3 Gaia

Erregresio linealeko eredu orokorra

Aurkibidea

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:

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

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

1. Aldagai azaltzaileak X_{2t},..., X_{Kt} ez dira estokastikoak.
2. Aldagai azaltzaileak linealki independenteak dira.
3. Koefiziente guztiak ($\beta_j, \quad j = 1, \cdots, K$) 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, $X_{ji} \quad j = 1, \cdots, K$ 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 etxebizitzen 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

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

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

Eredu hau matrizialki idatzi daiteke:

$$\underset{(14\times1)}{Y} = \underset{(14\times4)}{X} \underset{(4\times1)}{\beta} + \underset{(14\times1)}{u} \qquad (3.2)$$

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 Txikienen Arruntetako estimazioa

Karratu Txikienen 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 Txikienen 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 Txikienen Arruntetako estimatzailea lortzen dugu:

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

Gure adibideko eredua

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

KTA bitartez estimatzean lortzen diren emaitzak honakoak dira:

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

Aldagaia	Koefizientea	Desb. Tipikoa	t-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:

$$\widehat{\text{PRICE}}_i = \underset{(1,462)}{129,062} + \underset{(4,847)}{0,154800}\,\text{SQFT}_i - \underset{(-0,799)}{21,5875}\,\text{BEDRMS}_i - \underset{(-0,282)}{12,1928}\,\text{BATHS}_i$$

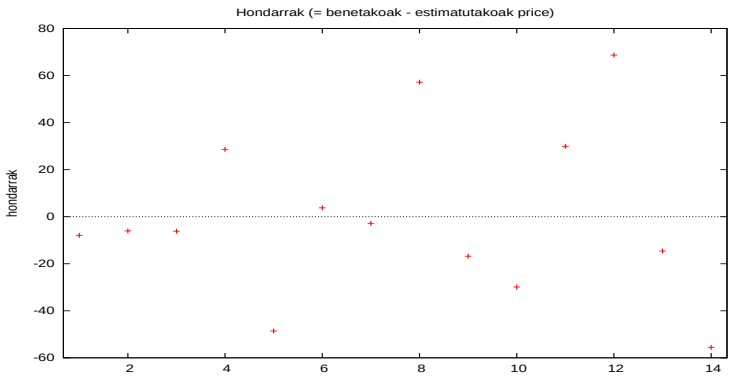
$$T = 14 \quad R^2 = 0,8359 \quad \bar{R}^2 = 0,7868 \quad F(3, 10) = 16,989$$

(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.

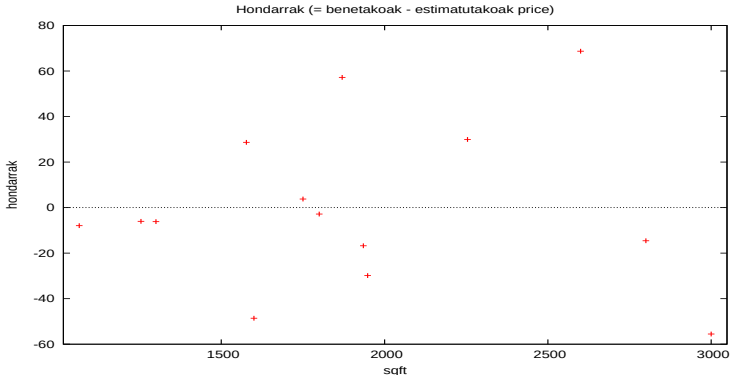
Estimazio emaitzen grafiko batzuk. Estimazio emaitzen leihatilan grafiko desberdinak ateratzeko aukera dago. Hondarren kasuan, hauek laginean zehar edo aldagai azaltzaile batekiko irudikatzeko aukera dago. Adibidez, hondarrak laginean zehar irudikatuz lortzen den grafikoa honakoa da:

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 dispertsioari dagokionez, azken behaketen dispertsioa hasierakoena baino handiagoa dela ikusten da. Hondarrak *sqft* aldagai azaltzailearekiko irudikatzerakoan lortzen den grafikoa ondorengoa da:

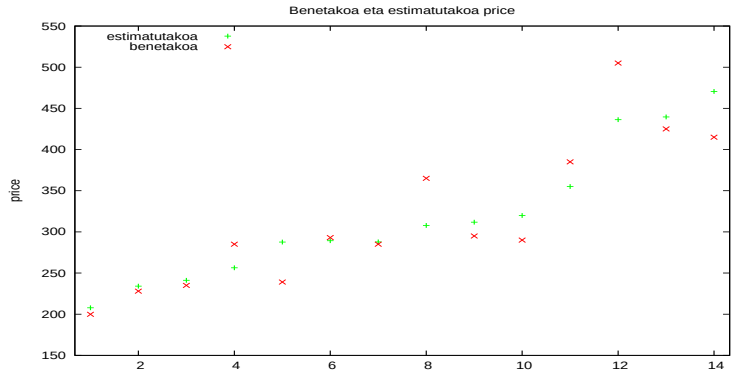
3.2 Irudia: Hondarrak sqft aldagai azaltzailearekiko



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

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

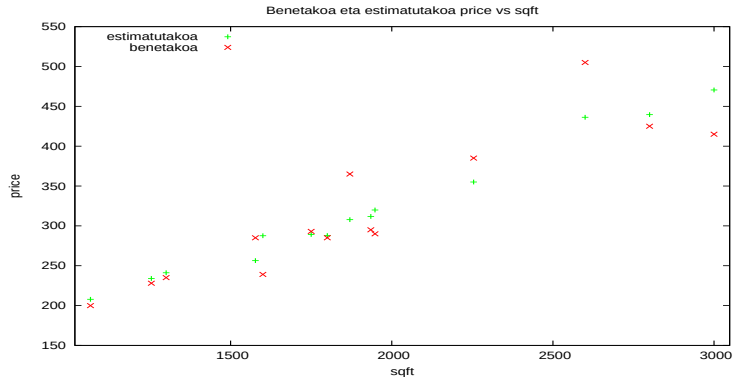
3.3 Irudia: Benetako eta estimatutako prezioak laginean zehar



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 etxebizitzen doikuntza berriz, txarragoa dela.

3.2.1 Koefizienteak

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

$$\widehat{\text{PRICE}}_i = 129,062 + 0,1548\,\text{SQFT}_i - 21,588\,\text{BEDRMS}_i - 12,193\,\text{BATHS}_i$$

Zeinuak aztertzerakoan, hasiera batean SQFT aldagaiaren koefizientearen zeinu positiboarekin ados gaude, zenbat eta etxebizitza handiagoa izan garestiagoa 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" neurtzen baitute beste aldagaiak **konstante mantenduz**. Ondorioz, etxebizitzaren azalera eta komun kopurua konstante mantenduz, gela bat gehiago izateak gela guztiak txikiagoak izan behar direla inplikatzen du. Azkenean, etxebizitzaren azalera zati gehiagotan banatu behar denez, etxebizitzaren kalitatea jaitsi egiten da, bere batezbesteko salmenta prezioa jaitsiz. 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, etxebizitzen batezbesteko salmenta prezioa 12,193 mila dolarretan gutxierazten du.

Zer gertatuko litzateke BEDRMS-ren koefizientearen zeinuarekin SQFT eta BATHS aldagaiak ereduan 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 suposatu behar. Hau da, zehaztutako eredua ondorengoa izanik, dagokion KTA estimazioaren emaitzetan ikusi daiteke zeinu aldaketa.

B Eredua $PRICE_i = \lambda_1 + \lambda_2 BEDRMS_i + u_i \qquad i = 1, \ldots, N$

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

Aldagaia	Koefizientea	Desb. Tipikoa	t-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:

C Eredua $PRICE_i = \theta_1 + \theta_2 BATHS_i + u_i \qquad i = 1, \ldots, N$

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?

D Eredua $PRICE_i = \delta_1 + \delta_2\, BEDRMS_i + \delta_3\, BATHS_i + u_i \qquad i = 1, \ldots, N$

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 estimatzerakoan? 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 baditugu, 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 $\left(E(\hat{\beta}_{KTA}) = \beta\right)$ eta bariantza minimodunekoa izatea nahi dugu. Eredu orokorreko koefizienteguztien populazio bariantzak eta kobariantzak ateratzeko ondorengo adierazpena dugu:

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

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

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

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 $\hat{Y}_i = \hat{\beta}_1 + \ldots + \hat{\beta}_K X_{Ki}$ denez, hondar hauek $\hat{u}_i = Y_i - \hat{\beta}_1 - \ldots - \hat{\beta}_K X_{Ki}$ bezala kalkulatzen 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:

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

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

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

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

Adibidez, atera ditugun estimazioetan oinarrituz eta tauletako balioa $t_{(14-4)0,05/2} = 2,228$ dela jakinik, β_2 koefizientearen %95eko konfidantza tartea ondorengoa 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 kalkulatzea da, aurreko gaian ikusi den bezala, eredua KTA bitartez estimatu ondoren lortzen den lehiatilan, *Analisia → Koefizienteen konfidantza tarteak* aukera erabiliz. Irtetzen den emaitza ondorengoa 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 $0,95$eko 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:

$$H_0\colon\ \beta_j = 0$$
$$H_a\colon\ \beta_j \neq 0$$

Eta erabiliko dugun kontrasterako estatistikoa berriz:

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

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 $\mid t_{est\ j} \mid < t_{(N-K)\alpha/2}$ 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 $\mid 4,847 \mid > 2,228$ 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 $\mid -0,799 \mid < 2,228$ eta $\mid -0,282 \mid < 2,228$ 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:

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

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 tixiagoa 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 $\alpha = 0,1$eko 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 dituenez, $H_0 : \beta_2 = 0$ 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 ($\mid t_{est\ j} \mid$) tauletako balioekin konparatuko dugu:

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

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

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

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, ereduan 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, ereduan 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:

$$H_0 \colon\ \beta_2 = \beta_3 = \beta_4 = \ldots = \beta_K = 0$$
$$H_a \colon\ \text{berdintzaren bat ez da ematen}$$

non termino konstanteari dagokion koefizientea **ez** den barneratzen. Kontrasterako estatistikoa aldiz ondorengoa da:

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

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, eredua 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

$$\begin{aligned} H_0&: \ \beta_2 = \beta_3 = \beta_4 = 0 \\ H_a&: \ \text{berdintzaren bat ez da ematen} \end{aligned}$$

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:

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

eta alboan, dagokion p-balioa agertzen da. Horrela, kontrastea burutzeko bi aukera ditugu. Lehena, $F = 16,9889 > \mathcal{F}_{(3,10)0,05} = 3,71$ denez hipotesi hutsa baztertuko 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-mailekin.

3.3 Doikuntzaren ontasuna eta ereduen sailkapena

Aurreko gaietan, ereduan 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 daitekela 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 eredua 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 berdinak. Nahiz eta horrela izan, mugatze koefizienteak arazo bat aurkezten 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:

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

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 ereduaren 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 ($\mid \hat{\beta}_j/\widehat{des}(\hat{\beta}_j) \mid > 1$), aldagai hori ereduan mantentzean, nahiz eta banakako esanguratasun kontrastean ez nabaria dela $\left(\mid \hat{\beta}_j/\widehat{des}(\hat{\beta}_j) \mid > 1 < t_{(N-K)\alpha/2}\right)$ 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, etxebizitzen 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 eredutik eredura, aldagai horren eragina aldagai azalduan desberdina delako eta beraz, koefiziente estimatuaren balioa ez da zertan berdina izan behar.

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

hasiko gara. Taulan eredu bakoitzean barneratutako aldagaien esanguratasuna jarri da, dagokion 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 gehigarririk 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 Ereduarekin** hasiko gara, bera baita aldagai azaltzaile guztiak barneratzen duen eredua. Bertako mugatze koefizientea interpretatuz, SQFT, BEDRMS eta BATHS aldagai azaltzaileen bariantzarekin PRICE aldagai azalduaren bariantzaren $\%83,5$a azaltzea lortzen da era lineal batean. Ikusi daitekeenez, guztietatik **A Eredukoa** da mugatze koefizienterik handiena. Emaitza 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, alderantzizko 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 graduen galera konpentsatzen. Horrela bada, zuzendutako mugatze koefizientean oinarrituz, eredu hauetatik etxebizitzen 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, eredua okerrago zehazten ari garela adierazten dute. Hortaz, irizpide hauek erabiliz, konklusio berdinera heldu gara: **E eredua** hoberen zehaztuta dagoen eredua da.

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

Bibliografia

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