen efektuak hobeto eta errazago ikusteko asmoarekin *benetako zehazpena* ezagutzen dela **suposatuko** dugu. Honela, “ondo zehaztutako benetako eredu” eta “guk gaizki zehaztu dugun ereduaren” arteko konparaketa egiterakoan, zehazpen errorearen efektuak erraz nabaritzen dira. Kotsideratuko ditugun zehazpen erroreak hauek dira:

- a) Aldagai azaltzaile nabarien omisioa. Omisioak KTAko estimatzailearengain eta banakako kontrastengain duen eragina analizatuko dugu. Hondarren grafikoa nola aztertu eta zehazpen txarraren kontrasteren bat ikusiko dugu adibide enpirikoen bitartez.
- b) Aldagai ez nabarien barnerapena. Ereduan agertu behar ez diren aldagaiak barneratzean sortzen diren ondorioak aztertuko ditugu. Aldagai bat ez nabaria den nola erabaki ikusiko dugu.

Nahiz eta errealitatean bi efektu hauek nahasturik agertzea posiblea izan, efektu biak banaka aztertuko dira.

5.2 Aldagai nabarien omisioaren ondorioak

Aurreko gaiko adibide berdinarekin¹ jarraituko dugu, hau da, etxebizitzaren salmenta prezioa (mila dolarretan) aztertu nahi da etxebizitzak duen azalera (*SQFT*), logela kopurua

¹Ramanathaneko (2002) data4-1.gdt datu-fitxategia.

(*BEDRMS*) eta komun kopuruaren (*BATHS*) menpean, konkretuki, aurreko gaian aztertutako *A* eta *D* ereduak.

Kontsidera dezagun hasiera batean etxebizitzaren salmenta prezioa finkatzeko “ondo zehazturiko benetako erdua” ondorengoa dela (*A* erdua):

$$PRICE_i = \beta_1 + \beta_2 SQFT_i + \beta_3 BEDRMS_i + \beta_4 BATHS_i + u_i \quad i = 1, \dots, N$$

non oinarritzko hipotesi guztiak betetzen diren. Guk zehaztu eta KTAn bitartez estimatzen dugun erdua berriz, honako beste hau izan da (*D* erdua):

$$PRICE_i = \beta_1 + \beta_3 BEDRMS_i + \beta_4 BATHS_i + v_i \quad i = 1, \dots, N$$

Guk proposatu eta estimatu dugun erduan azalera neurtzen duen aldagai azaltzailea (*SQFT*) omititu egin da. Aldagai hau benetan nabaria denez, $\beta_2 \neq 0$ ematen da eta guk zehazturiko ereduko perturbazioak (v_i) aldagai honen eragina bilduko du, hau da $v_i = \beta_2 SQFT_i + u_i$. Ondorioz, gure perturbazioaren batezbestekoa ateratzerakoan, $E(v_i) = \beta_2 SQFT_i \neq 0$ izango da eta hortaz, perturbazioari buruzko oinarritzko hipotesiren bat ez da beteko.

Bestalde, omisioa egiterakoan, erduan barneratutako aldagai azaltzaileen eta perturbazio berriaren (v_i) arteko kobariantzak ($kob(BEDRMS, v)$, $kob(BATHS, v)$) ez dira zertan zero izan behar. Izatez kobariantza hauen balioak, barneko aldagai azaltzaileen eta omititutako aldagai azaltzailearen arteko kobariantzaren ($kob(BEDRMS, SQFT)$, $kob(BATHS, SQFT)$) menpekoak dira. Azken kobariantza hauek zero ez badira, orduan KTAko estimatzaileak alboratuak izango dira ($E(\hat{\beta}_i) \neq \beta_i$ $i = 3, 4$) eta alborapena kobariantza hauen menpekoea izango da. Honela, omisioa daukagun *D* erdua KTAn bidez estimatzerakoan, estimatzaileek izango duten alborapena honakoa da:

$$E(\hat{\beta}_3) - \beta_3 = \beta_2 \frac{S_{23}S_{44} - S_{24}S_{34}}{S_{33}S_{44} - S_{34}^2} \quad E(\hat{\beta}_4) - \beta_4 = \beta_2 \frac{S_{24}S_{33} - S_{23}S_{34}}{S_{33}S_{44} - S_{34}^2} \quad (5.1)$$

non $S_{js} = N^{-1} \sum_i (X_{ji} - \bar{X}_j)(X_{si} - \bar{X}_s)$, X_j eta X_s bi aldagaien arteko lagin kobariantza den baldin eta $j \neq s$ bada eta lagin bariantza $j = s$ bada. Alborapenaren zeinua jakiteko, koefizientearen zeinua eta baita kobariantzen zeinua ere zeintzuk diren begiratu behar dira.

Estimatzaile bien alborapenaren adierazpenak begiratzuz, $S_{23} = kob(BEDRMS, SQFT)$ eta $S_{24} = kob(BATHS, SQFT)$ kobariantzen menpean daudela ikus daiteke. Kobariantza hauek zero balira (edo gauza bera dena aldagaiak koerlatu gabeak izango balira, hau da, $koer(BEDRMS, SQFT) = 0$ eta $koer(BATHS, SQFT) = 0$) orduan alborapena anulatzen da eta estimatzaile alboragabeak lortuko lirake. Baina baldintza hau, logelen eta komun tamaina eta etxebizitzaren azalera koerlatu gabeak izatea, ez da errealitatean oso posiblea.

Bestalde, termino konstantearen koefizientearen alborapena aztertzean:

$$E(\hat{\beta}_1) - \beta_1 = \beta_2 \left(\bar{X}_2 - \frac{S_{23}S_{44} - S_{24}S_{34}}{S_{33}S_{44} - S_{34}^2} \bar{X}_3 - \frac{S_{24}S_{33} - S_{23}S_{34}}{S_{33}S_{44} - S_{34}^2} \bar{X}_4 \right) \quad (5.2)$$

estimatzaile hau ere alboratua dela ikus daiteke. Izatez, nahiz eta beste biak alboragabeak izan, hau da, $S_{23} = S_{24} = 0$ eman, $\hat{\beta}_1$ estimatzaileak alboratua izaten jarraitzen du, omititutako aldagai azaltzailearen lagin batezbestekoaren ($\bar{X}_2$) menpekoea baita eta batezbesteko hau ez baita zertan zero izan behar. Horrela bada, omititutako aldagai azaltzailearen efektu batzuk termino konstantera pasatzen dira nahiz eta malden estimazioan arazorik ez izan. Arrazoi honegatik eta teoria ekonomiko konkretu batek termino konstantearen existentzia ezinezkoa dela ez badu baieztatzen, ereduetan beti jartzen da konstante bat.

Dena den, koefizienteen estimatzailea, batez ere maldena, alboratua baldin bada, estimatzaile alboratu hauetan oinarrituriko banakako eta baterako esanguratasun kontrasteak eta koefizientei buruzko beste edozein kontraste oker daude. Gainera, nahiz eta malden estimatzaileak alboragabeak izan koerlazio ezaren aurrean egotean, kontrasteak ez lirakeke baliagarriak izango, kontrasteetan erabiltzen diren desbideratze edo bariantza estimatuak ez baitira alboragabeak. Bariantza eta kobariantza matrizea ($\hat{\sigma}^2(X'X)^{-1}$) alboratuta estimatzearen arrazoia $\hat{\sigma}^2$ estimatzailean datza. Ereduan omisioa dagoenean estimatzaile hau alboratua da, hondar karratuen batura alboratuta egongo baita. Laburtuz, aldagai azaltzaileen omisioaren ondorioak latzak dira, inferentzian bereziki.

Nola hauteman aldagai nabariaren omisioa?

Aurreko gaietan sakondu dugun etxebizitzaren adibidearekin jarraituz, beste azterketa xume bat egingo dugu, konkretuki, lehen aipatutako A eta D ereduarekin. Aurreko gaian ikusi genuenez, *SQFT* aldagai nabariaren omisioak doikuntza (R^2 eta $\bar{R}^2$) txarragoa dakar eta errore neurriak (*AIC* eta *BIC*) handitu egiten dira. Bestalde, omisioaren aurrean, koefiziente estimatuen balioak eta esanguratasunak aldatzen dira. Adibide honetan, *BATHS* aldagaiaren koefizientearen balioa positiboa bihurtu da eta gainera orain aldagai hau nabaria bezala agertzen da. Dirudienez, *SQFT* aldagaia kentzerakoan, gelditzen diren aldagai bietatik *BATHS* aldagaiak *BEDRMS* aldagaiak baino zerikusi handiagoa du azalerarekin.

Bestalde, omititutako aldagaiak barneratutako aldagaiekiko duen koerlazioa aztertzeko, *SQFT*, *BEDRMS* eta *BATHS* aldagaiak azpimarratu ondoren, $Ikusi \rightarrow Koerlazio\ matrizea$ aukerarekin irtengo den emaitza honakoa da:

5.1 Taula: **A ereduko** aldagai azaltzaileen koerlazio matrizea

Koefizienteen koerlazioa, 1–14 behaketekin
% 5 esangura-maila (alde-bikoa) = 0,5324 eta $n = 14$

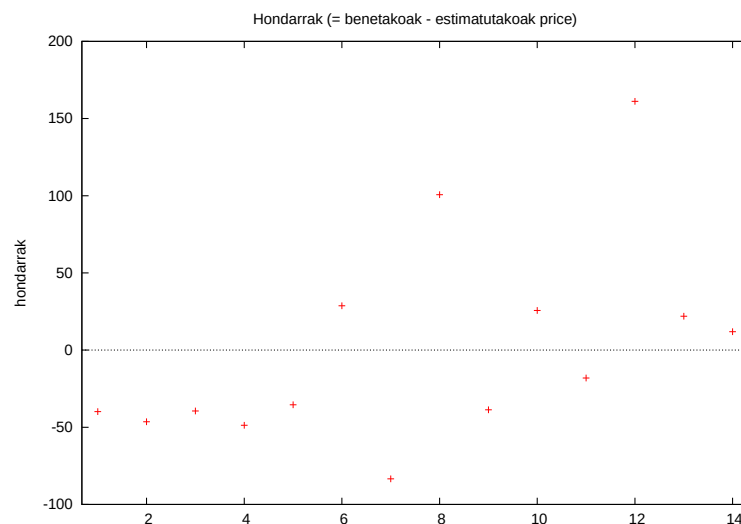
sqft	bedrms	baths	
1,0000	0,4647	0,7873	sqft
	1,0000	0,5323	bedrms
		1,0000	baths

Koerlazio hauek begiratzean, bai *BEDRMS* aldagaiak eta bai *BATHS* aldagaiak koerlazio positiboa daukate *SQFT* aldagaiekin. Hori espero genuen. Baina aldagai bietatik *BATHS* aldagaiak du koerlazioarik handiena *SQFT* aldagaiekin. Logikoa al da? Bai. Demagun bi etxebizitzaren artean aukeratu behar duzuela. Batak hiru logela eta komun bat ditu eta besteak logela bat eta hiru komun. Zein nahiago duzue? Gaur egun, etxebizitza batean

logela bat gehiago izatea, nahiz eta logelen tamaina txikia izan, nahiko normala da. Baina komun kopurua handitzen bada, ziur etxebizitzaren azalera handitu egiten dela, normalean logela edo komun baten artean nahiago izaten baita logela bat gehiago izatea. Gainera, *BEDRMS* aldagaiak duen koerlazioa *BATHS* aldagaiarekin *SQFT* aldagaiarekin duena baino handiago da. Zergatik? Logelen kopurua handitzean (normalean etxebizitzan pertsona gehiago daudelako) komun falta areagotu egiten da. Honela, logela gehiago izatean, komun gehiago behar dira eta etxebizitzaren azalera handitu beharko litzateke, hau da, *BEDRMS* aldagaiak etxebizitzaren salmenta prezioarengain duen eragina *BATHS* aldagaiaren bitartez biltzen dela dirudi, horrela *BEDRMS* aldagaia banaka ez esanguratsua irteten da.

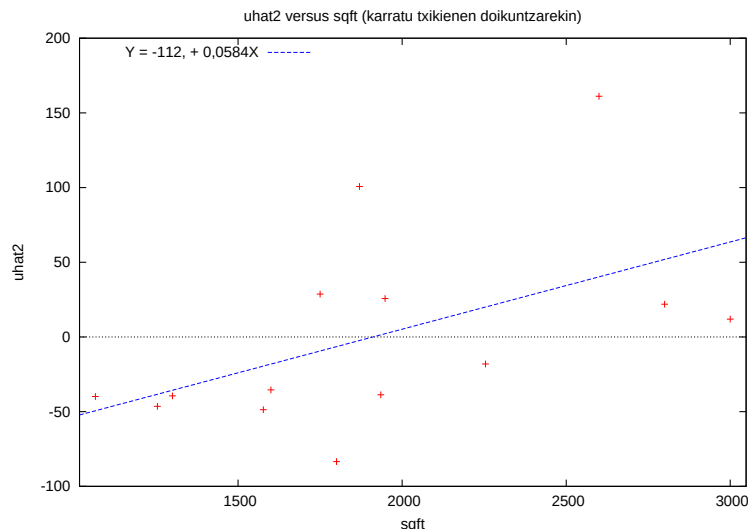
Hondarrak formalki analizatzeaz gain, hondarren behaketen grafikoan edo aldagai azaltzaile batekiko grafikoan joera sistematikoren baten existentziak, omititutako aldagai azaltzailearen bat faltan dagoela susmaerazi ahal digu. Jarraian, *SQFT* aldagaia omititu izan den D ereduaren hondarrak aztertuko ditugu. Horretarako KTAko estimazio lehiatilan: *Grafikoak* $\rightarrow$ *Hondarren grafiko* $\rightarrow$ *Behaketa zenbakiarekiko* eginez ateratzen zaigun grafiko aztertuko dugu. Grafikoan zera ikus daiteke: laginaren hasieran hondar negatiboen multzo bat dagoela eta lagina aurreratzen den heinean balio hauek etengabe handituz doazela, joera positibo bat izanez. Datuak nola ordenatuta dauden ikusirik, hau da azalera txikitik handira doazela ikusirik (ikusi *SQFT*ren balioak), azalera txikiko etxebizitzaren hondarrak dira negatiboak direnak eta azalera handitzean doan heinean, hondarrak gero eta handiagoak agertzen dira. Beraz, hondarrak ez dira zero inguruan aleatorioki banatzen, baizik eta patroi konkretu batekin: joera gorakor bat azaltzen dute.

5.1 Irudia: D Ereduko hondarren behaketen grafiko



SQFT aldagai azaltzailearekiko hondarren grafikoa lortzeko egin behar den lehenengo pausua hondar hauek gordetzea izango da. KTAko estimazio lehiatilan *Gorde* $\rightarrow$ *Hondarrak* egin eta hondarrak *uhat2* bezala gordeko ditugu. Ondoren, hasierako lehiatilan, *Ikusi* $\rightarrow$ *Grafikoak* $\rightarrow$ *X-Y Grafikoa* eginez hondarrak eta *SQFT* aldagaien arteko grafikoa ateratzen dugu, non *X*-ardatzaren aldagaia aukeran, *SQFT* aldagaia jarriko dugun eta *Y*-ardatzaren aldagaian, *D* ereduko KTAko hondarrak. Lortzen den grafikoa ondorengo da:

5.2 Irudia: **D Ereduko** hondarren grafikoa *SQFT* aldagaiarekiko



Bertan, hondarren eta *SQFT* aldagaiaren arteko erlazio lineal positibo baten existentzia argi ikusten da. Izatez, grafiko bertan bi aldagai hauen arteko erregresio zuzen estimatua, hau da, lagin erregresio zuzena marrazten da. Erlazio positibo honen esanahia zera da: omititutako *SQFT* aldagai azaltzaileak hondarrak azaltzeko informazioa dauka.

Grafiko honetan susmatzen dena formalki kontrastatu daiteke Engle-k (1982) proposatutako kontrastearekin Kontraste honen hipotesi hutsean kanpoan utzi den aldagai azaltzaileak *SQFT*, hondarrak $\hat{u}_i$ azaltzeko informaziorik ez duela daukagu, hau da, *SQFT* ez da esanguratsua hondarrak azaltzeko ($\delta_2 = 0$). Kontraste hau aurrera eramateko ondorengo erregresio laguntzailea estimatzen da

$$\hat{u}_i = \delta_1 + \delta_2 SQFT_i + \xi_i \quad i = 1, \dots, N \quad (5.3)$$

Mugatze koefizientea (R^2) kalkulatzeko da eta bere balioa altua bada *SQFT* aldagaiak hondarrak azaltzen dituela adieraziko du eta ez baldin badaukate inolako erlazorik, orduan mugatze koefizientearen balioa txikia izango da. Kontrastea formalki egiteko erabili behar den estatistikoa $\lambda = NR^2$ da zeinak hipotesi hutsaren menpean eta laginaren tamaina handia izanik, $\chi^2_{(q)}$ banaketa jarraitzen duen eta askatasun graduak (5.3) erregresio laguntzailean sartutako aldagai azaltzaileen kopurua den (termino konstantea ezik). Kasu honetan $q = 1$ izango da *SQFT* bakarrik sartu baitugu. Erabaki araua ohikoa da, kalkulaturako estatistikoa-ren balioa tauletako balioa baino txikiagoa bada ($\lambda < \chi^2_{(q)\alpha}$) hipotesi hutsa ez da baztertzen

5.2 Taula: Engle kontrastearen emaitzak

Chi-karratua(1)

eskubiko kolako probabilitatea = 0,05

osagarritzko probabilitatea = 0,95

Balio kritikoa = 3,84146

α esangura-mailarentzat eta kasu honetan *SQFT* aldagaiak hondarrekin ez duela inolako erlaziorik baieztatuko dugu. Lagin handietako kontrastea denez, behaketen kopurua handitzen doan neurrian, kontrastearen fidagarritasuna ere handitzen joango da.

Gure adibidean erregresio laguntzailearen emaitzak honakoak dira:

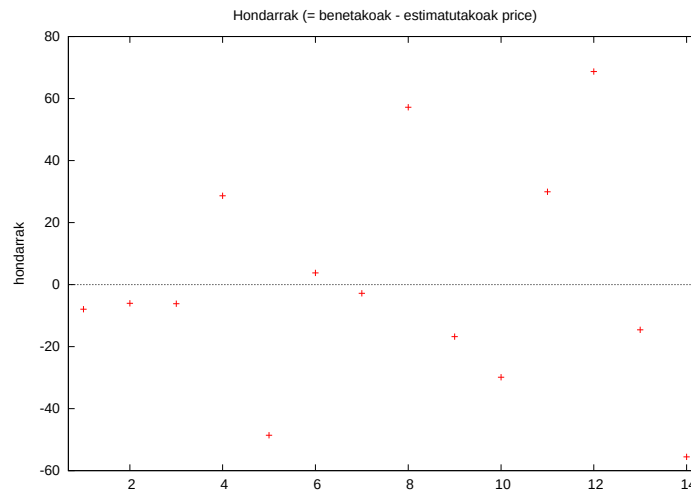
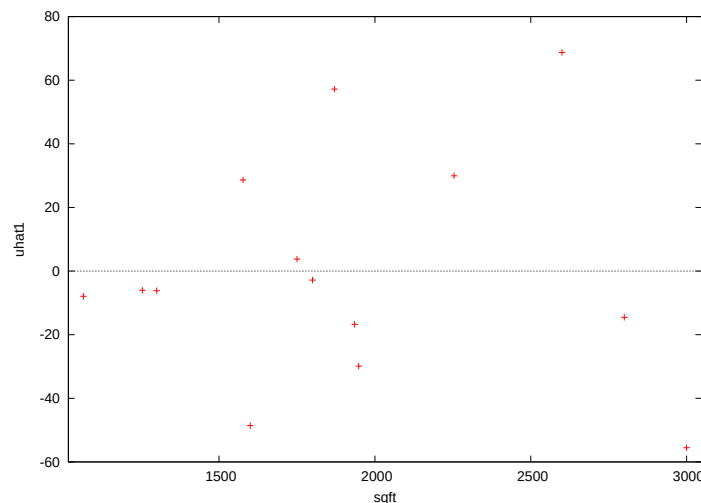
$$\begin{array}{c} \hat{u}_i = -111,588 + 0,0583946 \text{ SQFT}_i \\ (t_{est.}) \quad (-1,995) \quad (2,078) \\ N = 14 \quad R^2 = 0,264584 \end{array}$$

beraz estatistikoaren balioa $\lambda = NR^2 = 3,70417$ da. Balio hau gorde nahi izanez gero, erregresio honetako KTAko estimazioaren lehiatilan *Gorde* ataleko T^*R -karratua aukeratzen dugu eta jarraian, kontrastea burutzeko tauletako balioa libururen batean bilatzea ez badugu nahi, Gretl programaren barneko *Tresnak* aukeran *Estatistika-taulak* atalean *chi-karratua* aukeratzen dugu, askatasun graduak ($q = 1$) zeintzuk diren *ag* kaxan adierazi eta nahi dugun esangura-maila ere ($\alpha = 0.05$) jartzerakoan, ondorengo emaitza irteten da:

Aukeratutako $\alpha = \%5$ eko esangura-mailarekin, estatistikoa tauletako balioa baino txikiagoa da baina justu-justuan eta $\alpha = \%10$ eko esangura-maila hartu izan bagenu ordea, hipotesi hutsa baztertu egiten da. Guzti hau dela eta, badirudi *SQFT* aldagaiak zerbait badaukala hondarrei buruz esateko.

Emaitza hauek aztertuz, kontrastea $\alpha = 0,05$ eko esangura-mailarekin burutzean pantailan agertzen den p-balioa baino txikiagoa denez hipotesi hutsa ez da baztertzen **baina** $\alpha = 0,0542767$ koa baino apur bat handiagoa hartzen badugu (adibidez $0,0542767 + 0,0000001$) hipotesi hutsa baztertuko litzateke. Horregatik, lehen esan den bezala, badirudi *SQFT* aldagaia badela hondarrak azaltzeko gai. Hau horrela bada aldagai azaltzaile esanguratsua dela baieztatu dezakegu eta ereduaren barneratu beharko litzateke.

Nola aldatuko lirateke lorturiko emaitzak zuzeneko ereduko hondarrak aztertuko bagenu? Ondorengo (5.3) irudiko grafikoan A eredu zuzenaren hondarrak erakusten dira eta hondarren portaera aleatorioa dela ikusi daiteke. Eta *SQFT* aldagaiarekiko hondarren grafikoaren itxurara? Lehen azaldutako pausuak jarraituz, (5.4) irudiko grafikoaren lortuko genuke eta joera edo patroia argi bat ez dela ikusten esango genuke, honela *SQFT* aldagaia hondarrekin erlazionatu gabe dagoela esanez.

5.3 Irudia: **A Ereduko** hondarren behaketen grafikoa5.4 Irudia: **A Ereduko** hondarren grafikoa SQFT aldagaiarekiko

5.3 Aldagai ez nabarien barnerapenaren ondorioak

Demagun oraingoan “ondo zehaztuta dagoen benetako eredua” honakoa dela (E eredua):

$$PRICE_i = \beta_1 + \beta_2 SQFT_i + u_i \quad i = 1, \dots, N$$

non oinarritzko hipotesi guztiak betetzen diren. Baina guk zehaztu eta estimatzen dugun ereduan berriz $BEDRMS$ aldagai ez nabaria barneratu da (F eredua):

$$PRICE_i = \beta_1 + \beta_2 SQFT_i + \beta_3 BEDRMS_i + u_i \quad i = 1, \dots, N$$

Azken eredu honetan $BEDRMS$ aldagaiaren koefizientearen benetako balioa zero dela dakigu. Baina eredua estimatzerakoan, aldagai ez nabariaren koefizientea ere estimatzen da

eta bere puntuzko estimazioa ($\hat{\beta}_3$) ez da zertan zero irten behar. Zehazpen errore honen ondorioak zeintzuk dira?

- Eredu honetako estimatzaile guztiak alboragabeak direla froga daiteke: $E(\hat{\beta}_j) = \beta_j \forall j$. Adibidean, $E(\hat{\beta}_3) = 0$ da. Honela, perturbazioen batezbestekoak zero izaten jarraituko du.
- Estimatzailen bariantza eta kobariantza matrizearen ohiko estimatzailea alboragabea izango da, $\hat{\sigma}^2$ estimatzailea alboragabea baita. Honela bada, konfidantza tarteak eta ohiko kontrasterako estatistikoak fidagarriak izango dira.
- Zehazpen errore honen koste edo galera askatasun graduak galtzean datza eta baita estimazioen efizientzia ere. Guk zehaztu dugun ereduaren koefizienteen estimatzaileen bariantzak benetakoarenak baino handiagoak dira. Eta diferentzia hau gero eta handiagoa izango da barneratutako aldagai azaltzaile ez nabaria eta barneratutako aldagai azaltzaile nabariaren arteko koerlazioa handitzen doan heinean.

Gure adibidean, *BEDRMS* aldagai ez nabaria sartzerakoan galtzen den efizientzia handia izango da baldin eta aldagai ez nabari honen (*BEDRMS*) eta eredu aldagai azaltzaile nabariaren (*SQFT*) arteko koerlazioa handia bada. 3. gaian erakutsitako erregresioen emaitzetan, gaizki zehaztuta dagoen ereduaren (*F*) desbideratzeak eredu zuzenekoak (*E*) baino handiagoak direla, ikus daiteke. Bestalde, *SQFT* aldagaiari dagokion koefizientearen estimatzailearen bi bariantzen arteko zatidura edo ratioa begiratzean:

$$\frac{Bar(\hat{\beta}_2)^F \text{ eredu}}{Bar(\hat{\beta}_2)^E \text{ eredu}} = \frac{1}{1 - koer(SQFT, BEDRMS)^2} = 1,4 > 1$$

bat baino handiagoa denez efizientzia galdu egiten da. Hortaz, *BEDRMS* aldagai azaltzaile ez nabariaren barnerapenak β_2 koefizientea zehaztasun gutxiagorekin estimatzea erakartzen du eta berdina gertatzen da β_1 koefizientearen estimazioan.

Nola hauteman aldagai ez nabariaren presentzia?

Jarraibiderik hoberena eredu orokor batetik hasi eta aldagaien esanguratasuna analizatzea da. Honetarako zuzendutako mugatze koefizientea, *AIC* eta *BIC* neurriak erabili daitezke. Jarraibide honek desabantailak izan ditzake: barneratutako aldagaien kopurua handia bada askatasun graduen ($N - K$) galera izaten da eta askoz txarragoa dena, barneratuta dauden aldagaien arteko koerlazioarengatik efizientzia galdu egiten da, hau da, bariantza handia egotearen probabilitatea handia izan daiteke. Eta bariantza hauek inferentziaren emaitzetan eragina dutenez, baliteke aldagai azaltzaile bakoitzaren banakako esanguratasuna ondo bereiztea ezinezkoa izatea².

Alderantzizko bidea jarraitzen badugu eta eredu bakun batetik orokorrera joaten bagara zenbait arazo ager daitezke. Arazo hauek “*data mining*” (*datu-ustiaketa*) bezala ezagutzen dira. Prozedura honen pausuak honakoak dira: eredu bakunean aldagai bat gehiago sartu eta esanguratsua baldin bada barnean utzi eta bestela kendu, beste bat sartu eta analisi berdina,

²Aldagai azaltzaileen arteko koerlazioaren ondorioak hurrengo gaian ikusiko dira.

ondoren hurrengoa ... Hau egitean, azkenean erabili den esangura-maila ez da α izango baizik eta α bider egin diren kontrasteen kopurua izango da eta esangura-maila handitzean badakigu hipotesi hutsa benetakoa izanik, hipotesi huts hori baztertearen probabilitatea handitzen doala. Bestaldetik, eredu bakunetik hasterakoan, baliteke nabaria den aldagaiaren bat oraindik barneratu gabe egotea eta ondorioz, aldagai nabari hori sartu arte egindako kontraste guztiak ez dira fidagarriak izango, omisioaren aurrean egongo ginateke eta.

Bibliografia

Engle, R. F. (1982), "A general approach to Lagrangian Multiplier Modelo Diagnostics", *Journal of Econometrics*, 20, 83-104 orr.

Ramanathan, R. (2002), *Introductory Econometrics with Applications*, 5. ed., South-Western, Ohio.

6 Gaia

Kolinealitate anizkoitza

Aurkibidea

6.1	Sarrera	98
6.2	Kolinealitate anizkoitz zehatza	98
6.3	Kolinealitate anizkoitz altua	100

6.1 Sarrera

Orokorrean, erregresio eredu bat estimatzeko eskuragarri diren datuak eta ereduaren oinarrizko teoria ekonomikoaren arteko baterakuntza ez da beti posible izaten, laginean datu batzuk falta direlako, datuak taldekatuta daudelako, neurketa erroreak dituztelako edota aztergai ditugun aldagai azaltzaileen arteko koerlazioa handiegia delako.

Atal honetan, azken arazo hau aztertuko dugu, hau da, kolinealitate anizkoitza edo aldagai azaltzaileen arteko **lagin** koerlazioa. Erregresio ereduko koefizienteen interpretazioa egi-terakoan, konkretuki malden interpretazioa, X_{ji} ($j > 2$) unitate batean handitzean aldagai azalduaren (Y_i) batezbesteko gehikuntza β_j unitatekoa dela aipatzen genuen, **beste aldagai azaltzaile guztiak konstante mantenduz**. Kolinealitate anizkoitz zehatza dugunean ordea, ezinezkoa da gainontzeko aldagai azaltzaile guztiak konstante mantentzea, beraien arteko erlazio lineala baitago. Kasu hauetan, ondoren ikusiko dugun bezala, ezinezkoa izaten da, aldagai bakoitzaren eragin isolatua bereiztea.

Bereziki bi kasu analizatuko ditugu: kolinealitate anizkoitz zehatza eta altua.

6.2 Kolinealitate anizkoitz zehatza

Ereduaren zehazpen zuzena emanik, aldagai azaltzaileetarikoa bat besteen konbinazio lineal zehatz bezala adierazi badaiteke, orduan kolinealitate anizkoitz zehatza dugula esango dugu. Ikusiko dugunez, ereduaren zehazpena zuzena izan arren, bere estimazioan eragina izango du.

Izan bedi hurrengo erregresio eredu orokorra:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_i X_{Ki} + u_i \quad i = 1, \dots, N$$

non bi aldagai azaltzaileen artean konbinazio lineal zehatza ematen den, adibidez $X_{3i} = cX_{2i}$, c konstante ezaguna izanik. Informazio honekin X datu matrizea osatuko bagenu, hirugarren zutabeko balioak bigarren zutabekoak bider c eginez lortuko genituzke eta ondorioz, X matrizea ez litzateke zutabeetan hein osokoa izango ($h(X)=K-1$), hau da ez litzuzke K zutabe independente izango. Hondar Karratuen Batura (HKB) minimo eginez lortzen diren ekuazio normalen sistema ($X'Y = X'X\beta$) ezin izango litzateke era bakar batean askatu, $X'X$ matrizearen determinantea zero izango baita eta hortaz, bere alderantzizkoa ezin izango litzateke lortu. Ondorioz, ereduko koefizienteak banaka estimatzea $\hat{\beta} = (X'X)^{-1}X'Y$ erabiliz ez litzateke posible izango.

Koefizienteen interpretazioaren ikuspuntutik ere arazoa aztertu dezakegu. Alde batetik, lehen bezala, termino independentea, aldagai azalduaren batezbestekoa izango da aldagai azaltzaile guztiak zero direnean. Bestetik ordea, β_2 -ren kasuan, $E(Y)$ -ren aldakuntza izango da X_2 unitate batean handitzean eta X_3 konstante mantenduz, baina kolinealitate zehatza dugunean eta gure adibidearekin jarraituz, ezin daiteke X_3 konstante mantendu X_2 unitate batean handitzean, zeren eta $X_{3i} = cX_{2i}$ baita.

Ikusi dugunez, ezin izango litzateke koefiziente guztiak estimatu banan-banan, baina bai ordea beraien konbinazio linealen bat. Gure adibidearekin jarraituz, $X_{3i} = cX_{2i}$ (c =konstantea)

ereduan barneratzen dugunean eta ordenatzerakoan ondorengo eredua lortuko genuke:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 (cX_{2i}) + \dots + \beta_K X_{Ki} + u_i$$

$$Y_i = \beta_1 + (\beta_2 + c\beta_3)X_{2i} + \dots + \beta_K X_{Ki} + u_i$$

$$Y_i = \beta_1 + \gamma X_{2i} + \dots + \beta_K X_{Ki} + u_i$$

Lortutako eredu honetan ez dugu kolinealitate arazorik eta bertako koefiziente guztiak interpretagarriak dira eta estimagarriak KTA bitartez. Eredu honetako koefiziente guztiak estimatu daitezke eta horien artean γ , hau da, hasierako ereduko koefizienteen konbinazio lineal hau, baina ez ordea β_2 eta β_3 bi koefizienteak banaka. Koefiziente guztiak estimatzeko informazio independente nahiko ez dago, zeren eta **lagina izanik**, X_3 aldagaiak ez baitu X_2 aldagaiak eskeintzen duen informazioaren gehigarri esanguratsurik ematen.

Nola estimatzen du Gretlek kolinealitate anizkoitz zehatza duen eredu bat?

Demagun Ramanathaneko (2002) data4-1 fitxategian SQFT aldagaiaren konbinazio lineala den aldagai berri bat sortzen dugula, $SQFT1 = 5 \times SQFT$ eta hurrengo eredua zehazten dugula:

$$PRICE_i = \beta_1 + \beta_2 SQFT_i + \beta_3 BEDRMS_i + \beta_4 SQFT1 + u_i$$

Kolinealitate anizkoitz zehatza dagoenean, Gretlek kolinealitatea sortarazten duen aldagaie-tariko bat omititzen du, hurrengo emaitza erakutsiz:

Eredua 1: KTA estimazioak 14 behaketak erabiliz 1-14

Aldagai azaldua: price

Omitituta kolinealitate zehatzagatik sqft1

ALDAGAIA	KOEFIZIENTEA	DESB.TIP	T ESTAT	P-BALIOA
const	121,179	80,1778	1,511	0,15888
sqft	0,148314	0,0212080	6,993	0,00002 ***
bedrms	-23,9106	24,6419	-0,970	0,35274

Aldagai azalduaren batezbestekoa = 317,493

Aldagai azalduaren desbiderazio tipikoa = 88,4982

Hondar Karratuen Batura = 16832,8

Hondarren desbiderazio tipikoa = 39,1185

R-karratua = 0,834673

Zuzendutako R-karratua = 0,804613

F-estatistikoa (2, 11) = 27,7674 (p-balioa = 5,02e-005)

Log-egiantza = -69,5093

Akaike Informazio Irizpidea (AIC) = 145,019

Schwarz Bayesian Irizpidea (BIC) = 146,936

Hannan-Quinn Irizpidea (HQC) = 144,841

Hau da, Gretlek ohartarazten du erregresioan aldagai bat omititu duela, gure adibidean SQFT1, eta aldagaia ez duen ereduaren estimazioaren emaitzak ematen ditu. Teorikoki, $SQFT1 = 5 \times SQFT$ ereduan barneratu ondoren lortzen den eredua honakoa da:

$$PRICE_i = \beta_1 + (\beta_2 + 5\beta_4)SQFT_i + \beta_3BEDRMS_i + u_i$$

non β_1 eta β_3 banaka estimatu daitezkeen baina β_2 eta β_4 ez. Beraz, estimazio emaitzetan, SQFT-ren koefiziente estimatua (0,148314) $(\beta_2 + 5\beta_4)$ konbinazioari dagokio.

6.3 Kolinealitate anizkoitz altua

Errealitatean, aldagai ekonomikoen artean nolabaiteko erlazioa izaten da, nahiz eta erlazio hori zehatza izatea ohikoa ez izan. Oso garrantzitsua da erlazio hori zein gradukoa den aztertzea, oso sakona baldin bada, hau da, kolinealitate anizkoitz altua baldin badago, ondorioak izango baititu adibidez eredu koefizienteen zehaztasunean eta kontrasteetan. Hala ere, kontuan izan behar da askotan datuetan ematen den arazoa izaten dela, hau da, aukeratutako laginekoa eta ez ereduakoa.

Aldagai azaltzaile batzuren arteko erlazioa sakona baldin bada baina ez zehatza, X datu matrizea hein osokoa izango da zutabeetan. Hau da, ereduan K koefiziente baldin baditugu, $h(X) = K$ izango da eta hortaz, $X'X$ matrizearen determinante zeroren desberdina izango denez, koefiziente guztiak estimagarriak dira.

Aurreko adibidean oinarrituz, kolinealitate anizkoitz altua hurrengo moduan ikusi daiteke:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

$$X_{3i} \approx cX_{2i}$$

non X_2 eta X_3 aldagaien arteko erlazioa zehatza ez den. Aldagai azaltzaileen artean kolinealitate altua denean, ohikoa izaten da ondorengo **arazoak** izatea:

- Datuetan aldaketa txiki batzuk, koefizienteen estimazioetan aldaketa handiak sortaraztea.
- Koefizienteen estimazioek espero den zeinua ez izatea edota magnitude sinesgaitza.
- Koefizienteen estimatzaileek bariantza handia izaten dute, hau da, zehaztasun txikiarekin estimatuta daude. Ondorioz, aldagaien banakako esanguratasun kontrastea egitean, estatistikoaren balioa oso txikia izango da eta ondorioz, hipotesi hutsa ez baztertzearen probabilitatea izugarri handitzen denez, aldagaia banaka nabaria ez dela ondorioztatuko dugu. Hau da, *gezurrezkoak* diren hipotesi hutsak ez baztertzearen probabilitatea handiagoa izatea eragingo du. Ondorioz, koefiziente baten konfidantza tartea lortuz gero, tartea oso zabala izango da.
- Eredua estimagarria da eta erregresioaren mugatze koefizientea altua izaten da. Horrela, aldagai azaltzaileen baterako esanguratasun kontrastea burutuz gero, hipotesi hutsa ez da baztertzen eta aldagaiak batera nabariak direla ondorioztatzen da, baina lehen ikusi dugunez, banaka ez esanguratsuak direla irtengo da.

Kolinealitatearen arazo nagusienak hauek izanik, emaitza horiek izatea lagungarri izan daiteke kolinealitatea dagoen edo ez susmatzeko. Ondorio edo arazo horien arrazoitariko bat $X'X$ matrizearen determinantean datza. Aldagaien arteko erlazioa sakona izan arren, ez da zehatza eta beraz, X matrizearen zutabeen artean ez dagoenez konbinazio lineal zehatzik, determinantea ez da zero izaten. Hala ere, kolinealitate altuagatik, determinantea zerotik oso hurbileko balio bat izango da.

Dena den, kontu handia izan behar da, gertagarria delako aldagaien neurri unitateagatik determinantea ia zero izatea eta ez kolinealitate arazoagatik. Bestalde, posible da aldagaien banakako esanguratasun kontraste batean hipotesi hutsa ez baztertzea, benetan aldagai ez nabaria delako aldagai azaldua azaltzeko. Garbi izan behar dugu, egoera hauek ematen diren guztietan ez dagoela beti kolinealitate anizkoitzaren arazoa eta beraz, kontu handia izan behar da emaitzak aztertu eta interpretatzerakoan.

Kolinealitatea hautematzeko modu eragingarri bat, aldagai azaltzaile bakoitza besteekiko eginiko erregresio laguntzaileen mugatze koefizienteak erabiliz izango da. Erregresio laguntzaile hauetako R_j^2 -tariko bat edo batzuk handiak badira, erregresioan kontuan hartutako aldagaien arteko koerlazioa handia dela esan nahiko du eta beraz, aldagai hauen arteko kolinealitate anizkoitz posiblea dugula. Gure adibidean, oso lagungarria izango litzateke X_2 eta X_3 ren arteko erregresio laguntzaile bat burutzea:

$$X_{2i} = \alpha_1 + \alpha_2 X_{3i} + v_i$$

Kolinealitate altua badute, erregresio honen mugatze koefizientea oso altua izango da, erregresio lineal eredu bakuna izanik, mugatze koefizientea X_2 eta X_3 aldagaien arteko lagineko koerlazio koefiziente linealaren karratua baita ($R^2 = (r_{X_2X_3})^2$).

Beste aukera bat aldagaien arteko lagineko koerlazio koefizienteak kalkulatzean datza. Aldagaien arteko koerlazio koefizienteek -1 eta 1 bitarteko balioak har ditzakete, positiboa baldin bada erlazio zuzena dela adierazten duelarik eta negatiboa bada orde, alderantzizkoa. Bestalde, balio absolutuan bat baliotik hurbileko balioa bada, erlazioa sakona da eta zero baldin bada, koerlazio linealik ez dutela esango dugu. Aurreko gaietan ikusi bezala, etxebizitzaren prezioaren adibideko kasuan (**A eredu**a hartuz), ondorengo koerlazioak lortuko genituzke:

6.1 Taula: **A eredu**ko aldagaien koerlazio matrizea

Koerlazio Koefizienteak, 1 - 14 behaketak erabiliz

%5eko esanguratasuna (alde biko) = 0,5324 n = 14 -rentzat

price	sqft	bedrms	baths	
1,0000	0,9058	0,3156	0,6696	price
	1,0000	0,4647	0,7873	sqft
		1,0000	0,5323	bedrms
			1,0000	baths

Ikus dezakegunez, aldagai azaltzaileen arteko erlazio linealak zuzenak dira, koefiziente guztien zeinuak positiboak baitira. Erlazio sakonena, SQFT eta BATH-en artekoa litzateke 0.7873 balioarekin, zuzena eta nahiko sakona. Hala ere, BATH eta BEDRMS aldagaien koerlazio koefizientea ikusirik (0.5323) eta aurreko ikasgaian egin ditugun banakako eta baterako

esanguratasun kontrasteak gogoratu, azterketa sakonago bat egingo dugu. Ikusi dugunez, eredu honetan aldagai azaltzaileak batera nabariak direla ondorioztatu dugu eta SQFT aldagaia ezik, beste biak banaka ez nabariak direla. Kolinealitate arazoak izan ditzakegu eta horren ondorioz eman al dira emaitza hauek? Formalki aztertuko dugu Gretleko aukerak aprobetxatuz.

Badira kolinealitatea **hautemateko prozedura formalak** ere. Belsley, Kuh eta Welsch (1980) egileen prozedura adibidez, matrize baten Baldintza Zenbakian (γ) oinarritzen da, hau da, matrize horren balio propio handienaren eta txikienaren ($\lambda_{max}, \lambda_{min}$) zatiduraren erro karratuan: $\gamma = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}}$. Gure kasuan, $X'X$ matrizearen balio propioak kalkulatu genituzke eta aipaturiko Baldintza Zenbakia kalkulatu. Aldagaien arteko koerlazioa handituz gero, zenbaki hau handiagoa izango litzateke. Egile hauek diotenez, zenbaki hau 30 baino handiago denean, gradu altuko kolinealitate arazoak egon daitezke.

Gai honetan proposatu eta erabiliko duguna ordea, Neter, Wasserman eta Kutner (1990) proposaturiko Jasankortasuna (JAS) eta Bariantzaren Puztea (VIF) izango dira, Gretlek prozedura hau jarraitzeko aukera ematen baitu.

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_i X_{Ki} + u_i \quad i = 1, \dots, N.$$

Aurreko erregresio lineal eredu orokorreko koefiziente baten ($\hat{\beta}_j$) bariantza ondorengo espresioaren bitartez kalkulatu daiteke:

$$var(\hat{\beta}_j) = \frac{\sigma^2}{\sum (X_{ji} - \bar{X}_j)^2} \frac{1}{(1 - R_j^2)} = \frac{\sigma^2}{\sum (X_{ji} - \bar{X}_j)^2} VIF_j$$

non β_j , X_j aldagaiaren koefizientea den, R_j^2 koefizientea X_j aldagaia beste aldagai azaltzaile guztien funtzioan adierazitako erregresio laguntzailearen mugatze koefizientea den eta azkenik, $VIF_j = \frac{1}{(1 - R_j^2)}$ bariantzaren puztea den. Era laburrean azalduz, R_j^2 batetik gero eta hurbilago badago, hau da, X_j aldagaiak beste aldagai azaltzaileekin duen kolinealitatea gero eta handiagoa bada, VIF_j ere handitu egiten da ($1 \leq VIF_j \leq \infty$), ikusi dugunez, kolinealitateak bariantza “puztu” egiten baitu. Egile hauen arabera, $VIF_j > 10$ bada, X_j aldagaiaren kolinealitatea bestetik handia dela esango dugu. (OHARRA: jasankortasuna VIF -aren alderantzizkoa da: $JAS_j = \frac{1}{VIF_j}$).

Aipatu beharra dago, kolinealitatearen azterketa VIF -ak edota jasankortasuna erabiliz, zenbait kritika jasan dituela, askotan hauen arabera lortutako emaitzak ez baitira erabatekoak. Ikusi bezala, koefizienteen estimatzaileen bariantzak, VIF_j , σ^2 eta $\sum (X_{ji} - \bar{X}_j)^2$ -en arabera daudenez, VIF_j altu bat ez da baldintza nahikoa eta beharrezkoa estimatzaileen bariantza handiak lortzeko. Nahiz eta VIF_j altu bat izan, baliteke σ^2 txikia edota $\sum (X_{ji} - \bar{X}_j)^2$ oso handia izatea eta kolinealitatearen ondorioen arraztorik ez izatea kontraste eta tartezko estimazioetan adibidez.

Etxebizitzaren prezioen adibidearekin jarraituz, aldagaien artean kolinealitatea dugun edo ez aztertzeko bariantzaren puzte koefizientea kalkulatu dugu. Horretarako, eredu KTA bitartez estimatutako lehiatilan *Kontrasteak* $\rightarrow$ *Kolinealitatea* klikatu dugu, bertan azaltzen baitira kolinealitate arazoa emateko aztertu eta konparatu beharreko balioak. Ondorengo lortzen da:

Ikus dezakegunez, VIF_j -aren arabera, kolinealitate arazorik ez dugula ondorioztatuko genuke. Hortaz, eredu BATHS eta BEDRMS aldagaiak ez nabariak dira etxebizitzaren prezioa

6.2 Taula: Kolinealitate kontrastearen emaitzak

Bariantzaren Inflazio Faktoreak

Balio minimo posiblea = 1.0

Balioak > 10.0 badira, agian kolinealitate arazoren bat dago

2)	sqft	2,651
3)	bedrms	1,406
4)	baths	2,900

$VIF(j) = 1/(1 - R(j)^2)$, non $R(j)$ j aldagaiaren eta beste aldagai azaltzaileen arteko koerlazio koefiziente anizkoitza den

$X'X$ matrizearen propietateak:

1-norma = 55654161

Determinantea = 1,366856e+008

Elkarrekiko baldintza zenbakia = 2,808412e-009

azaltzeko eta SQFT ordea, nabaria da eta aurreko gaian ikusi bezala, proposatzen genuen **Eeredua** erabiliko genuke etxebizitzaren prezioa azaltzeko.

Kolinealitate arazoa zuzentzeko proposaturiko **soluzioak** asko dira, baina orokorrean ez dira asebetegarriak.

- Lagin horretan ematen den arazoa bada, datuak aldatuz posible da (ez da beti ematen) kolinealitate arazoa konpontzea. Ideia, aurrekoak baino gutxiago erlazionaturiko datuak barneratzea izango da, datu bakar batzuk barneratuz edota lagina aldatuz. Edonola ere, oso gutxitan izan ohi da hobeagoak diren datuak lortzeko aukera.
- Askotan, koefizienteen informazioa barneratuz arazoa desagertu egiten da. Hala ere, informazio gehigarri hori kolinealitate arazoa hauteman baino lehen kontuan hartu beharko genuke eta ez ondoren, informazio gehiagorekin eredua era efizienteagoan estimatuko baitugu.
- Kolinealitate anizkoitza sortarazten duten aldagaien bat eredutik kendu. Hala ere, soluzio honekin kontu handia izan behar da, ereduaren zehazpen oker bat izan dezakegu eta, hau da aldagai nabari baten omisioa. Ondorioz, lortuko genituzkeen koefizienteen KTA estimatzaileak eta perturbazioaren bariantzaren estimatzailea alboratuak izango lirarteke eta proposatutako kontrasteak (inferentzia) ez lirarteke baliogarriak izango.
- Aldagaien arteko kolinealitatea altua bada, intuitiboki baliogarria eta ulargarria den soluzio posible bat **Osagai Nagusiko Metodoa** erabiltzean datza. Hau da, metodo honekin X matrizeatik aldagai kopuru txikiago bat kanporatuko da, konkretuki aldagai azaltzailearen aldakuntza edo bariantza gutxia edo gehiena jasoko duten aldagaiak. Horrela lortzen diren estimatzailearen bariantza, KTA estimatzailearena baino txikia-

goa izan arren, alboratua izango da eta gainera, aldagaien neurri unitateen menpekoa. Bestalde, osagai nagusiak aukeratzeko ez dugu aldagai azaltzaile eta azalduaren arteko erlazioa kontuan hartzen eta azkenik, horrela lortzen diren estimatzaileak, hasierako ereduko koefizienteen konbinazioak dira eta ondorioz, oso zaila izaten da emaitzak interpretatzea.

- **Cresta Erregresio Estimatzailerak:** $(\hat{\beta}_\lambda) = [X'X + \lambda D]^{-1} X'Y$ non D matrize diagonal bat den $X'X$ matrizeko diagonal nagusiko elementuak izanik, λ parametro ezezaguna aukeratu egin behar da. Normalean parametro hau aukeratzeko balio txiki batekin hasten gara $\lambda = 0,01$ eta bere balioa handitzen joaten gara estimatzailea egonkortu arte. Aurkeztutako estimatzailea alboratua izan arren, bere bariantza eta kobariantza matrizea, KTA estimatzailearena baino “txikiagoa” da eta Batezbesteko Errore Koadratiko (BEK) txikiagoa du. KTAko estimatzailea kasu berezi bat bezala lor daiteke $\lambda = 0$ denean. Estimatzaileraren batezbestekoa eta bariantza honakoak dira:

$$E(\hat{\beta}_\lambda) = \sigma^2 [X'X + \lambda D]^{-1} X'X \beta$$

$$Var(\hat{\beta}_\lambda) = \sigma^2 [X'X + \lambda D]^{-1} X'X [X'X + \lambda D]^{-1}$$

Bibliografia

- Belsley, D., E. Kuh eta R. Welsch** (1980). *Regression Diagnostics*. Wiley, New York.
- Neter, J., W. Wasserman eta M.H. Kutner** (1990). *Applied Linear Statistical Models*, 3. ed., Boston, M.A.: Irwin, Boston.
- Ramanathan, R.** (2002). *Introductory Econometrics with Applications*, 5. ed., South Western, Ohio.

7 Gaia

Aldagai koalitatiboak

Aurkibidea

7.1	Sarrera	106
7.2	Aldagai koalitatibo bakarra duen eredua	106
	Jatorri aldaketa	112
	Jatorri eta malda aldaketa	114
7.3	Aldagai koalitatibo bi edo gehiago dituen eredua	116
7.4	Egitura aldaketa	119

7.1 Sarrera

Orain arte aldagai koantitatiboak dituzten ereduak besterik ez dugu zehaztu. Hau da, zenbakizko balioak hartzen dituzten aldagaiak erabili ditugu. Hala ere, aldagaiak koalitatiboak ere izan daitezke, hau da, zenbakizko balioak hartzen ez dituztenak, halaber kategoriak, klaseak edo atributuak dituztenak. Besteak beste, banakoen generoa, egoera zibila, arraza, kokapena edo gertakizun historiko baten barnerapena, urtaroak,... aldagai koalitatiboak dira. Adibidez, langileen soldata generoaren funtzioan egon daiteke, banakoen arrazak edo bizilekuak kriminalitate-tasa azaldu dezakete, gerrateek edo beste gertakizun historikoek eragina izan dezakete herrialdeen BPGean, produktu baten salmentak ezberdinak izan daitezke urtaroen arabera, etab.

Nahiz eta aldagai azaldua koantitatiboa bezala mantentzen jarraitu, gai honetan aldagai azaltzaile koalitatiboak edota koantitatiboak dituzten ereduak kontsideratuko ditugu.

Aldagaien kategoriak koantitatiboki neurgarriak ez direnez, zenbakizko aldagaiak sortu behar dira, fikzio-aldagaiak izenekoak. Fikzio-aldagaiek 1 balioa hartzen dute banakoa kategorian konkritu batean barneratua baldin badago eta 0 balioa bestelako kasuetan¹.

$$D_i = \begin{cases} 1 & \text{i banakoa kategorian baldin badago} \\ 0 & \text{en caso contrario} \end{cases}$$

Gai honetan, estimazioan, koefizienteen interpretazioan eta hipotesien kontrasteetan oinarrituko gara aldagai azaltzaile koalitatiboak dituzten ereduentzat. Ramanathaneko (2002) data7-3 eta data7-19 fitxategiak erabiliz, zenbait eredu aztertuko ditugu.

7.2 Aldagai koalitatibo bakarra duen ereduak

Bi kategoriako aldagai koalitatiboa

Aldagai koalitatibo bakarra duen ereduak, kasurik errezena da. Suposa dezagun etxebizitzaren prezioa azaltzeko etxebizitzak igerilekua izatea bakarrik kontuan hartzen dugula². Horretarako, ondoko fikzio-aldagaia definitzen dugu:

$$POOL_i = \begin{cases} 1 & \text{i-etxebizitzak igerilekua badu} \\ 0 & \text{en caso contrario} \end{cases}$$

Ramanathaneko (2002) data7-3 fitxategia irekitzen dugu. Fitxategi honek 14 etxebizitzaren hurrengo datuak ditu: aurreko gaietan erabilitako etxebizitzaren salmenta prezioa (PRICE),

¹ Fikzio-aldagaiek edozein balio har ditzakete, baina koefizienteen interpretazioa errezagoa da 0 eta 1 balioak hartzen baldin badituzte.

² Errazteko asmoz, etxebizitzaren prezioari eragiten dioten gainontzeko aldagaiak alde batera uzten ditugu.

oin karratu bizigarriak (SQFT), gela kopurua (BEDRMS) eta komun kopurua (BATHS) eta aurrean definitutako POOL fikzio-aldagaiaren erantzunak.

PRICE eta POOL aldagaiak aukeratu eta beraien balioak behatuz hurrengo balioak ditugu:

7.1 Taula: Pool fikzio-aldagaiaren balioak

Obs	price	pool
1	199,9	1
2	228,0	0
3	235,0	1
4	285,0	0
5	239,0	0
6	293,0	0
7	285,0	0
8	365,0	1
9	295,0	0
10	290,0	0
11	385,0	1
12	505,0	1
13	425,0	0
14	415,0	0

Adibidez, laginaren lehenengo etxebizitzaren prezioa 199900 dolarrekoa da eta igerilekua dauka POOL aldagaiak 1 balioa hartzen duelako. Bigarrenak, berriz, ez dauka igerilekurik (POOL aldagaiak 0 balioa hartzen du) eta bere salmenta prezioa 228000 dolarrekoa da.

Aurreko PRICE aldagaiari dagozkion oinarritzko estatistikoak ateraz, oso erraz lortu daiteke etxebizitzaren batezbesteko salmenta prezioa 317493 dolarrekoa dela:

7.2 Taula: Price aldagaiaren estatistiko nagusiak

Estatistikoen laburpena, 1 - 14 behaketak erabiliz
'price' aldagaiarentzat (14 behaketa baliagarriak)

Batezbestekoa	317,49
Mediana	291,50
Minimoa	199,90
Maximoa	505,00
Desbiderazio tipikoa	88,498
Aldakuntza Koefizientea (A.K.)	0,27874
Asimetria	0,65346
Kurtosis soberakina	-0,52983

Hala ere, alde batetik, igerilekurik ez duten etxebizitzaren batezbesteko prezioa eta bestaldek,

igerilekua dutenena kalkulatu daitezke. Horretarako, igerilekua duten etxebizitzaren salmenta prezioa hautatzen dugu, lehendabizi, PRICE aldagaia aukeratzen dugu, gero *Lagina* → *Definitu*, *fikzio-aldagaian oinarrituz* ... klikatzen dugu eta POOL aldagaia aukeratu ondoren *Ados* klikatu. Ondoren, oinarritzko estatistikoak lortzen ditugu:

7.3 Taula: Igerilekua duten etxebizitzaren *price* aldagaiaren estatistiko nagusiak

Estatistikoen laburpena, 1 - 5 behaketak erabiliz
'price' aldagaiarentzat (5 behaketa baliagarriak)

Batezbestekoa	337,98
Mediana	365,00
Minimoa	199,90
Maximoa	505,00
Desbiderazio tipikoa	122,99
Aldakuntza Koefizientea (A.K.)	0,36390
Asimetria	0,15896
Kurtosis soberakina	-1,2798

Lagin osoaren tamaina berreskuratzeko *Lagina* → *Berrezarri lagin osoa* klikatu.

Igerilekurik ez duten etxebizitzaren prezioa aukeratzeko, berriz, *Lagina* → *Murritz honako irizpidean oinarrituz*... klikatu, *pool=0* baldintza sartu eta adostu. Horrela, NOPOOL aldagaia sortzen dugu, zeinak 1 balioa hartzen duen igerilekurik gabeko etxebizitza bada eta 0 balioa bestelako kasuan. PRICE aldagaiaren oinarritzko estatistikoak ondokoak dira:

7.4 Taula: Igerilekua ez duten etxebizitzaren *price* aldagaiaren estatistiko nagusiak

Estatistikoen laburpena, 1 - 9 behaketak erabiliz
'price' aldagaiarentzat (9 behaketa baliagarriak)

Batezbestekoa	306,11
Mediana	290,00
Minimoa	228,00
Maximoa	425,00
Desbiderazio tipikoa	68,959
Aldakuntza Koefizientea (A.K.)	0,22527
Asimetria	0,87575
Kurtosis soberakina	-0,52255

Hortaz, igerilekua duten etxebizitzaren batezbesteko prezioa 337980 dolarrekoa da eta igerilekurik ez dutenena, aldiz, 306110ekoa. Hala ere, emaitza hauetatik ezin da ondorioztatu igerilekua duen etxebizitzaren batezbesteko prezioa 31870 dolar garestiagoa denik zeren eta etxebizitzaren prezioan eragina izan dezaketen bestelako faktoreak (oin karratu bizigarriak, gela kopurua,...) ez baitira kontuan hartu.

POOL aldagai azaltzailea erabiliz, eredu ekonometriko bat zehaztu, estimatu, kontrasteak egin eta etxebizitzaren prezioetan eragina izan dezaketen bestelako ezaugarriak barneratu ditzakegu. Hasteko, ondoko eredua kontsideratzen dugu:

$$PRICE_i = \alpha + \beta POOL_i + u_i \quad i = 1, \dots, 14 \quad (7.1)$$

Koefizienteen interpretazioa eta estimazioa

Gure adibidean, populazioko erregresio funtzioa aldatzen da etxebizitzak igerilekua izatearen arabera:

- $E(PRICE_i | i \text{ etxebizitzak igerilekua du}) = \alpha + \beta$ da, POOL aldagaiak 1 balioa hartzen duelako eta $E(u_i) = 0$ delako.
- $E(PRICE_i | i \text{ etxebizitzak ez du igerilekurik}) = \alpha$ da, POOL aldagaiak 0 balioa hartzen duelako eta $E(u_i) = 0$ delako.

Beraz, koefizienteak horrela interpretatzen dira:

- α : igerilekurik gabeko etxebizitzaren batezbesteko prezioa.
- $\alpha + \beta$: igerilekua duen etxebizitzaren batezbesteko prezioa.

Hortaz:

- β : igerilekua duen etxebizitzaren batezbesteko prezioaren diferentzia igerilekurik ez duenarekiko.

Eredua KTAen bidez estimatuz, ondoko emaitza lortzen dugu Gretlen bitartez:

Eredua 1: KTA estimazioak 14 behaketak erabiliz 1–14
Aldagai azaldua: price

Aldagaia	Koefizientea	Desb. Tipikoa	<i>t</i> -estatistikoa	p-balioa
const	306,111	30,2077	10,1335	0,0000
pool	31,8689	50,5471	0,6305	0,5402

Aldagai azalduaren batezbestekoa	317,493
Aldagai azalduaren Desb. Tip.	88,4982
Hondar Karratuen Batura	98550,5
Hondarren desbiderazio tipikoa ($\hat{\sigma}$)	90,6231
R^2	0,0320632
Zuzendutako $\bar{R}^2$	-0,0485982
Askatasun graduak	12
Log-egiantza	-81,880
Akaike Informazio Irizpidea	167,760
Schwarz Bayesian Irizpidea	169,038
Hannan–Quinn Irizpidea	167,642

Erregresio bakuneko eredua estimatzeko, 2. gaian lortu genituen ekuazio normalak erabiliz eta POOL fikzio-aldagaiaren karratua aldagai bera dela kontuan hartuz (0 eta 1 balioak hartzen dituelako), (7.1) ereduko koefizienteen estimatzaileak lagin batezbestekoen bitartez lortzen dira³:

- $\hat{\alpha} = \overline{PRICE}_{nopool} = 306,11 \Rightarrow$ igerilekurik gabeko etxebizitzaren batezbesteko prezio estimatua.
- $\hat{\beta} = \overline{PRICE}_{pool} - \overline{PRICE}_{nopool} = 337,98 - 306,11 = 31,87 \Rightarrow$ igerilekua duen etxebizitzaren batezbesteko prezioaren diferentzia estimatua igerilekurik ez duenarekiko.

Etxebizitzaren prezio aldakuntzak azaltzeko etxebizitzak igerilekua izatearen funtzioan jartzeari ez da zehazpen zuzen bakarra. NOPOOL fikzio-aldagaiaren erabiliz ondorengo bi eredu baliokideak zehaz daitezke:

$$PRICE_i = \gamma_1 + \gamma_2 NOPOOL_i + u_i \quad i = 1, \dots, 14 \quad (7.2)$$

$$PRICE_i = \beta_1 POOL_i + \beta_2 NOPOOL_i + u_i \quad i = 1, \dots, 14 \quad (7.3)$$

Koefizienteak interpretatzeko, (7.1), (7.2) eta (7.3) ereduetako baliokidetasuna ondorengo da:

- $E(PRICE_i | i \text{ etxebizitzak igerilekua du}) = \alpha + \beta = \gamma_1 = \beta_1$
- $E(PRICE_i | i \text{ etxebizitzak ez du igerilekurik}) = \alpha = \gamma_1 + \gamma_2 = \beta_2$

Oker logikean zehazpen bat ondorengo litzateke:

$$PRICE_i = \alpha + \beta_1 POOL_i + \beta_2 NOPOOL_i + u_i \quad i = 1, \dots, 14$$

zeren eredu honetan X datu matrizea aztertzerakoan, bigarren eta hirugarren zutabeen batuketaren lehenengoa baita eta aipatutako oinarritzko hipotesi bat, $h(X) < k$, ez da betetzen. Kolinealitate anizkoitz zehatza dago, hau da, aldagaien arteko konbinazio lineal bat dago. Ondorioz, $X'X$ singularra denez ezin dugu bere alderantzizkoa atera eta ezin izango genuke eredua KTAen bitartez era bakar batean estimatu.

³ $\overline{PRICE}_{pool}$ igerilekua duen etxebizitzaren lagin batezbesteko prezioa da eta $\overline{PRICE}_{nopool}$ igerilekurik gabeko etxebizitzaren lagin batezbesteko prezioa da.

$$X = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix}$$

Hipotesien kontrastea

Hipotesien kontrasteak aurreko gaitan ikusitako metodologia erabiliz egiten dira. Adibidez, etxebizitzaren batezbesteko prezioaren diferentzia igerilekua izateagatik, esanguratsua dela kontrastatzeko hipotesi hutsa (7.1) ereduan $\beta = 0$ da⁴. Kontraste hau egiteko t -estatistikoa erabili daiteke. Kasu honetan bere p -balioa 0,5402 denez, ez dugu hipotesi hutsa baztertzen %5eko esangura-mailarekin, hau da, etxebizitzaren batezbesteko prezioaren diferentzia igerilekua izateagatik ez da esanguratsua. Kontraste bera hondar karratuen baturetan oinarritutako F -estatistikoa erabiliz ere egin daiteke. Honela, (7.1) murriztu gabeko eredua eta $PRICE_i = \alpha + u_i \quad i = 1, \dots, 14$ murriztutako eredua izanik.

Kategoria bi baino gehiagoko aldagai koalitatiboa

Aurrean ikusitako kasua orokortu daiteke kategoria bi baino gehiagoko aldagai koalitatiboa ereduan barneratzean.

Demagun igerilekua duten etxebizitzaren prezioa azaldu nahi dugula eta, errazteko asmoz, igerilekuaren tamaina dela aldagai azaltzaile bakarra. Horretarako, ondoko fikzio-aldagaiak definitzen ditugu:

$$D_{1i} = \begin{cases} 1 & \text{i etxebizitzaren igerilekua txikia baldin bada (10 m}^2 \text{ baino gutxiago)} \\ 0 & \text{en caso contrario} \end{cases}$$

$$D_{2i} = \begin{cases} 1 & \text{i etxebizitzaren igerilekua ertaina baldin bada (10 eta 20 m}^2 \text{ artean)} \\ 0 & \text{en caso contrario} \end{cases}$$

$$D_{3i} = \begin{cases} 1 & \text{i etxebizitzaren igerilekua handia baldin bada (20 m}^2 \text{ baino gehiago)} \\ 0 & \text{en caso contrario} \end{cases}$$

⁴Era berean, $H_0 : \gamma_2 = 0$ edo $H_0 : \beta_1 = \beta_2$ (7.2) eta (7.3) ereduetan, hurrenez hurren.

Eredua zehazterakoan, termino konstantea eta hiru fikzio-aldagaiak barneratzen baditugu kolinealitate anizkoitzaren arazoa daukagu. Arazo hau ekiditzeko termino konstantea edo fikzio-aldagaietatik bat kendu behar da. Adibidez, zehazpen baliokide posible batzuk ondo-koak dira:

$$PRICE_i = \alpha_1 D_{1i} + \alpha_2 D_{2i} + \alpha_3 D_{3i} + u_i \quad i = 1, \dots, N \quad (7.4)$$

$$PRICE_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + u_i \quad i = 1, \dots, N \quad (7.5)$$

non $\alpha_1 = \beta_1$ igerileku txikia duen etxebizitzaren batezbesteko prezioa den, $\alpha_2 = \beta_1 + \beta_2$ igerileku ertaina duenarena eta $\alpha_3 = \beta_1 + \beta_3$ igerileku handia duenarena. Hortaz, β_2 igerileku ertaina duen etxebizitzaren batezbesteko salmenta prezioaren diferentzia igerileku txikia duenarekiko da eta β_3 igerileku handia duen etxebizitzaren batezbesteko salmenta prezioaren diferentzia igerileku txikia duenarekiko.

Lehen aipatu dugun bezala, hipotesien kontrasteak aurreko gaitan ikusitako metodologia erabiliz egiten dira. Adibidez, (7.4) ereduan igerilekuaren tamainak ez daukala eraginik etxebizitzaren batezbesteko prezioan kontrastatzeko hipotesi hutsa $H_0 : \alpha_1 = \alpha_2 = \alpha_3$ da, (7.4) murriztu gabeko eredua izanik eta $PRICE_i = \alpha + u_i \quad i = 1, \dots, N$ murriztutako eredua. Era berean, (7.5) ereduan hipotesi hutsa $H_0 : \beta_2 = \beta_3 = 0$ da, (7.5) murriztu gabeko eredua izanik eta $PRICE_i = \beta_1 + u_i \quad i = 1, \dots, N$ murriztutako eredua.

Aldagai koantitatiboen barnerapena

Orain arte etxebizitzaren prezioa azaltzeko aldagai azaltzaile bakarra koalitativoa izan dugu. Hala ere, askotan aldagai koalitativo batez gain, aldagai koantitatiboak barneratzen dira aldagai endogenoa azaltzeko. Gure adibidean, etxebizitzaren prezioa etxebizitzak igerilekua izatearen funtzioan zehaztea ez da oso errealista, beraz, SQFT (etxebizitzaren oin karratu bizigarriak) aldagai koantitatibo aldagai azaltzaile moduan erantsiko dugu ereduan. Horretarako, bi aukera ikusiko ditugu. Lehenengoan, SQFT aldagaiak $E(PRICE_i)$ zuzenean jatorri aldaketa eragiten du. Bigarrean, berriz, bai jatorrian baita zuzenaren maldan ere aldaketa eragiten du.

Jatorri aldaketa

SQFT aldagaia barneratzen dugu gure ereduan aldagai azaltzaile moduan:

$$PRICE_i = \alpha_1 + \alpha_2 POOL_i + \beta_1 SQFT_i + u_i \quad i = 1, \dots, 14 \quad (7.6)$$

Koefizienteen interpretazioa eta estimazioa

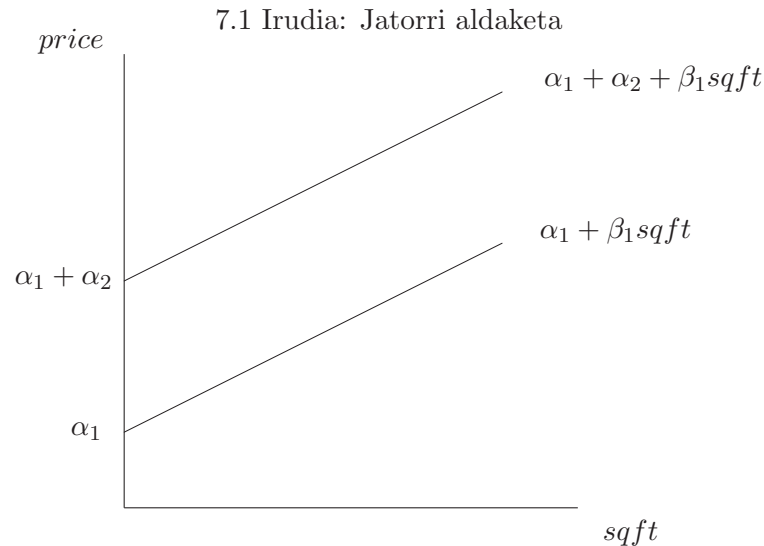
Populazioko erregresio funtzioa ondokoa da:

- $E(PRICE_i | i \text{ etxebizitzak igerilekua dauka}) = \alpha_1 + \alpha_2 + \beta_1 SQFT_i$

- $E(PRICE_i | i \text{ etxebizitzak ez dauka igerilekurik}) = \alpha_1 + \beta_1 SQFT_i$

Beraz, α_1 igerileku gabeko eta oin karratu bizigarririk gabeko etxebizitzaren batezbesteko prezioa da, α_2 igerilekua duen etxebizitzaren batezbesteko prezioaren diferentzia igerilekurik ez duenarekiko, oin karratu bizigarrien kopurua berdina izanik eta β_1 -k etxebizitza oin karratu bizigarri batean handitzen denean batezbesteko prezioaren gehikuntza adierazten du.

Grafikoki, malda berdineko (β_1) eta jatorri ezberdineko zuzen bi dauzkagu:



Ondoren (7.6) ereduko KTA bidezko estimazioaren emaitzak ematen dira:

Eredua 2: KTA estimazioak 14 behaketak erabiliz 1–14

Aldagai azaldua: price

Aldagaia	Koefizientea	Desb. Tipikoa	<i>t</i> -estatistikoa	p-balioa
const	22,6728	29,5058	0,7684	0,4584
pool	52,7898	16,4817	3,2029	0,0084
sqft	0,144415	0,0141849	10,1809	0,0000

Aldagai azalduaren batezbestekoa	317,493
Aldagai azalduaren Desb. Tip.	88,4982
Hondar Karratuen Batura	9455,36
Hondarren desbiderazio tipikoa ($\hat{\sigma}$)	29,3186
R^2	0,907132
Zuzendutako $\bar{R}^2$	0,890247
$F(2, 11)$	53,7238
Log-egiantza	-65,472
Akaike Informazio Irizpidea	136,944
Schwarz Bayesian Irizpidea	138,861
Hannan–Quinn Irizpidea	136,767

Emaitza hauetan oinarrituriko lagineko erregresio funtzioa hurrengoa da:

$$\widehat{PRICE}_i = 22,673 + 52,790 POOL_i + 0,144 SQFT_i$$

$(t_{est}) \quad (0,768) \quad (3,203) \quad (10,181)$

Aldagaiak aztertzerakoan, aldagai azaltzaile biak esanguratsuak direla kontrastatu daiteke etxebizitzaren batezbesteko prezioa azaltzeko. Izatez, dagozkien koefizienteen t -estatistikoak Student- t tauletako $t_{(N-K)0,05/2} = t_{(14-3)0,05/2} = 2,201$ balioa baino handiagoak dira. Beraz, alde batetik, etxebizitzaren batezbesteko prezioen diferentzia etxebizitzak igerilekua izateagatik esanguratsua da. Bestalde, etxebizitzaren oin karratu bizigarrien kopurua garrantzitsua da prezioa azaltzeko.

Koefiziente estimatuen interpretazioa ondorengoa da:

- $\hat{\alpha}_1 = 22,673 \Rightarrow$ igerileku gabeko eta oin karratu bizigarri gabeko etxebizitzaren batezbesteko prezio estimatua 22673 dolarrekoa da.
- $\hat{\alpha}_2 = 52,790 \Rightarrow$ igerilekua duen etxebizitzaren batezbesteko prezio estimatua 52790 dolar garestiagoa da igerilekua ez duen etxebizitzaren batezbesteko prezio estimatuarekiko, beste guztia konstante mantenduz, hau da, oin karratu bizigarrien kopuru berdina izanik.
- $\hat{\beta}_1 = 0,144 \Rightarrow$ etxebizitza oin karratu bizigarri batean handitzen denean batezbesteko prezio estimatua 144 dolarretan gehitzen da.

Jatorri eta malda aldaketa

Etxebizitzaren azalera oin karratu batean handitzerakoan, dakartzan prezioaren batezbesteko aldakuntza, igerilekua izatearen menpekoea dela pentsa daiteke. Hala bada, POOL fikzioaldagaiak eragina dauka bai zuzenaren jatorrian eta baita maldan ere. Ondoko ereduak aipatutako eragina jasotzen du:

$$PRICE_i = \alpha_1 + \alpha_2 POOL_i + \beta_1 SQFT_i + \beta_2 POOL \times SQFT_i + u_i \quad i = 1, \dots, 14 \quad (7.7)$$

$POOL \times SQFT$ terminoak igerilekua duten etxebizitzaren oin karratu bizigarriak biltzen ditu eta igerilekurik ez dutenentzat berriz, 0 balioa hartzen du.

Koefizienteen estimazioa eta interpretazioa

$POOL \times SQFT$ aldagai berria definitu ondoren, (7.7) eredu estimatzen dugu:

Eredua 3: KTA estimazioak 14 behaketak erabiliz 1–14
Aldagai azaldua: price

Aldagaia	Koefizientea	Desb. Tipikoa	t -estatistikoa	p-balioa
const	77,1332	25,6379	3,0086	0,0131
pool	-82,648	39,7759	-2,0779	0,0644
sqft	0,116667	0,0125934	9,2641	0,0000
poolxsqft	0,0722955	0,0203274	3,5566	0,0052

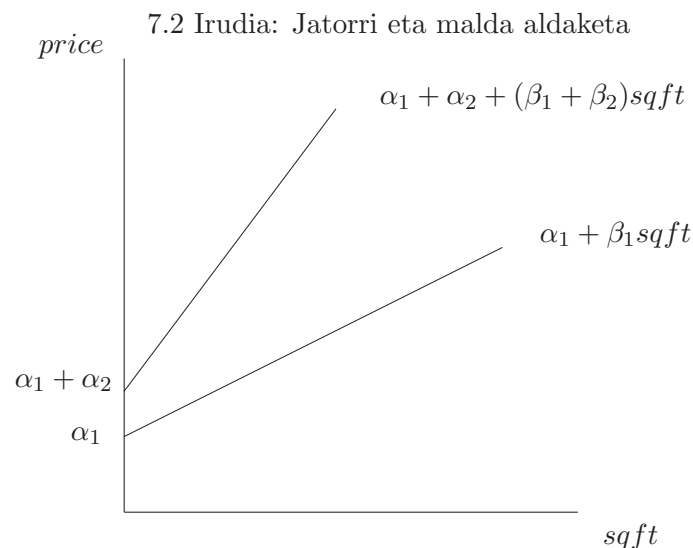
Aldagai azalduaren batezbestekoa	317,493
Aldagai azalduaren Desb. Tip.	88,4982
Hondar Karratuen Batura	4174,72
Hondarren desbiderazio tipikoa ($\hat{\sigma}$)	20,4321
R^2	0,958997
Zuzendutako $\bar{R}^2$	0,946696
$F(3, 10)$	77,9615
Log-egiantza	-59,749
Akaike Informazio Irizpidea	127,499
Schwarz Bayesian Irizpidea	130,055
Hannan–Quinn Irizpidea	127,262

Populazioko erregresio funtzioak:

- $E(PRICE_i | i \text{ etxebizitzak igerilekua du}) = \alpha_1 + \alpha_2 + (\beta_1 + \beta_2)SQFT_i$
- $E(PRICE_i | i \text{ etxebizitzak ez du igerilekurik}) = \alpha_1 + \beta_1 SQFT_i$

α_1 parametroa aurreko kasuan ikusi dugun moduan interpretatzen da. α_2 azalera zero eta igerilekua duen etxebizitzaren batezbesteko salmenta prezioaren diferentzia igerilekurik ez duenarekiko da, β_1 -ek etxebizitzak oin karratu bat gehiago izaterakoan, igerilekurik ez duen etxebizitzaren batezbesteko prezioaren gehikuntza neurtzen du, $(\beta_1 + \beta_2)$ -k etxebizitzak oin karratu bat gehiago izaterakoan, igerilekua duen etxebizitzaren batezbesteko prezioaren gehikuntza neurtzen du eta hortaz, β_2 -k etxebizitzak oin karratu bat gehiago izaterakoan, igerilekua duen etxebizitzaren batezbesteko prezioaren diferentziaren gehikuntza neurtzen du.

Grafikoki, malda eta jatorri ezberdineko bi zuzen dauzkagu:



Koefiziente estimatuen interpretazioa:

- $\hat{\alpha}_1 = 77,133 \Rightarrow$ etxebizitzaren batezbesteko salmenta prezio estimatua, azalera zero denean eta igerilekurik ez duenean, 77133 dolarrekoa da.

- $\hat{\alpha}_2 = -82,648 \Rightarrow$ azalera zero eta igerilekua duen etxebizitzaren batezbesteko salmenta prezioaren diferentzia estimatua igerilekurik ez duenarekiko -82648 dolarrekoa da.
- $\hat{\beta}_1 = 0,117 \Rightarrow$ etxebizitzak oin karratu bat gehiago izaterakoan, igerilekurik ez duen etxebizitzaren batezbesteko gehikuntza estimatua 117 dolarrekoa da.
- $\hat{\beta}_2 = 0,072 \Rightarrow$ etxebizitzak oin karratu bat gehiago izaterakoan, igerilekua duen etxebizitzaren batezbesteko prezio estimatuaren gehikuntza, igerilekua ez duen etxebizitzaren batezbesteko prezio estimatuaren gehikuntzarekiko diferentzia 72 dolarrekoa da.

Hipotesien kontrastea

Etxebizitzaren batezbesteko prezioa desberdina dela igerilekua izateagatik kontrastatzeko hipotesi hutsa $H_0 : \alpha_2 = \beta_2 = 0$ eta aurkakoa $H_a : \alpha_2 \neq 0$ edota $\beta_2 \neq 0$ dira.

Kontrastearen emaitza ondokoa da:

Hipotesi hutsa: aldagaien erregresio koefizienteak zero dira

```
pool
poolxsqft
```

Estatistikoa: $F(2, 10) = 16,886$, p-balioarekin = 0,000622329

hortaz, %5eko esangura-mailarekin hipotesi hutsa baztertzen dugu eta igerilekua izateak eragina dauka etxebizitzaren batezbesteko prezioetan.

Halaber, banakako esanguratasun kontrastearen bidez, kontrastatu daiteke ea etxebizitzak igerilekua izateak batezbesteko prezioan eragiten duen jatorrian bakarrik ($H_0 : \alpha_2 = 0$) edo maldan bakarrik ($H_0 : \beta_2 = 0$).

7.3 Aldagai koalitatibo bi edo gehiago dituen eredua

Zehaztutako (7.6) ereduan, etxebizitzaren prezioa azaldu lezaketeen aldagai azaltzaile gehiago barneratu daitezke. Adibidez, etxebizitzak egongela izatea, tximinia izatea, logela kopurua eta komun kopurua. Lehenengo bi aldagaiak (egongela eta tximinia izatea) koalitatiboak dira eta azken biak (logela eta komun kopurua), berriz, koantitatiboak dira. Demagun ondoko eredua zehazten dugula:

$$PRICE_i = \alpha_1 + \alpha_2 POOL_i + \alpha_3 FAMROOM_i + \alpha_4 FIREPL_i + \beta_1 SQFT_i + \beta_2 BEDRMS_i + \beta_3 BATHS_i + u_i \quad i = 1, \dots, 14 \quad (7.8)$$

non FAMROOM eta FIREPL hurrengo fikzio-aldagaiak diren:

$$FAMROOM_i = \begin{cases} 1 & \text{i etxebizitzak egongela baldin badu} \\ 0 & \text{en caso contrario} \end{cases}$$

$$FIREPL_i = \begin{cases} 1 & \text{i etxebizitzak tximinia baldin badu} \\ 0 & \text{en caso contrario} \end{cases}$$

BEDRMS_{*i*}-k, i etxebizitzaren logela kopurua adierazten du eta BATHS_{*i*}-k, i etxebizitzaren komun kopurua adierazten du.

Aukeratutako zehazpenean, (7.8) ereduan, hiru aldagai koalitatibo (igerilekua izatea, tximinia izatea eta egongela izatea) daude eta jatorrian bakarrik eragiten dute. Zehazpenean, termino konstantea mantendu da eta aldagai koalitatibo bakoitzatik kategoria bat kendu da. Koefizienteak aurreko ataletan ikusi dugun moduan interpretatzen dira. Adibidez, α_3 koefizienteak egongela duen etxebizitzaren batezbesteko prezioaren diferentzia egongelarik ez duenarekiko adierazten du beste aldagai azaltzaile guztiak konstante mantenduz. Estimazioaren emaitzak ondokoak dira:

Eredua 4: KTA estimazioak 14 behaketak erabiliz 1–14
Aldagai azaldua: price

Aldagaia	Koefizientea	Desb. Tipikoa	<i>t</i> -estatistikoa	p-balioa
const	39,0571	89,5397	0,4362	0,6758
pool	53,1958	22,0635	2,4110	0,0467
famroom	−21,344	42,8734	−0,4979	0,6338
firepl	26,1880	53,8454	0,4864	0,6416
sqft	0,146551	0,0301014	4,8686	0,0018
bedrms	−7,0455	28,7363	−0,2452	0,8134
baths	−0,263691	41,4547	−0,0064	0,9951

Aldagai azalduaren batezbestekoa	317,493
Aldagai azalduaren Desb. Tip.	88,4982
Hondar Karratuen Batura	9010,24
Hondarren desbiderazio tipikoa ($\hat{\sigma}$)	35,8773
R^2	0,911504
Zuzendutako $\bar{R}^2$	0,835650
$F(6, 7)$	12,0166
p-balioa $F()$	0,00221290
Log-egiantza	−65,134
Akaike Informazio Irizpidea	144,269
Schwarz Bayesian Irizpidea	148,743
Hannan–Quinn Irizpidea	143,855

Koefiziente estimatuen interpretazioa:

- $\hat{\alpha}_1 = 39,057$: etxebizitzaren batezbesteko salmenta prezio estimatua, azalera zero duen eta igerilekurik, komunik, logelik eta tximinirik ez dituenen, 39057 dolarrekoa da.
- $\hat{\alpha}_2 = 53,1958$: igerilekua duen etxebizitzaren batezbesteko salmenta prezioaren diferentzia estimatua igerilekurik ez duenarekiko 53196 dolarrekoa da, gainontzeko ezaugarriak (azalera, logelak, komunak, egongela eta tximinia) berdinak izanik.
- $\hat{\alpha}_3 = -21,34$: egongela duen etxebizitzaren batezbesteko prezio estimatua egongelarik ez duenarena baino 21340 dolar merkeagoa da, gainontzeko ezaugarriak berdinak izanik. Zeinu hau izatearen arrazoia, etxebizitzaren azalera, logela eta komun kopurua konstante mantenduz, egongela izateak gelak edota komunak txikiagoak izan behar direla inplikatzeko duela, da.
- $\hat{\alpha}_4 = 26,188$: tximinia duen etxebizitzaren batezbesteko prezio estimatua tximinirik ez duenarena baino 26188 dolar garestiagoa da, gainontzeko ezaugarriak berdinak izanik.
- $\hat{\beta}_1 = 0,147$: etxebizitzak oin karratu bat gehiago izaterakoan, gainontzeko aldagaiak konstante mantenduz, espero den salmenta prezioaren batezbesteko gehikuntza 147000 dolarrekoa da.
- $\hat{\beta}_2 = -7,046$: etxebizitzak logela bat gehiago izaterakoan, gainontzeko aldagaiak konstante mantenduz, espero den salmenta prezioaren batezbesteko jaitsiera 7046 dolarrekoa da. Honen arrazoia azalera eta komun kopurua mantentzen direnez, logelak txikiagoak izan beharko dutela, da.
- $\hat{\beta}_3 = -0,264$: etxebizitzak komun bat gehiago izaterakoan, gainontzeko aldagaiak konstante mantenduz, espero den salmenta prezioaren batezbesteko jaitsiera 264 dolarrekoa da. Lehen bezala, logelak txikiagoak izango baitira.

Hipotesien kontrastea

Adibide moduan, etxebizitzaren batezbesteko prezioaren diferentzia etxebizitzak tximinia izateagatik esanguratsua dela kontrastatzeko, FIREPL aldagaiaren banakako esanguratasun kontrastea egin behar da. Kasu honetan, dagokion p-balioa 0,6416 denez, %5eko esanguramailarekin diferentzia ez dela esanguratsua ondorioztatu daiteke.

Aurreko (7.6) eta (7.8) ereduak konparatuz, azken ereduaren erantsitako aldagaiak ez dira banaka adierazgarriak. Halaber, $\bar{R}^2$ txikiagoa da. Erantsitako aldagaien baterako esanguratasuna kontrastatzeko hipotesiak $H_0 : \alpha_3 = \alpha_4 = \beta_2 = \beta_3 = 0$ eta $H_a : \alpha_3 \neq 0$ edota $\alpha_4 \neq 0$ edota $\beta_2 \neq 0$ edota $\beta_3 \neq 0$ dira. Kontrastea hondar karratuen baturetan oinarritutako F -estatistikoa erabiliz egin daiteke, (7.6) murriztutako eta (7.8) murriztu gabeko ereduak izanik.

Kontrastearen emaitza ikusirik:

Hipotesi hutsa: aldagaien erregresio koefizienteak zero dira

```
famroom
firepl
bedrms
baths
```

Estatistikoa: $F(4, 7) = 0,0864517$, p-balioarekin = 0,983881

Ez dugu hipotesi hutsa bazterten %5eko esangura-mailarekin, hau da, (7.6) eremuan erantsitako aldagaiak ez dira batera esanguratsuak. Aldagai horiek omititzen baditugu, lortutako eredua hobeagoa da koefizienteen esanguratasun eta $\bar{R}^2$ koefizientearen aldetik. Beraz, POOL eta SQFT aldagaiak eremuan mantenduz, gainontzekoak (FIREPL, FAMROOM, BATHS, BEDRMS) ez dira etxebizitzaren prezioaren azalpenean lagungarriak.

7.4 Egitura aldaketa

Batzuetan, aldagai azaldua eta azaltzaileen arteko erlazioa laginean zehar aldatu daiteke, hau da, egitura aldaketa bat gerta daiteke. Adibidez, tabakoaren kontsumoa analizatu nahi badugu eta laginean zehar, tabakoaren kalte informatzeko asmoz, kanpaina bat abian jartzen bada, kanpainaren ondorioz tabakoaren eskaera murriztu daitekela pentsa dezakegu. Hala bada, ezin da eskaera funtzio bakarra zehaztu, bi baizik. Bata kanpaina egiten den momenturaino eta bestea, kanpainatik aurrera. Beraz, egitura aldaketaren bat dagoela susmatzen badugu funtzioaren koefizienteen egonkortasuna kontrastatu beharko dugu.

Egitura aldaketaren kontrastea, Chow-en kontrastea deritzona, era desberdinetara burutu daiteke. Alde batetik, hondar karratuen baturetan oinarritutako F estatistikoa erabiliz eta bestaldetik, fikzio-aldagaiak erabiliz.

Ramanathaneko (2002) data7-19 fitxategian 1960 eta 1988 tartean Turkian tabakoaren eskaerari buruzko ikasketa bat egiteko datuak daude:

Q: Tabakoaren kontsumoa kilotan (Ibiltartea 1,86 - 2,723)
 Y: NPG per capita 1968ko Turkiako liritan (Ibiltartea 2560 - 5723)
 P: Tabakoaren salmenta prezioa Turkiako liritan (Ibiltartea 1,361 - 3,968)
 D82: 1982 urtetik 1 balioa hartzen duen fikzio-aldagaia

Tabakoaren eskaera azaltzeko, ondoko eredua zehazten dugu:

$$\ln Q_t = \alpha + \beta \ln Y_t + \gamma \ln P_t + u_t \quad t = 1960, \dots, 1988 \quad (7.9)$$

Hala ere, 1981 urtean Turkiako gobernuak tabakoaren aurkako kanpaina bati hasiera eman zion eta horren ondorioz, tabakoaren kontsumoa aldatu zen. Gure helburua aldaketa hori nabaria izan zen ala ez aztertzea da. Nabaria izan bazen, ondoko ereduak zehaztu beharko genituzke:

$$\ln Q_t = \alpha_1 + \beta_1 \ln Y_t + \gamma_1 \ln P_t + u_{1t} \quad t = 1960, \dots, 1981 \quad (7.10)$$

$$\text{Ln}Q_t = \alpha_2 + \beta_2 \text{Ln}Y_t + \gamma_2 \text{Ln}P_t + u_{2t} \quad t = 1982, \dots, 1988 \quad (7.11)$$

Egitura aldaketarik ez dagoelaren hipotesi hutsa $H_0 : \alpha_1 = \alpha_2, \beta_1 = \beta_2$ eta $\gamma_1 = \gamma_2$ da eta hondar karratuen baturetan oinarritutako F -estatistikoa erabil daiteke kontrastea burutzeko. Kasu honetan, (7.10) eta (7.11) ekuazioek murriztu gabeko ereduaosatzen dute eta (7.9) murriztutako eredua da.

Adibide honetarako Chow-en kontrastea burutuko dugu Gretl erabiliz. Data7-19 fitxategia ireki eta aldagaiak eraldatu ondoren, (7.9) eredua estimatzen dugu KTA bitartez. Estimazioaren emaitzen leihoan *Kontrasteak* klikatu, gero *Chow kontrastea* aukeratu eta *Lagina zatitzeko behaketa* leihoan 1982 idatzi. Estimazio eta kontrastearen emaitza hurrengoa da:

Eredua 1: KTA estimazioak 29 behaketak erabiliz 1960–1988

Aldagai azaldua: LnQ

Aldagaia	Koefizientea	Desb. Tipikoa	t -estatistikoa	p-balioa
const	−4,5898	0,724913	−6,3316	0,0000
LnY	0,688498	0,0947276	7,2682	0,0000
LnP	−0,485683	0,101394	−4,7900	0,0001
Aldagai azalduaren batezbestekoa			0,784827	
Aldagai azalduaren Desb. Tip.			0,108499	
Hondar Karratuen Batura			0,0949108	
Hondarren desbiderazio tipikoa ($\hat{\sigma}$)			0,0604187	
R^2			0,712058	
Zuzendutako $\bar{R}^2$			0,689908	
$F(2, 26)$			32,1480	
Durbin–Watson estatistikoa			1,00057	
Lehen ordenako autokoerlazio koefizientea			0,489867	
Log-egiantza			41,8214	
Akaike Informazio Irizpidea			−77,642	
Schwarz Bayesian Irizpidea			−73,541	
Hannan–Quinn Irizpidea			−76,358	

Chow egitura aldaketaren estatistikoa (1982 behaketan) –

Hipotesi hutsa: egitura aldaketarik ez

Estatistikoa: $F(3, 23) = 20,1355$

p-balioarekin = $P(F(3, 23) > 20,1355) = 1,25619\text{e-}006$

Emaitzetan ikusten dugunez, kalkulaturako F -estatistikoa $20,1355 > F_{(3,23)0,05}$ da, beraz hipotesi hutsa baztertzen dugu %5eko esangura-mailarekin. Honen ondorioz, Turkiako gobernuak hasitako tabakoaren aurkako kanpainak eragina izan zuela kontsumoan esan dezakegu eta (7.10) eta (7.11) ekuazioek azaltzen dutela tabakoaren eskaera. Estimazioaren emaitzak ondokoak dira:

$$\widehat{\text{Ln}Q}_t = -5,024 + 0,735 \text{Ln}Y_t - 0,381 \text{Ln}P_t \quad t = 1960, \dots, 1981 \quad R^2 = 0,9456$$

$(t_{est}) \quad (-10,614) \quad (11,587) \quad (-4,227)$

$$\widehat{LnQ}_t = 8,837 - 0,953 LnY_t + 0,108 LnP_t \quad t = 1982, \dots, 1988 \quad R^2 = 0,6203$$

$(t_{est}) \quad (2,170) \quad (-1,941) \quad (0,654)$

Egitura aldaketaren kontrastea fikzio-aldagaiak erabiliz

Aurreko kontrastea D82 fikzio-aldagaia erabiliz burutu daiteke. Horretarako, hurrengo eredua zehazten dugu:

$$LnQ_t = \beta_1 + \beta_2 LnY_t + \beta_3 LnP_t + \beta_4 D82_t + \beta_5 D82_t \cdot LnY_t + \beta_6 D82_t \cdot LnP_t + u_t \quad (7.12)$$

$$t = 1960, \dots, 1988$$

non egitura aldaketarik ez dagoelaren hipotesi hutsa $H_0 : \beta_4 = \beta_5 = \beta_6 = 0$ den. Hondar karratuen baturetan oinarritutako F -estatistikoa erabil daiteke kontrastea burutzeko, (7.12) murriztu gabeko eredua izanik eta

$$LnQ_t = \beta_1 + \beta_2 LnY_t + \beta_3 LnP_t + u_t \quad t = 1960, \dots, 1988 \quad (7.13)$$

murriztutako eredua.

Ondoren, Gretlen jarraitu behar diren pausuak egitura aldaketaren kontrastea burutzeko aztertuko dugu. Aldagaiak eraldatu ondoren⁵, (7.12) eredua estimatzen dugu KTAren bitartez. Estimazioaren emaitzen leihuan *Kontrasteak* klikatu, gero *Omititu aldagaiak* eta *D82*, *D82 × LnY* eta *D82 × LnP* aldagaiak aukeratzen ditugu. Kontrastearen emaitza ondokoa da:

Eredua 2: KTA estimazioak 29 behaketak erabiliz 1960–1988

Aldagai azaldua: LnQ

Aldagaia	Koefizientea	Desb. Tipikoa	t-estatistikoa	p-balioa
const	−5,0248	0,541262	−9,2837	0,0000
LnY	0,735837	0,0726077	10,1344	0,0000
LnP	−0,381857	0,103289	−3,6970	0,0012
D82	13,8623	2,85197	4,8606	0,0001
D82LnY	−1,6892	0,345466	−4,8898	0,0001
D82LnP	0,490526	0,153989	3,1855	0,0041

Aldagai azalduaren batezbestekoa	0,784827
Aldagai azalduaren Desb. Tip.	0,108499
Hondar Karratuen Batura	0,0261724
Hondarren desbiderazio tipikoa ($\hat{\sigma}$)	0,0337332
R^2	0,920598
Zuzendutako $\bar{R}^2$	0,903336
$F(5, 23)$	53,3328
Durbin–Watson estatistikoa	2,02153
Lehen ordenako autokoerlazio koefizientea	−0,0136939
Log-egiantza	60,5008
Akaike Informazio Irizpidea	−109,00
Schwarz Bayesian Irizpidea	−100,79
Hannan–Quinn Irizpidea	−106,43

⁵Adibidez, LnQ sortzeko Gretlen, *Aldagaia* → *Definitu aldagai berria...* klikatu eta agertzen den kaxan $LnQ = \ln(Q)$ idatzi.

Omititutako aldagaien kontrastea –

Hipotesi hutsa: aldagai guztientzat parametroak zero dira

D82

D82LnY

D82LnP

Estatistikoa: $F(3, 23) = 20,1355$

p-balioarekin = $P(F(3, 23) > 20,1355) = 1,25619\text{e-}006$

Emaitzetan ikus daitekenez, p-balioa $1,25619\text{e-}006$ da eta hipotesi hutsa baztertzen dugu %5eko esangura-mailarekin. Beraz, Turkiako gobernuak hasitako tabakoaren aurkako kanpainak eragina izan zuen kontsumoan. Honen ondorioz, (7.10) eta (7.11) ekuazioek azaltzen dute tabakoaren kontsumoa.

Bibliografia

Ramanathan, R. (2002), *Introductory Econometrics with applications*, 5. ed., South-Western, Ohio.

A Eranskina: Errepasoak

A.1 Probabilitatearen errepasoa

Zorizko saiakuntza baten ondorioz lortutako datuen interpretazioan oinarritzen dira metodo ekonometrikoak. Aldagai ekonomikoek zati edo osagai sistematiko bat eta aleatorio bat izaten dute, bere behaketa eman aurretik ezin baita zehaztasun osoz hartuko dituzten balioak aurrezan. Atal honetan, kurtsoan zehar aplikatuko ditugun probabilitate kontzeptuak gogoratuko ditugu: zorizko aldagaia (*aleatorioa edo estokastikoa*) eta bere propietateak, baita probabilitate banaketa erabilgarrienak ere.

A.1.1 Zorizko aldagaia edo aldagai aleatorioa

Aldagai aleatorio baten (X) balioa ez da ezagutzen bere behaketa eman aurretik eta probabilitatea, emaitzaren ziurgabetasuna adierazteko baliabide bat da. Bi motatako aldagai aleatorioak ezagutzen dira: *diskretuak*, har ditzaken balio posible guztien multzoa finitua edo zenbagarria denean eta *jarraiak*, egite multzoa infinitoki zatigarria denean eta beraz, ez zenbagarria. Adibidez, etxebizitza baten azalera aldagai jarraia da baina komun kopurua berriz, diskretua. Orokorrean, kurtso honetan aldagai jarraiekin lan egingo dugu.

Baldin eta X aldagai diskretua bada, emaitza posible bakoitzari (x_i) probabilitate bat ezarri diezaiokegu $p(x_i) = \text{Prob}(X = x_i)$. *Probabilitate funtzioak* $p(x_i) \geq 0 \forall x_i$ eta $\sum_i p(x_i) = 1$ bete beharko ditu.

Baldin eta X jarraia bada, puntu zehatz bati dagokion probabilitatea zero da eta X aldagaiaren balioak tarte batean egotearen probabilitateari buruz jardungo dugu, $a \leq X \leq b$.

Horrela, X aldagai jarrai baten *dentsitate funtzioak*, $f(x)$, ondorengo betetzen du:

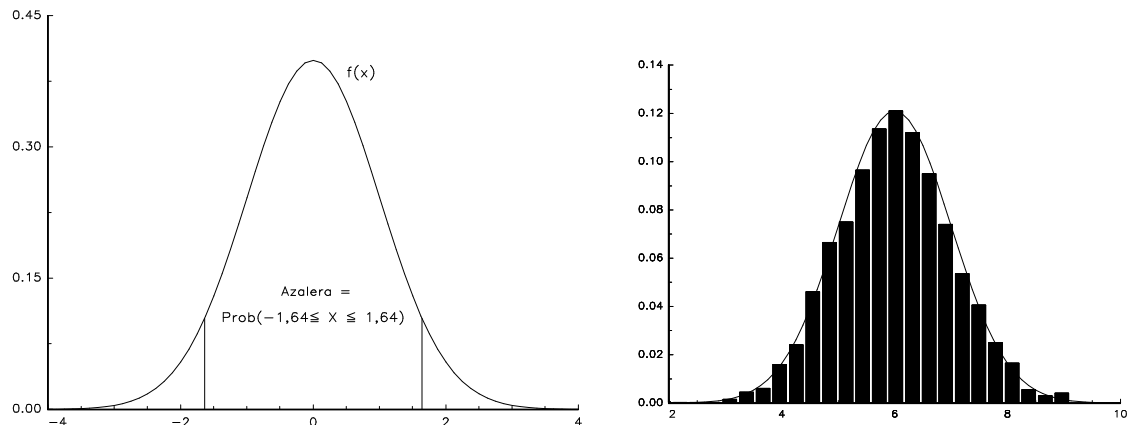
$$\text{Probabilitatea}(a \leq X \leq b) = \int_a^b f(x) dx$$

Hau da, a eta b bi puntuen artean funtzioaren azpian gelditzen den azalera, aldagaiak hartuko dituen balioak $[a, b]$ tartean egotearen probabilitatea da (ikusi A.1 grafikoko ezkerreko panela). Dentsitate funtzioak $f(x) \geq 0 \forall x$ eta $\int_{-\infty}^{\infty} f(x) dx = 1$ bete beharko du.

Aldagai aleatorio jarrai baten adibide bat **aldagai Normala** da. Bere dentsitate funtzioak kanpai forma du (ikusi A.1 grafikoko ezkerreko panela). Erdiko balio baten inguruan simetriko koki banatzen diren aldagaiak egituratzeko praktikan asko erabiltzen den funtzioa da, erdiko

balio horren inguruko balioetan probabilitate handiena pilatzen delarik eta gutxi urruneko balioetan. A.1 grafikoko eskubiko panelak datuen histograma eta dentsitate funtzioaren ar-

A.1 Irudia: *Normalaren* dentsitate funtzioa eta histograma.



teko erlazioa irudikatzen du. Peña eta Romo (1997) adierazten duten bezala: “*La función de densidad constituye una idealización de los histogramas de frecuencia o un **modelo** del cual suponemos que proceden las observaciones. El histograma representa frecuencias mediante áreas; análogamente, la función de densidad expresa probabilidades por áreas. Además, conserva las propiedades básicas del histograma: es no negativa y el área total que contiene es uno.*”.

Aldagai aleatorio baten banaketa funtzioa, lekuko neurriak (batezbestekoa, mediana eta moda), sakabanatze neurriak (bariantza, desbideratze tipikoa eta aldakuntz koefizienteak) edota forma neurriak (asimetria eta kurtosis koefizienteak) erabiliz laburbildu daiteke. Kontzeptu hauek, datu multzoen ezaugarriak laburbiltzeko erabilitako neurrien antzera definitzen dira. Jarraian, kurtsoan zehar erabiliko ditugun beste elementuak azalduko ditugu.

X aldagai baten **batezbestekoa edo itxarondako balioa** (μ) lekuko neurri bat da. Aldagai horrek har ditzaken balio posible guztien batezbesteko ponderatua da, non ponderazioa balio bakoitzaren probabilitatea den. Aldagaia jarraia bada:

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

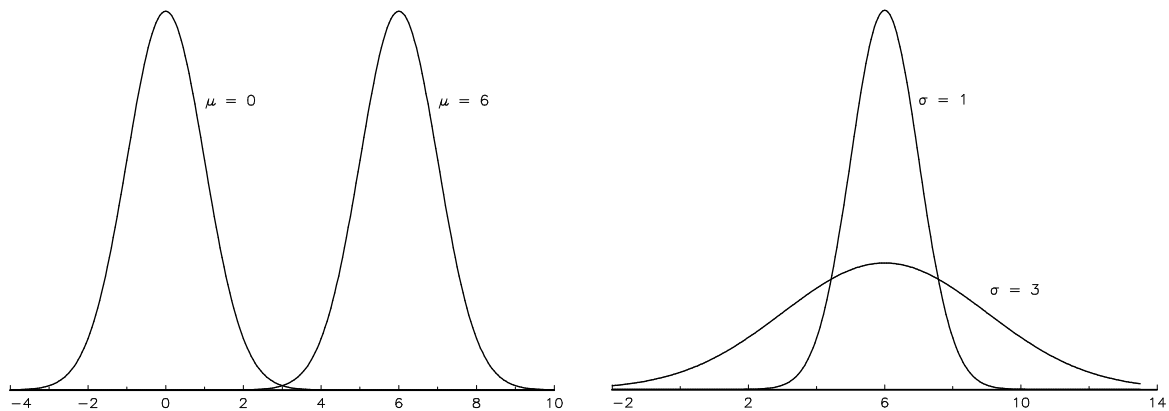
non E , itxaropenaren eragilea den. A.2 grafikoko eskubiko panelan ikusi daitekenez, itxaropen horrek aldagaiaren banaketaren zentrua jasotzen du. Horrela, itxaropena gero eta handiagoa bada, saiakuntzaren gauzatutako balioak handiagoak izatea espero da.

X aldagai aleatorio baten **bariantza**, bigarren ordenako batezbestekoarekiko momentu zentratua da. Hau da:

$$Var(X) = \sigma_X^2 = E[(X - \mu)^2] \geq 0$$

Banaketaren sakabanatze neurri bat da eta bere erro positiboa **desbideratze tipikoa** edo

A.2 Irudia: Banaketa normalaren adibideak

Batezbesteko desberdinak eta $\sigma = 1$ Dispersio desberdinak eta $\mu = 6$ 

desbideratze estandarra bezala ezagutzen da:

$$desb(X) = \sigma_X = \sqrt{Var(X)}$$

A.2 grafikoko eskubiko panelak adierazten duen bezala, aldagaiaren bariantza gero eta txikiagoa bada, batezbestekoaren inguruko probabilitatea handiagoa da.

Normal estandar edo sinplearen banaketa. Banaketa Normala bere batezbesteko eta bariantzaren balioengatik zehazten da. Baldin eta Z aldagai aleatoria zero batezbestekodun eta bat barintzaduneko Normala bada, orduan Z aldagai Normal estandarra dela esaten da eta $Z \sim N(0, 1)$ bezala adierazten da. Badira banaketa honen taulak, non x emaitz posible bakoitzari, puntu horretaraino metatutako probabilitatea egokitzen zaion, $Prob(X \leq x)$.

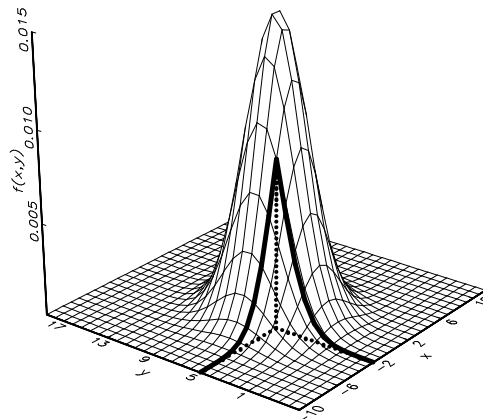
Orokorki, X aldagai Normala bada, orduan μ batezbestekoarekin eta σ^2 bariantzarekin, $X \sim N(\mu, \sigma^2)$ bezala adierazten da. $Z = (X - \mu)/\sigma$ eraldaketa Normal estandarra izango denez, banaketa Normal honen taularekin $Prob(X \leq x)$ metatutako probabilitatea lortzen da.

A.1.2 Zorizko bi aldagai edo gehiago

Ekonometrian aldagai multzo baten propietateak ikertu nahi dira eta zorizko bi aldagai edo gehiagoei buruzko galderei erantzuteko, baterako dentsitate funtzioa ezagutu behar da. Bi aldagaiak (X eta Y) diskretuak badira, (x_i, y_j) emaitza posible bakoitzari probabilitate bat egokitu diezaiokegu, $p(x_i, y_j)$. Probabilitate multzoa *baterako probabilitate funtzioa da*, non $0 \leq p(x_i, y_j) \leq 1$ eta $\sum_i \sum_j p(x_i, y_j) = 1$ betetzen den.

Aldagaiak jarraiak badira, baterako banaketa *baterako dentsitate funtzioarekin* ($f(x, y)$) lortzen da. Bi aldagaiek banaketa Normala jarraitzen badute, baterako densitate funtzioa ondorengo grafikoan agertzen da:

A.3 Irudia: Aldagai biko banaketa normala.



Funtzio honen azpiko azalera totala probabilitate masa deitzen da eta bere balioa bat izango da, hau da, $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$. Gainera, funtzioak ez duenez balio negatiborik hartzen ($f(x, y) \geq 0$), (a, b) puntuen artean definitutako laukiak zehazten duen azalerak, X aldagaiak a baino balio handiagoak eta Y adagaiak b baino balio txikiagoak hartzeko probabilitatea neurtzen du. Hau da:

$$Probabilitatea(X \leq a, Y \leq b) = \int_{-\infty}^a \int_{-\infty}^b f(x, y) dx dy$$

Adibide gisa, A.3 grafikoan markatutako azaleraren azpian jasotako bolumena, $X \leq -2$ eta $Y \leq 4.5$ izatearen probabilitatea da.

Hala ere, aldagai bakoitzaren **dentsitate funtzio marginala** ondorengo integralekin kalkulatzen daiteke:

$$f(x) = \int_{-\infty}^{\infty} f(x, y) dy \quad f(y) = \int_{-\infty}^{\infty} f(x, y) dx \quad (\text{A.14})$$

Zorizko bi aldagaien baterako banaketa horrela laburbildu daiteke:

- Aldagai bakoitzaren grabitate zentruak, hau da, μ_X eta μ_Y , (A.14) banaketa marginaletatik lortzen dira.
- Aldagai bakoitzaren sakabanatzea batezbestekoaren inguruan, adibidez X eta Y -ren bariantzak (σ_X^2 eta σ_Y^2), (A.14) banaketa marginaletik eratortzen dira.
- Kobariantzaren bitartez bi aldagaien arteko erlazioa jasotzen da:

$$Kob(X, Y) = \sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] \quad (\text{A.15})$$

edota aldagaien koerlazio koefizientearen bitartez

$Koer(X, Y) = \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \in [-1, 1]$. Kobariantzak eta zorizko aldagaien koerlazioak, datuekin egindako antzeko interpretazioa dute. Horrela, $\sigma_{XY} = \rho_{XY} = 0$ bada, X eta Y aldagaiak koerlatugabeak direla esaten da.

Baterako banaketa, batezbesteko bektorean (μ) eta Σ bariantza eta kobariantza matrizean laburtzen da:

$$\mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix} \quad \Sigma = \begin{pmatrix} Bar(X) & Kob(X, Y) \\ Kob(X, Y) & Bar(Y) \end{pmatrix} = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix} \quad (\text{A.16})$$

Baldintzatutako banaketak. Aldagai multzo bat aztertzean, gertaera bat egin ondoren beste gertaera bat betetzearen posibilitatea ebaluatzea komeni da. Adibidez, zein da ezkondu-tako eta seme-alabak dituen emakume batek lan merkatuan parte hartzearen probabilita-tea? **Baldintzatutako probabilitateak** horrelako galderei erantzuteko aukera ematen du. Aldagaiak diskretuak badira, X aldagaia x_i balioa denean, Y aldagaiaren baldintzatutako banaketa horrela definitzen da:

$$Prob(Y = y_j | X = x_i) = \frac{Prob(Y = y_j, X = x_i)}{Prob(X = x_i)} = \frac{p(y_j, x_i)}{\sum_j p(y_j, x_i)} \quad \text{non } Prob(X = x_i) > 0$$

Aldagaiak jarraiak badira, X aldagaiak x_i balioa hartzen duelarik ($f(x) > 0$ kasuetan) baldintzatutako Y aldagaiaren dentsitate funtzioa horrela definitzen da:

$$f(y|X = x) = \frac{f(y, x)}{f(x)}$$

Era honetan banaketa berri bat lortzen da, jada ikusitako propietateekin. Banaketa honen momentu interesgarriak, $X = x$ balioari baldintzatutako Y -ren batezbestekoa eta bariantzak dira eta $E(Y|X = x)$ eta $Bar(Y|X = x)$ bezala izendatzen dira.

Independentzia. Bi aldagai aleatorio X eta Y estatistikoki independenteak dira edo independenteki banatuak daude, baldin eta aldagai batek hartzen duen balioa ezaguturik, ez badu besteak har dezaken balioaren informaziorik ematen. X eta Y aldagaiak independenteak badira, orduan baterako dentsitate funtzioa atalbanatu daiteke jarraian adierazten den bezala:

$$f(x, y) = f(x)f(y) \quad -\infty < x, y < \infty$$

Gainera, $f(y|X = x) = f(y)$ betetzen da. Bestalde, X eta Y independenteak badira, orduan $Kob(X, Y) = 0$ izango da. Azkenik, X eta Y aldagaiak batera Normalki banatzen badira eta $Kob(X, Y) = 0$, orduan X eta Y independenteak dira.

Bi aldagai baino gehiago daudenean, aurreko emaitzak n aldagai aleatorioen multzoari orokortu daitezke, hau da, $X_1, X_2, \dots, X_n$ izanik, bektore batean biltzen dira:

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}.$$

Orokorrean, bere baterako banaketaren momentuak **batezbestekoen** bektorea

$$E(\mathbf{X}) = \mu = \begin{pmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_n) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix}$$

da eta **bariantza eta kobariantza** matrizea Σ_X da:

$$\Sigma_X = \begin{pmatrix} Bar(X_1) & Kob(X_1, X_2) & \dots & Kob(X_1, X_n) \\ Kob(X_1, X_2) & Bar(X_2) & \dots & Kob(X_2, X_n) \\ \vdots & \vdots & & \vdots \\ Kob(X_1, X_n) & Kob(X_2, X_n) & \dots & Bar(X_n) \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} & \dots & \sigma_{1,n} \\ \sigma_{1,2} & \sigma_2^2 & \dots & \sigma_{2,n} \\ \vdots & \vdots & & \vdots \\ \sigma_{1,n} & \sigma_{2,n} & \dots & \sigma_n^2 \end{pmatrix}$$

non Σ_X matrizea, n ordenako matrize karratua eta definitu ez negatiboa den. Honek, diagonal nagusiko elementuak ez negatiboak direla adierazten du, $\sigma_i^2 \geq 0$, $\forall i$.

Aldagaiak elkarrekiko independenteak badira, orduan elkarrekiko koerlatugabeak dira, hau da, $\sigma_{i,j} = 0, \forall i \neq j$, eta hortaz Σ_X diagonal izango da:

$$\Sigma_X = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n^2 \end{pmatrix}$$

Gainera, $X_1, \dots, X_n$ aldagaiek banaketa berdina jarraitzen badute, batezbesteko eta bariantza berdinekin

$$E(\mathbf{X}) = \begin{pmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{pmatrix} \quad \Sigma_X = \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix} = \sigma^2 \mathbf{I}$$

orduan aldagai aleatorioak identikoki eta independenteki banatuta egongo dira μ batezbesteko eta σ^2 bariantzarekin: $X_i \sim iib(\mu, \sigma^2), \forall i = 1, \dots, n$.

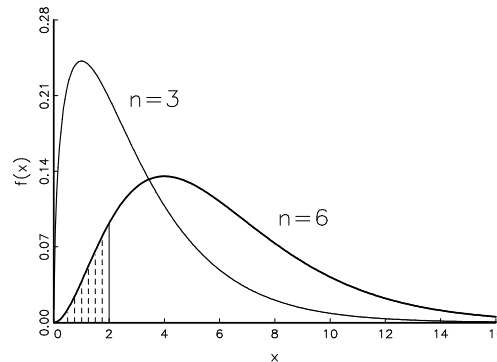
Azkenik, $X_1, \dots, X_n$ zorizko aldagai Normalak badira, $\mathbf{X}$ bektoreak **Banaketa Normal anizkoitza** jarraitzen du eta bere batezbesteko bektorea eta bariantza eta kobariantza matrizeagatik zehazten da, $\mathbf{X} \sim N(\mu, \Sigma_X)$. Aldagaiak independenteak badira, batezbesteko eta bariantza berdinarekin, $X_i \sim NIB(\mu, \sigma^2)$ idatziko dugu.

A.1.3 Zenbait probabilitate banaketa

Kurtsoan zehar, aipaturiko banaketa Normalaz gain, berekin erlazionaturiko beste banaketa batzuk erabiliko ditugu.

Chi-karratu banaketa. $(Z_1, \dots, Z_n)$, zero batezbestekoa eta bat bariantzako banaketa Normaldun zorizko aldagai independenteak badira, hau da $Z_i \sim NIB(0, 1)$, orduan $X =$

A.4 Irudia: Chi karratu banaketaren dentsitate funtzioa



$\sum_{i=1}^n Z_i^2$ aldagaia, n askatasun graduko chi-karratu banaketako zorizko aldagaia da eta $X \sim \chi^2(n)$ adierazten da. Bere dentsitate funtzioak ondoko itxura du:

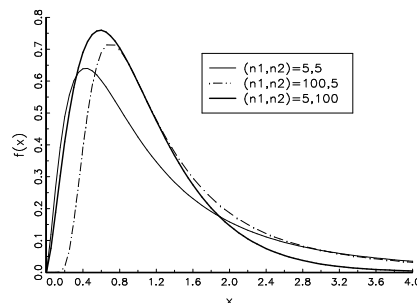
X aldagaiaren balio negatiboentzat $f(x) = 0$ dela ikusten da. Banaketa asimetrikoa da, batezbestekoa n da eta bariantza $2n$. Badira puntu bateraino metatutako probabilitatea ($Prob(X \leq x)$) jasotzen dituen taulak, hau da, n askatasun graduen funtzioan egindako grafikoko marraztutako azalera.

Snedecorren-F banaketa. Izan bitez $Z_1 \sim \chi^2(n_1)$ eta $Z_2 \sim \chi^2(n_2)$ independenteki banatutako aldagaiak, orduan $X = (n_2/n_1)(Z_1/Z_2)$ aldagaiak, n_1 eta n_2 askatasun graduko Snedecorren F banaketa jarraituko du:

$$X = \frac{Z_1/n_1}{Z_2/n_2} \sim \mathcal{F}(n_1, n_2)$$

Hurrengo grafikoak, askatasun gradu desberdinekin dentsitate funtzioak jasotzen ditu:

A.5 Irudia: Snedecorren-F banaketaren dentsitate funtzioa



Zuzen errealearen alde positiboan ($x > 0$) metatzen da probabilitatea. Izendatzailearen aska-

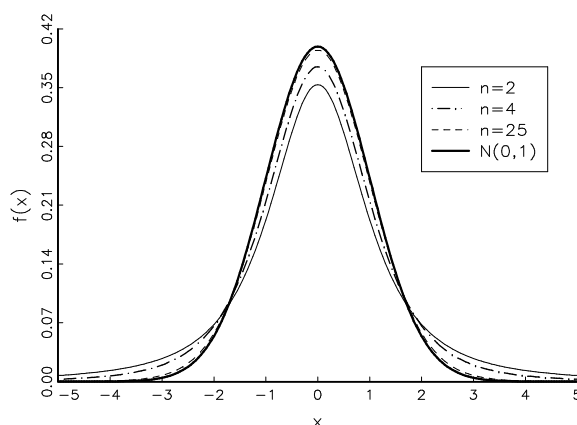
tasun graduak handitzen diren heinean, $n_2 \rightarrow \infty$, orduan $n_1 \mathcal{F}(n_1, n_2)$ -ren banaketak $\chi^2(n_1)$ banaketara jotzen du.

Student- t banaketa. Izan bitez $Z \sim N(0, 1)$ eta $Y \sim \chi^2(n)$ aldagaiak independenteak, orduan $X = Z/\sqrt{Y/n}$ aldagai aleatorioak, n askatasun graduako Student- t banaketa izango du:

$$Y = \frac{X}{\sqrt{Z/n}} \sim t_{(n)}$$

Hurrengo grafikoan, Student- t banaketaren dentsitate funtzioaren zenbait adibide jasotzen dira banaketa normal estandarrakin batera:

A.6 Irudia: Studenten- t banaketaren dentsitate funtzioa



Banaketa guztiak zero balioaren inguruan simetrikoki banatzen dira. Banaketaren batezbestekoa zero da $n > 1$ bada eta $n > 2$ denean, bere bariantza $n/(n-2)$ da. Student- t banaketaren buztanak, Normalarenak baino lodiagoak dira baina askatasun graduak handitzen diren heinean, Normal estandarrera konbergitzen du.

A.2 Inferentzia estatistikoaren errepeaso

Suposa dezagun lizentziatu berrien batezbesteko soldata aztertu nahi dela. Populazioa edo banakoen multzoa zabala denez, lagin bat aukeratzen da, hau da, zoriz aukeratutako lizentziatu berrien azpimultzo bat. Populazioa ikerketaren helburu diren elementuez osatutako multzoa da (herrialde bateko familiak, lizentziatu berriak, herri bateko etxebizitzak, enpresa bateko bezeroak, etab.) Datuak ondo erabiliz, posible da susmoen azterketa sakon bat egitea eta lagin honetako emaitzak ondorioz, populaziora orokortzea.

Inferentzia estatistikoaren helburu bat lagin bat aztertuz populazioaren ezaugarriak ezagutzea da. *Populazioa*, aztertu nahi den elementuetaz osatutako eta ondo definitutako multzoa da.

Lagina berriz, populazioaren azpimultzo adierazgarri bat da.

Populazioa zehaztu ondoren, interesatzen zaizkigun ezaugarriak jasotzen dituzten datuentzat eredu bat zehaztu behar da. Ekonometrian $y_1, y_2, y_3, \dots, y_N$, Θ parametro ezezagunen menpeko baterako banaketako N aldagai aleatorioen gauzatzeak direla suposatzen da. **Eredu** batek, banaketa eta Θ parametro bektorearen ezaugarri nagusiak jasotzen ditu. Adibidez, hiri bateko etxebizitza baten metro karratuko batezbesteko prezioa ezagutzea interesatzen bazailugu, suposa dezagun jaso ditugun 50 etxebizitzaren metro karratuko prezioak ($y_1, y_2, y_3, \dots, y_{50}$) independenteki eta identikoki banatutako aldagai normalen gauzatzeak direla. Hortaz, datuentzat zehaztutako eredua honako da:

$$Y_i \sim NID(\mu, \sigma^2)$$

Banaketa zehazten duten parametroak, hau da, metro karratuko prezioaren batezbestekoa eta bariantza, ($\Theta = (\mu, \sigma^2)$), ezezagunak dira. Gure ikerketan interesatzen zaiguna batezbestekoa denez, datuak erabiliz, beregaineko informazioa ondorioztatu nahi izango dugu.

Horretarako, estatistikako bi erraminta aplikatuko ditugu: estimazioa eta hipotesien kontrasteak. Estimazioarekin interesatzen zaigun parametroen balio posibleak kalkulatu ahal izango dugu eta hipotesien kontrasteekin berriz, populazioaren hipotesi edo susmo bat ezarri eta gero, datuak aztertuz susmo hori egiazkoa den edo ez ondorioztatu ahal izango dugu.

A.2.1 Estimazioa

Orain arte saiakuntza baten ondorioz datuak nola sortzen diren ikusi dugu. Datu hauek inferentzian erabiliko ditugu populazioko usteak eta erlazioak aztertzeko. Ekonometrian $y_1, y_2, \dots, y_N$ datuak, $f_{\Theta}(Y_1, Y_2, \dots, Y_N)$ -rekin izendatzen dugun Θ parametro desberdinen menpeko baterako banaketako N zorizko aldagaien egiteak edo gertaerak direla suposatzen da. Datuentzat zehaztutako *eredu* batek, banaketaren ezaugarri nagusiak jasotzen ditu, Θ parametro ezezagunen bektorearekin batera.

Adibideko 50 etxebizitzaren metro karratuko prezioaren datuak ($y_1, \dots, y_N$), independenteki banatutako aldagai Normalen egiteak badira, datuentzat zehaztutako eredua

$$Y_i \sim NIB(\mu, \sigma^2)$$

da. Banaketako parametro ezezagunak (batezbestekoa eta bariantza $\Theta = (\mu, \sigma^2)$), datuekin estimatu daitezke, lagineko datuen batezbesteko aritmetikoa eta lagin bariantza edo dispersioa ($\bar{y}$ eta S_y^2) erabiliz adibidez.

Estimazioaren helburua parametro ezezagun multzoaren balioei hurbiltzea da, lagineko behaketen bitartez. Parametro ezezagun bat θ bezala adieraziko dugu eta K parametro ezezaguneko bektorea $\Theta = (\theta_1, \theta_2, \dots, \theta_k)'$ bezala. **Estatistiko** bat, $g(y_1, \dots, y_N)$ datuen funtzio bat da eta θ -ren **puntuazko estimatzaile bat**. Parametro ezezagunaren hurbilketarentzat estatistiko bat, $\hat{\theta}$ adierazten delarik. Adibidez, lagineko batezbesteko aritmetikoa, aldagai aleatorioaren batezbestekoaren estimatzailea izan daiteke eta lagineko bariantza berriz, aldagaiaren bariantzaren estimatzailea. Hau da,

$$\hat{\mu} = \bar{y} = \frac{1}{N} \sum_{i=1}^N y_i \qquad \hat{\sigma}^2 = S_y^2 = \frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{y})^2$$

Behaketak izan aurretik osatutako funtzioa da estimatzailea eta datuak funtzio horretan barneratzerakoan lortzen den balioa *estimazioa* da. Horrela, gure adibideko etxebizitzaren metro karratuko batezbesteko prezioaren (*Prezioa/m2*) estimazioa ondorengo izango da:

$$\hat{\mu} = \frac{3,82 + 5,246 + \dots + 3,434 + 4,20}{50} = 3,91 \text{ mila euro}$$

Hau da, etxebizitza baten batazbesteko prezio estimatua 3910 euro metro karratuko ingurukoa izango da. Hala ere, emaitza honetan nolako konfidantza izan dezakegu? Kantitate berbera berdin baloratuko genuke 5 behaketako lagin batekin lortu izan bagenu? Emaitza noski EZ da, 50 behaketekin lortutako emaitza fidagarriagoa baita 5ekin lortutako baino. Beraz, estimatzaile bat (eta bere estimazioak) zehaztasun edo fidagarritasun neurri batekin batera aztertu behar da.

Estimatzaile bat, Y_i , $i = 1, \dots, N$ aldagaien menpeko zorizko aldagaia da. Bere probabilitate banaketa lagineko banaketa edo estimatzailearen banaketa enpirikoa bezala ezagutzen da. Aurreko adibidean, $Y_i \sim NIB(\mu, \sigma^2)$ bada, orduan $\hat{\mu} = \bar{y}$ estimatzailea N aldagai Normal independenteen konbinazio lineal bat denez, bere lagineko banaketa hurrengo da:

$$\hat{\mu} = \bar{y} \sim N(\mu, \sigma^2/N) \tag{A.17}$$

Lagineko batezbestekoa populazioko batezbestekoaren inguruan banatzen da eta lagingaren tamaina (N) gero eta handiagoa bada (hau da, bariantza txikiagoa), μ -ren inguruan probabilitate handiagoa bilduko da. Hortaz, μ -ren hurbileko estimazio bat lortzeko probabilitatea handiagoa da 50 behaketekin, 5ekin baino. Kasu honetan, zentzuduna da *zehazpenaren* indikatzaile moduan desbideratze tipikoa ($\sigma/\sqrt{N}$) erabiltzea, desbideratze tipiko txikiagoak zehaztasun handiagoa adierazten duelarik. Hala ere, normalean σ ezezaguna izaten denez, populazioko balioa laginekoarekin ordezkatzeko dugu, hau da S_y . Hortaz, $\bar{y}$ -ren lagineko banaketaren desbideratze tipikoaren estimazioa honakoa izango da:

$$S_{\bar{y}} = S_y/\sqrt{N}$$

$\bar{y}$ -ren *errore tipiko* bezala ezagutzen delarik. etxebizitzaren prezioen adibidean lortzen den estimazioaren errore tipikoa $0,993341/\sqrt{50} = 0,14$ da. Bestalde, erraz frogatu daiteke 5 behaketako lagin batekin $\bar{y}$ eta S_y lortuko bagenu, errore tipikoa hirukoiztu egingo litzatekeela: $S_{\bar{y}} = 0,993341/\sqrt{5} = 0,44$.

Estimatzaileak konparatzeko irizpideak

Lehen aipatu bezala, estimatzaileak ugari dira eta noski, batzuk besteak baino hobeagoak. Orokorrean, benetako balioei gehien hurbiltzen diren estimatzaileen bila gabiltza. Estimatzailerik, θ eta $\hat{\theta}$ -ren arteko distantzian oinarritzen diren zenbait propietate bete ditzaten espero

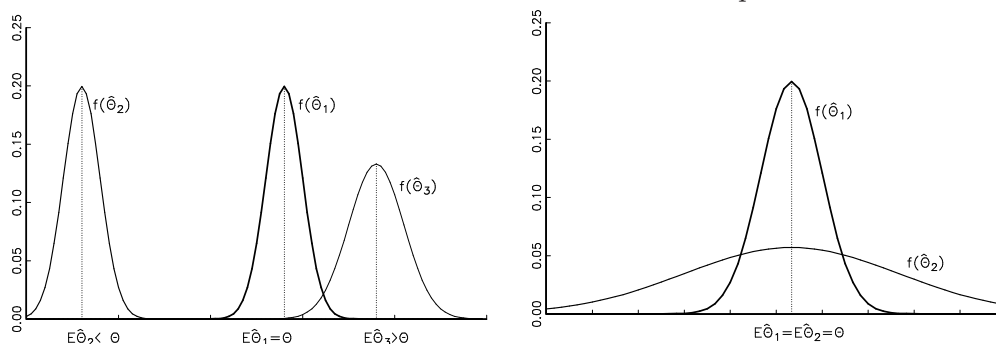
dugu. Horrela, lagin finituetako propietateak, estimatzaileen ezaugarriak lagineko tamainarekiko independenteki konpara daitezkenak dira. Asintotikoek ordea, portaerak konparatzen dituzte baina estimatzaileak kalkulatzaren behaketa kopurua *handia* denean. Kurtso honetan, estimatzaileen hautaketa lagin finitu edo txikietako propietateen arabera egingo dugu, hau da, alboragabetasuna, efizientzia eta batezbesteko errore kuadratikoko minimoaren arabera.

Alboragabetasuna. Estimatzailer bat alboragabea da baldin eta bere lagineko banaketaren itxaropena, parametroaren benetako balioa baldin bada, hau da:

$$E(\hat{\theta}) = \theta$$

Infinitu lagin lortuz gero eta lagin bakoitzarekin estimazioa kalkulatu, estimazio guzti hauen batezbestekoa parametroaren benetakoa balioa izango litzateke. Hau da, batezbestekoz, estimazioaren errorea ($\hat{\theta} - \theta$) ezereztatzen da. Kontrako kasuan, estimatzailea alboratua izango da. Horrela, estimatzaile baten alborapena $Alborapena(\hat{\theta}) = (\hat{\theta}) - \theta$ bezala definitzen da. Adibide gisa, A.7 grafikoaren ezkerreko zatiak θ parametroaren hiru estimatzaileen banaketak jasotzen ditu: $\hat{\theta}_1$ estimatzailea alboragabea da; $\hat{\theta}_2$ estimatzaileak alborapen negatiboa du, hau da, batezbestekoak parametroaren balioa azpiestimatzeko du eta azkenik, $\hat{\theta}_3$ estimatzailearena positiboa da, hau da, batezbestekoz, estimatzaileak parametroaren balioa gainestimatzeko du.

A.7 Irudia: Estimatzaileen alborapena



Efizientzia. Estimatzailer alboragabeetan bakarrik oinarritzen bagara, estimatzaileen artean bat aukeratzeko irizpide bat finkatu behar dugu. Adibidez, A.7 grafikoaren eskubiko aldean bi estimatzaile alboragabeen banaketak adierazten dira. Garbi ikus daitekeenez, $\hat{\theta}_1$ estimatzailea bariantza txikienekoa dela eta hortaz, parametroaren benetako balioaren *urrunko* balioak lortzeko probabilitate txikiagoa du. Horregatik, $\hat{\theta}_1$ estimatzailea $\hat{\theta}_2$ baino efizienteagoa dela esango dugu.

Orokorrean, klase bereko estimatzaileen artean bariantza txikiena duena, *klase hortako efizientea* dela esaten da. Horrela, $\hat{\theta}$ estimatzailea, alboragabeen artean efizientea dela esango dugu, baldin eta ez badago bariantza txikiagodun beste estimatzaile ($\tilde{\theta}$) alboragaberena bat.

$$Bar(\tilde{\theta}) \geq Bar(\hat{\theta}) \quad \forall \tilde{\theta} \text{ alboragabea}$$

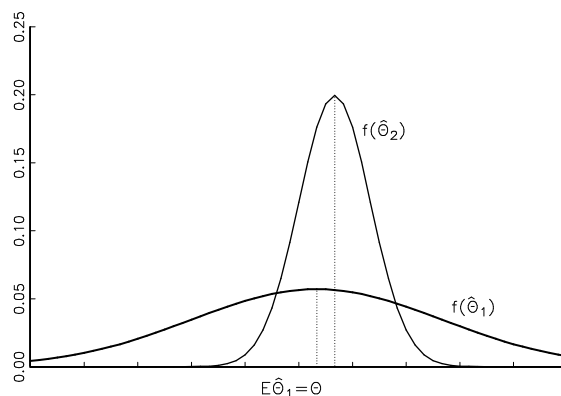
Adibidez, datuen batezbesteko aritmetikoa, aldagai Normal baten populazioko batezbestekoaren (μ) estimatzaile alboragabeen artean, efizientea da. Hau da, $Y_i \sim NIB(\mu, \sigma^2)$ bada, orduan μ -ren estimatzaile alboragabe guztientzat ($\tilde{\mu}$ non $E(\tilde{\mu}) = \mu$) ondorengoa beteko da:

$$Bar(\bar{y}) = \frac{\sigma^2}{N} \leq Bar(\tilde{\mu})$$

Azkenik, K parametrodun multzo bat (Θ) estimatzean, $\hat{\Theta}$ estimatzaile alboragabe bat beste estimatzaile alboragabe bat ($\tilde{\Theta}$) baino efizienteagoa dela esango dugu baldin eta $[Bar(\tilde{\Theta}) - Bar(\hat{\Theta})]$ matrizeen diferentzia matrize erdidefinitu positiboa bada. Honek $\hat{\Theta}$ bektoreko elementu bakoitzaren bariantza, dagokion $\tilde{\Theta}$ -ko elementuaren bariantza baino txikiagoa dela esan nahi du.

Batezbesteko errore kuadratikoa Alboragabetasuna desiragarria den propietatea izan arren, ez du esan nahi beti estimatzaile alboragabe bat, alboratu bat baino nahiago izango dugunik. Ondorengo grafikoaren arabera, $\hat{\theta}_1$ estimatzaile alboragabea baztertuko genuke $\hat{\theta}_2$ estimatzaile alboratuarekin alderatuz. Lehenak bariantza *oso handia* duenez, estimazioetan errore handiak egiteko probabilitate handiagoa du, bariantza txikiagoko estimatzailearekin ($\hat{\theta}_2$) konparatuz.

A.8 Irudia: Estimatzaileen banaketen adibideak



Honek, estimatzaileen aukeraketa irizpidetzat estimatzailearen errorearen neurri bat erabiltzea iradokitzen du. Jarraian, estimatzaile baten batezbesteko errore kuadratikoa definituko da:

$$BEK(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = Bar(\hat{\theta}) + [alborapena(\hat{\theta})]^2$$

Eta ondoren, estimatzaile multzo batean, batezbesteko errore kuadratikoa txikiena duena aukeratzen da.

A.2.2 Hipotesien kontrasteak

Aurretik aipatu bezala, Ekonometriaren helburu bat *hipotesiak kontrastatzea* da. Etxebizitzaren prezioen adibideko datuak 3000 euro/ m^2 batezbesteko banaketa konkretu batekin bateragarriak diren aztertzea adibidez (7. ariketa). Ikusi dugunez 6. ariketan, datuen batezbesteko aritmetikoa ez da populazioarenaren berdina izango, baina hipotesien kontraste batekin bi batezbesteko horien arteko diferentzia (populaziokoa 3000 euro eta laginekoka 3910 euro) datuen jatorri aleatorioagatik eratorria den edo ez ikusi dezakegu.

Hipotesien kontraste batek hiru etapa ditu (Ramanathan (2002)): (1) Aurkakoak diren bi hipotesi zehaztu; (2) kontrasterako estatistiko bat lortu bere lagineko banaketarekin eta (3) planteaturiko bi hipotesietatik bat aukeratzeko irizpidea ezarri.

Hipotesi bat, aldagai aleatorio bat edo batzuren banaketari buruzko baieztapen bat da. Kontraste batekin, planteatutako hipotesietatik datuei hobekien zein egokitzen den erabaki daiteke. Interesatzen zaigun hipotesia **hipotesi hutsa** (H_0) izango da eta **hipotesi alternatiboa edo aurkako hipotesiarekiko** (H_a) kontrastatzen da. Adibidez, etxebizitzaren prezioen adibidean, Y -ren batezbestekoa 3 (mila euro) delaren hipotesi hutsa planteatuko dugu 3-ren desberdina delaren hipotesiaren aurka, hau da:

$$H_0 : \mu = 3 \quad H_a : \mu \neq 3$$

Hipotesi hutsean μ -ren balio bakar bat zehazten da eta hipotesi bakuna dela esango dugu. Alternatiboa berriz, osagarritzko hipotesia izango da, gure kasuan ezberdintasuna, ondorioz, alde biko kontrastea da. Kontrastea horrelakoa bada ordea,

$$H_0 : \mu \geq 3 \quad H_a : \mu < 3$$

batezbesteko hiru edo hiru baino handiagoa dela kontrastatu nahi dugu, hiru baino txikiagoa delaren aurkako hipotesiaren aurka. Ondorioz, alde bateko kontrastea dela esango dugu.

Azkenean zein hipotesiarekin gelditzen garen **kontrasterako estatistikoaren** menpean egongo da, hau da, datuen funtzioaren arabera, konkretuki, datu eta H_0 -aren arteko diferentzien funtzioan. Batezbestekoaren bi aldetako kontrasteen adibidez, ondorengo diferentzia edo ezberdintasuna definitzen da:

$$\frac{\bar{y} - 3}{S_{\bar{y}}}.$$

Diferentzia hau kontrasteko estatistiko gisa erabiliko dugu, ez baita neurri unitateen menpekkoa eta $\bar{y}$ batezbesteko aritmetikokoa laburbiltzen den datu eta H_0 hipotesian lortutako balioen arteko diferentzia kontuan hartzen den. Gainera, hipotesi hutsa egiazkoa denean aldagai aleatorio honek izango duen banaketa ezagutu beharko dugu. Gure adibidean, $y_1, y_2, \dots, y_N$ datuak, $Y_i \sim NIB(\mu, \sigma^2)$ aldagaiaren lagina badira, non μ eta σ^2 ezezagunak diren, orduan:

$$\frac{\bar{y} - \mu}{S_{\bar{y}}} \sim t_{(N-1)}$$

eta $\mu = 3$ ordezkatzuz, H_0 hipotesi hutsaren menpean estatistikoaren lagineko banaketa izango dugu:

$$t_{est} = \frac{\bar{y} - 3}{S_{\bar{y}}} \stackrel{H_0}{\sim} t_{(N-1)} \quad (A.18)$$

Estatistiko hau praktikan asko erabiltzen da eta batezbestekoaren t estatistikoa bezala eza-gutzen da.

Azkenik, **erabaki irizpidea** zehazteko estatistikoak har ditzaken emaitza multzoa bi eremu-tan banatzen da: **eskualde kritikoa eta osagarritzkoa**. Eskualde horiek kontuan izanik, hipotesi hutsa baztertuko dugu baldin eta datuekin lortutako estatistikoaren balioa eskualde kritikoko balioa bada. Aurreko kasuan, $H_0: \mu = 3$ hipotesi hutsa dugu $H_a: \mu \neq 3$ hipotesiaren aurka eta H_0 baztertuko dugu batezbestekoaren estimazioa ($\bar{y}$) eta H_0 -ren menpean lortutako balioaren arteko distantzia *handia* bada, horrela,

$$|t_{est}| = \left| \frac{\bar{y} - 3}{S_{\bar{y}}} \right| > c \quad (\text{A.19})$$

non c balio kritikoa den eta $|\bar{y} - 3|/S_{\bar{y}} \leq c$ bada, hipotesi hutsa ez dela baztertzen ondoriozta-tuko dugu. Bestalde c -ren balioa, kontrasteiko estatistikoaren banaketaren menpean dago H_0 ematen denean eta baita, jasan nahi izango dugun errorearen menpean. Dena den, kontraste batean ondorengo erroreak egiteko posibilitatea izango da:

- Hipotesi hutsa batertzea egiazkoa denean (*I motako errorea*). Kontraste baten *tamaina edo esanguratasun maila*: I motako errorea egitearen probabilitatea da eta α izendatzen da.
- Hipotesi hutsa ez baztertu egiazkoa ez denean (*II motako errorea*). Kontraste baten *potentzia*: II motako errorea ez egitearen probabilitatea da H_a egiazkoa denean.

Orokorrean, errorerik txikiena egin nahi dugu, baina errealitatean ez da posible izaten goian aipaturiko bi motatako errore horiek ezabatzea, hau da tamaina 0 eta potentzia 1 izatea. Bestalde, I motako errorea murrizteak, II motakoa handitzea dakar. Adibidez, ez genuke I motako errorerik egingo baldin eta irizpide orokortzat hipotesi hutsa ez baztertzea bada; baina kasu horretan, kontrastearen potentzia 0 izango litzateke egiazkoa ez den H_0 ere baztertuko baitugu. Hala ere, I motako erroreari garrantzi gehiago emango diogu, kontrastearen tamaina aukeratuz. Gehien erabiltzen direnak honakoak dira:

$$0,10 \quad 0,05 \quad \text{eta} \quad 0,01$$

Horrela, tamaina aukeratu ondoren, potentzia handien duen kontrastea erabiltzen saiatuko gara.

Adibidea: banaketa Normalaren batezbestekoaren bi aldetako kontrastea. Suposa deza-gun $y_1, y_2, \dots, y_N$ datu multzoa, μ batezbestekoz eta σ^2 bariantzaduneko aldagai aleatorio Normal batetik eratortzen dela, hau da, $Y_i \sim NIB(\mu, \sigma^2)$. Demagun batezbestekoa zero den edo ez kontrastatzea interesatzen zaigula, horretarako $H_0: \mu = 0$ hipotesia planteatuko dugu $H_a: \mu \neq 0$ hipotesiaren aurka.

Kontrasterako estatistikoa osatzeko, μ -ren estimatzaile alboragabe batean oinarrituko gara, adibidez, datuen batezbesteko aritmetikoan ($\bar{y}$). Horrela, zentzuduna izango da H_0 bazter-

tzea, lagineko batezbestekoak muga bat gainditzen badu, hau da:

$$|\bar{y}| > c \quad (\text{A.20})$$

eta ez baztertzea kontrako kasuan, hau da, $-c \leq \bar{y} \leq c$. Muga hori zehazteko, estima-tzailearen banaketa erabiliko dugu:

$$\bar{y} \sim N\left(\mu, \frac{\sigma^2}{N}\right) \Rightarrow \frac{\bar{y} - \mu}{\sigma/\sqrt{N}} \sim N(0, 1)$$

Aldagai honek banaketa ezaguna du baina σ^2 bariantza ezezaguna duenez, datuak kontuan izanik, ezin izango genuke estatistikoaren egiterik edo gauzatzerik izan. Baina σ^2 bariantza bere estimatzaile alboragabeagatik ordezkari dezakegu, S^2 , ondorengoa lortuz:

$$\frac{\bar{y} - \mu}{S/\sqrt{N}} \sim t_{(N-1)}$$

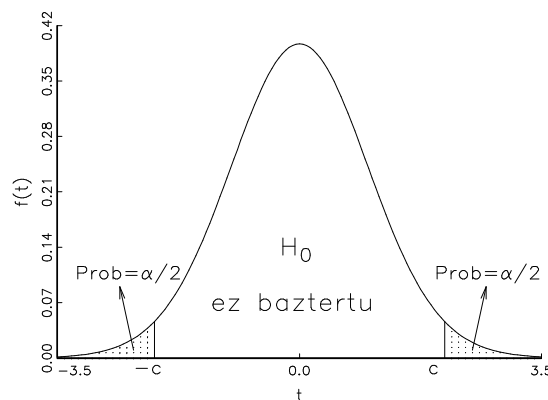
Baldin eta $H_0: \mu = 0$ egiazkoa bada, aldagai honen banaketa hau da:

$$t_{est} = \frac{\bar{y}}{S/\sqrt{N}} \stackrel{H_0}{\sim} t_{(N-1)} \quad (\text{A.21})$$

c baita kontrasterako estatistikoa ere. Estatistiko hau batezbestekoaren t estatistikoa bezala ezagutzen da. Azaldutako (A.20) irizpidea jarraituz, H_0 baztertuko dugu zerotik urruneko balioak hartzen dituenean, hau da $|t_{est}| > c$. Kontrasteko tamainaren arabera aukeratzen da c balioa, eta α esanguratasun maila izendatzen da. Azken finean, esanguratasun maila, egiazkoa den H_0 batertzearen probabilitatea da, hau da:

$$\alpha = \text{Prob}(|t_{est}| > c) \quad t_{est} \sim t_{(N-1)} \text{ denean}$$

A.9 Irudia: Alde biko kontrastearen eskualde kritikoak



c balioa, $t_{(N-1)}$ banaketaren jatorria da eta banaketako bi buztanetan α probabilitatea uzten du. Adibidez, $\alpha = 0,05$ eta $N = 50$ badira, orduan $c = 2,01$. Horrela, H_0 baztertuko dugu %5eko esangura-mailarekin, baldin eta $|t_{est}| > 2,01$ bada edota:

$$\bar{y} > 2,01 \frac{S}{\sqrt{N}} \quad \text{edota} \quad \bar{y} < -2,01 \frac{S}{\sqrt{N}}$$

A.9 grafikoko ezkerreko aldeak, estatistikoaren banaketa adierazten du $H_0: \mu = 0$ hipotesia egiazkoa denean. Banaketaren buztanetan adierazitako margotutako eremuak eskualde kritikoa osatzen dute, buztan bakoitzean $\alpha/2$ probabilitatea biltzen delarik. Kasu honetan, H_0 baztertuko dugu, datuekin lortutako t -estatistikoaren balioa, $\mu = 0$ batezbestekodun banaketan, *ia ziur* emango ez den balioa baita.

Adibidea: m^2 ko prezioaren batezbestekoaren kontrastea Gretlekin.

Demagun metro karratuko prezioa aldagaiak (*pr_m2*) banaketa Normala jaraitzen duela. Kontrasta ezazu $H_0: \mu = 3$ hipotesia $H_a: \mu \neq 3$ hipotesiaren aurka. Jarraitu beharreko pausuak honakoak dira:

1. Ondorengo estatistikoaren lagineko balioa kalkulatu $t = (\bar{y} - 3)/S_{\bar{y}}$, non $\bar{y}$, *pr_m2* aldagaiaren lagineko batezbestekoa den:

$$t = \sqrt{50}(3,9144 - 3)/0,99341 = 6,51$$

Gretleko ondorengo aukera jarraitu dezakegu:

Tresnak $\rightarrow$ Kontrasteen estatistikoaren kalkulagailua

Ondorengo lehiatilan *batezbestekoa* klikatu eta bertan:

- *Erabili datu-multzoko aldagaia* klikatu.
- Hemen *pr_m2* aldagaia aukeratuz, t kalkulatzeko behar diren estatistiko deskribatzaileak agertuko dira:

Lagineko batezbestekoa: 3,9144

Desb. tipikoa: 0,99341

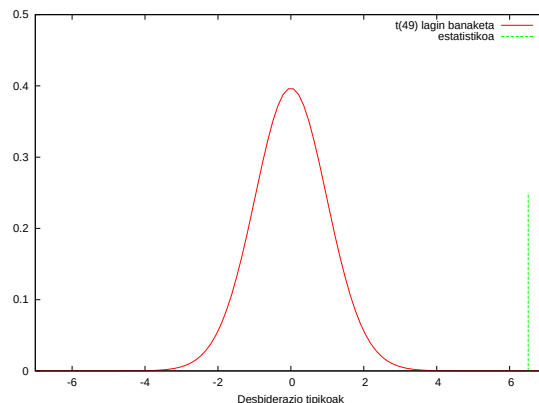
Laginaren tamaina: 50

- Kontrastatu nahi den hipotesi hutsa idatzi: H_0 : *batezbestekoa* = 3.
- Konprobatu *Suposatu desbideratze tipikoa populazioko balioa dela* aukera **ez** dagoela aktibaturik eta klikatu *Ados*.

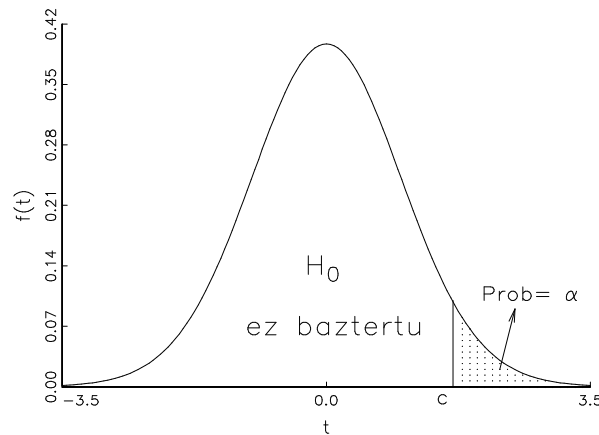
Eraitza honakoa da:

Hipotesi hutsa: *populazio batezbestekoa* = 3 *Lagin tamaina*: n =50
Lagin batezbestekoa = 3,91439, *desbideratze tip.* = 0,993407
Estatistikoa: $t(49) = (3,91439 - 3)/0,140489 = 6,50864$ *Alde-biko*
p-balioa = 3,83e-008 (*alde-batekoa* = 1,915e-008)

Dagokion grafikoa:



A.10 Irudia: Alde bateko kontrastearen eskualde kritikoak



non estatistikoaren balioa konfidantza tartetik kanpo dagoela ikusten den. Bestalde ondorengo grafikoan

estatistikoaren banaketa agertzen da hipotesi hutsaren menpean, kasu honetan $t(49)$, eta baita estatistikoaren lagineko balioa ere (lerro berdea). Ikus dezakegu estatistikoaren lagineko balioa goi buztanean erortzen dela, hau da, H_0 egiazkoa izatea ez da probablea. Beraz, hipotesi hutsa baztertuko dugu. Dena den, eskualde kritikoa kalkulatu dugu.

2. Eskualde kritikoa edo baztertze eskualdea. Gretl erabiliz c balio kritikoa aurkitu dezakegu, zehazki *Tresnak* $\rightarrow$ *Estatistika-etaulak* aukeran.

Lehiatila berrian t aukeratu behar da eta *ag* koadrotxoan askatasun graduak idatzi, gure kasuan 49, eta baita behar dugun esangura. Eraitza bezala 1,299 balioa irtetzen da.

Balioa interpretatuz, honakoa adierazten du: adibidez $X \sim t(49)$ bada orduan

$$Prob(X > 1,299) = 0,10$$

Hau da, $[1,299, \infty)$ tartean duen metatutako probabilitatea 0,1 da. Kontrastea bi aldetakoa denez $\alpha = 5\%$ esangura-mailarentzat, eskualde kritikoaren eskubiko zatian $\alpha/2$ probabilitatea metatuko da. Beraz, c balio kritikoa lortzeko 49 ag eta $\alpha/2 = 0,025$ barneratzen dugu. Ematen digun balioa $c = 2,010$ da.

3. Erabaki irizpidea aplikatu, hau da, $|6,51| > 2,010$ denez 5% esangura-mailarekin, batezbesteko prezioa 3000 euro direla hipotesi hutsa baztertzen da. Azkenik, itxi itzazu lehiatilak.

Adibidea: banaketa Normal baten batezbestekoaren alde bateko kontrastea. Askotan ikerketa ekonometrikoetan alde bateko kontrasteak burutu behar dira. Adibidez, soldatarikiko bereizketarik dagoen edo ez, aztertzeko eta konkretuki emakumena gizonena baino txikiagoa dela kontrastatzeko.

Suposa ezazu, aurreko kasuan batezbestekoa hiru edo txikiagoa dela kontrastatu nahi dela hiru baino handiagoa delaren aurka. Ondorioz, hurrengo hipotesiak planteatuko genituzke:

$$H_0: \mu \leq 3 \qquad H_a: \mu > 3.$$

Kontrastea alde batekoa denez, H_0 baztertuko dugu $H_a: \mu - 3 > 0$ hipotesiaren aurka, baldin eta $\bar{y} - 3$ diferentziak balio konkretu bat gainditzen badu, hau da $\bar{y} - 3 > c$. Hipotesi honen

kontrasterako estatistikoa ere $\bar{y}$ estimatzailearen banaketatik ondorioztatzen da. Baldin eta $H_0: \mu = 3$ egiazkoa bada, aldagai honen banaketa hurrengoa litzateke:

$$t_{est} = \frac{\bar{y} - 3}{S/\sqrt{N}} \stackrel{H_0}{\sim} t_{(N-1)} \quad (\text{A.22})$$

Beraz, H_0 baztertuko dugu $\bar{y} - 3$ diferentzia *handia* denean, hau da H_0 baztertuko dugu $t_{est} > c$ denean. Jakina da, c balioa ondorengoagatik zehazten dela: $\alpha = \text{Prob}(t_{est} > c)$; $t_{est} \sim t_{(N-1)}$ denean. Hortaz, kontrasteko eskualde kritikoa, A.9 grafikoko eskubiko aldean agertzen den paneleko margotutako eremuak osatuko du. Horrela, H_0 ez dugu baztertuko α esanguramailarekin, baldin eta t_{est} -ren balioa $t_{(N-1)}$ banaketako c balioa baino txikiagoa bada, honek α probabilitatea metatzen duelarik.

Orokorrean, H_0 *baztertu* edo *ez baztertu* hitzak erabiltzen dira. Kontraste batean H_0 mantendu ohi da, kontrakoaren ebidentzia nahiko ez dagoen bitartean. Datuek H_0 hipotesi hutsa baztertu dezakete baina ezin dute frogatu zuzena edo egiazkoa den eta ondorioz, ez da H_0 *onartu* esaten. Bestalde, H_0 ez baztertzeak, datuekin hipotesiaren faltsukeria ezin dutela frogatu esan nahi du.

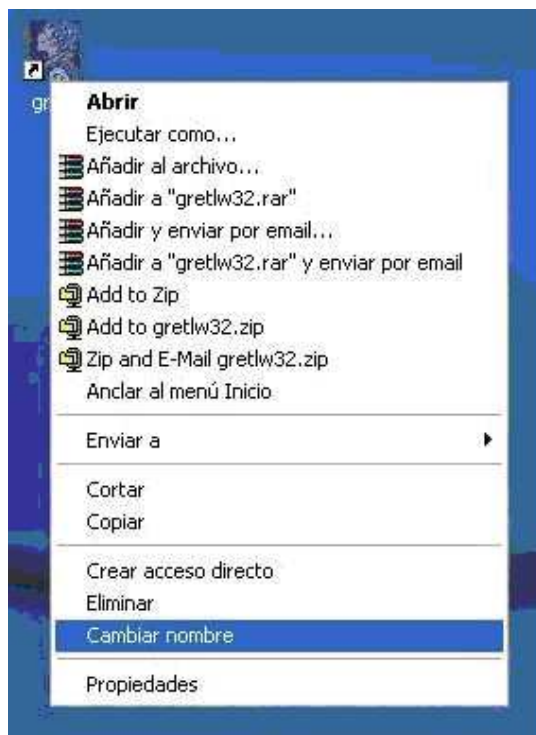
Bibliografia

- Peña, D. eta J. Romo** (1997). *Introducción a la Estadística para las Ciencias Sociales*. Mc-Graw Hill, Madrid.
- Ramanathan, R.** (2002). *Introductory Econometrics with Applications*. South western, Ohio.

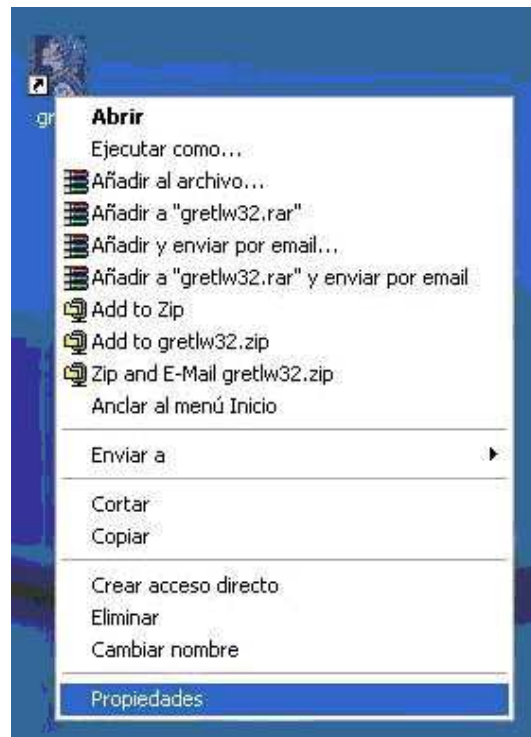
B Eranskina: Gretl euskeraz instalatzeko jarraibideak

Gretl euskeraz instalatzeko ondorengo pausuak jarraitu behar dituzue:

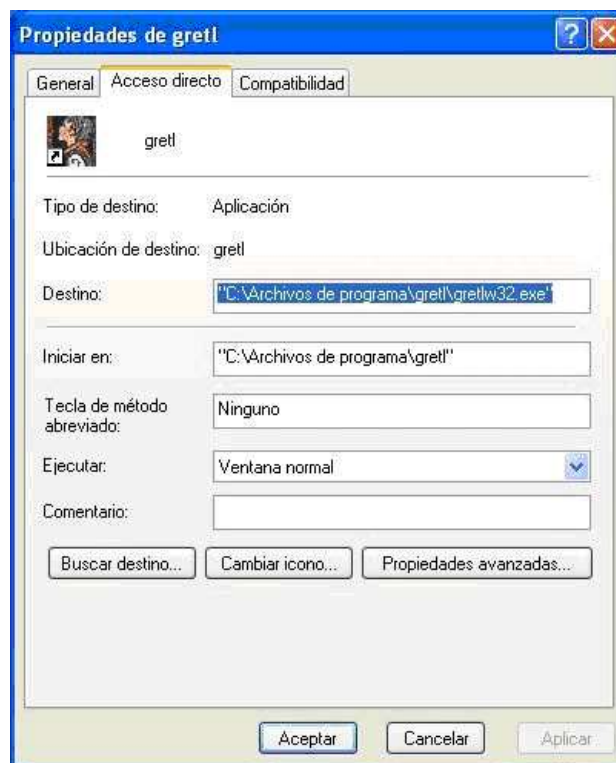
1. Gretl jaisteko web-gunea <http://gretl.sourceforge.net/win32/> da. Behin programa instalatzen duzuenean, gretl ikonoa agertuko zaizue. Ikonoaren izena aldatu nahi izanez gero (ez da beharrezkoa), ikono gainean jarritz saguaren eskumako botoiarekin hurrengo menua agertuko zaizue eta hemen Gretl-Basque jartzen duzue.



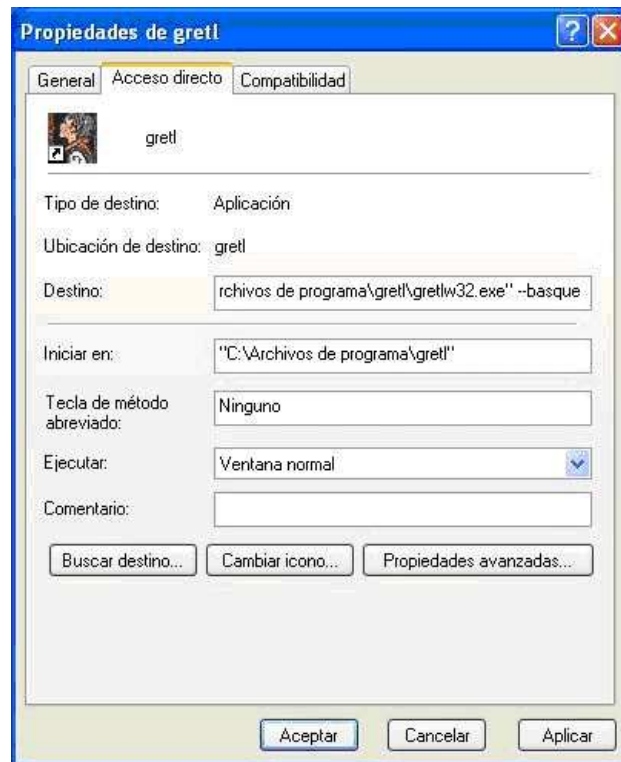
2. Euskeraren bertsioa lortzeko: berriro Gretl ikonoaren gainean jarri eta saguaren esku-mako botoiarekin irteten den menuan *Propiedades* aukeratu:



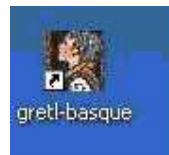
Hemen *Acceso directo* atalean *Destino* aldatu behar duzue:



Agertzen den helbideari eta azkeneko kakotxen atzetik hutsune bat utzi ondoren - - *basque* gehitu:



Hemetik aurrera euskerazko Gretlekin lan egiteko:



Ikonoari klik-bikoitza

3. Gretl lagindegian datu gehiago izateko Gretl web-gunean sartu:

[http : //gretl.sourceforge.net/gretl_data.html](http://gretl.sourceforge.net/gretl_data.html)

eta jaitsi nahi duzuen (wooldridge data.exe, stock watson.exe, ...).

Bibliografia

Belsley, D., E. Kuh eta R. Welsch (1980). *Regression Diagnostics*. Wiley, New York.

Davidson, D. eta J. Mackinnon (2004), *Econometric Theory and Methods*, Oxford University Press, New York.

Engle, R. F. (1982), “A general approach to Lagrangian Multiplier Modelo Diagnostics”, *Journal of Econometrics*, 20, 83-104 orr.

Greene, W. (2008), *Econometric Analysis*, 6, ed., Prentice-Hall, New Jersey.

Gujarati, D. (1997), *Econometría básica*, 4. ed., McGraw-Hill, México.

Heij, C., de Boer, P., Frances, P., Kloek, T. eta H. Van Dijk (2004), *Econometric Methods with Applications in Business and Economics*, Oxford University Press, Oxford.

Neter, J., W. Wasserman eta M.H. Kutner (1990). *Applied Linear Statistical Models*, 3. ed., Boston, M.A.: Irwin, Boston.

Peña, D. eta J. Romo (1997), *Introducción a la Estadística para las Ciencias Sociales*, McGraw-Hill, Madrid.

Ramanathan, R. (2002), *Introductory Econometrics with Applications*, 5. ed., South-Western, Ohio.

Stock, J. eta M. Watson (2003), *Introduction to Econometrics*, Addison-Wesley, Boston.

Verbeek, M. (2004), *A Guide to Modern Econometrics*, 2. ed., John Wiley, England.

Wooldridge, J. M. (2003), *Introductory Econometrics. A Modern Approach*, 2. ed., South-Western, Ohio.