A többváltozós lineáris regresszió 1.

2018. szeptember 17.

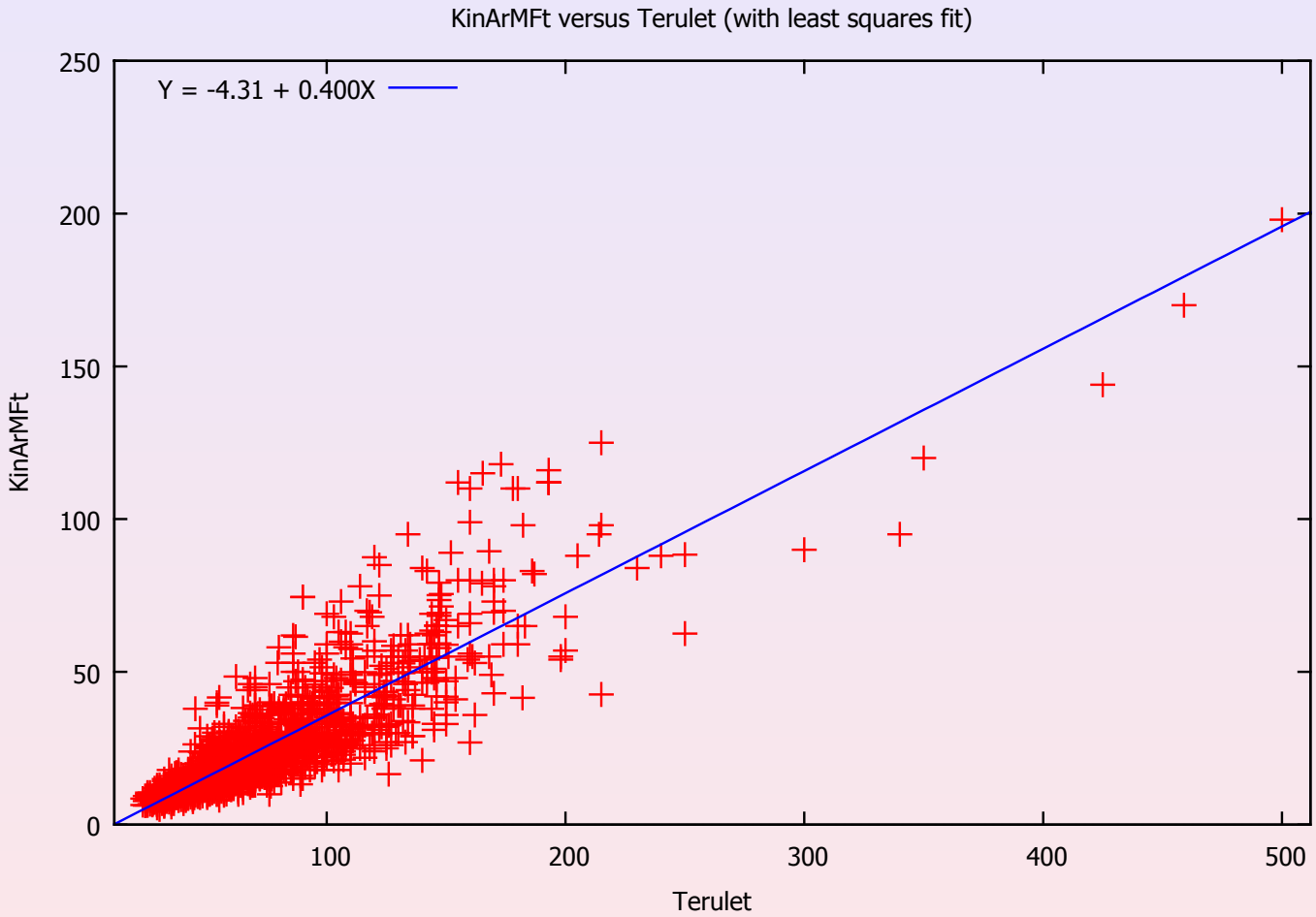
Lakásár adatbázis - részlet

KinArMFt	Terulet	Terasz	Szoba	Felszoba	Furdoszoba	Emelet	DeliTaj
10,70	32	0	1	0	1	2	0
10,00	32	0	1	0	1	2	0
10,50	32	0	1	0	1	2	0
12,00	34	0	1	0	1	1	1
13,00	34	0	1	0	1	1	1
13,90	35	0	1	0	1	0	1
17,20	43	0	2	0	1	1	0
15,90	44	3	1	0	1	2	0
13,90	46	0	2	1	1	3	1
18,00	47	0	1	0	1	2	0
14,90	47	3	1	0	1	2	0
20,00	47	0	2	0	1	1	0
15,50	48	0	2	1	1	0	0
28,90	50	15	2	1	1	1	1
19,90	50	0	3	0	1	1	1
15,90	51	0	1	0	1	2	1
16,50	52	3	2	0	1	3	1
20,00	53	0	2	0	1	0	1
17,30	55	0	2	0	1	2	0
17,90	56	0	2	0	1	1	0
16,90	56	0	2	0	1	2	0

- eredmény- és magyarázó jellegű változók
- Cél: egy eredményváltozó alakulásának jellemzése a magyarázó változók segítségével

Legegyszerűbb eset - kétváltozós kapcsolat modellezése

Magyarázó változó - terület vs. **eredmény változó - kínálati ár**



$corr(KinAr, Terulet) = 0,86$ – szoros lineáris kapcsolat. Hogyan lehetne ezt a kapcsolatot modellezni?

Modell

Legyenek tehát

- X_t: a magyarázó változó megfigyelései, $t = 1, \ldots, n$
- Y_t: az eredményváltozó megfigyelései, $t = 1, \ldots, n$
- u_t: a modell hibatagja, $t = 1, \ldots, n$.

Ekkor a modell

$$Y_t = \alpha + \beta X_t + u_t, \quad t = 1, \ldots, n$$

alakú, ahol

- X_t független u_t-től minden t esetén,
- (u_t) pedig i.i.d. (független, azonos eloszlású) sorozat 0 várható értékkel és σ szórással.

Modell

A hibatag tartalma:

- kihagyott változó(k) hatása
- helytelen függvényforma megválasztásából adódó eltérések
- mérési hibák
- modellezhetetlen véletlenség
- stb...

Feladat:

- az α és β valós paraméterek becslése a mintából
- a modell illeszkedésének vizsgálata, mérése
- előrejelzés

Paraméterbecslés – legkisebb négyzetek módszere (OLS)

Jelölje a és b az α és β paraméterek futó értékeit, és legyen

$$V(a,b) = \sum_{t=1}^{n} e_t^2 = \sum_{t=1}^{n} (Y_t - a - bX_t)^2$$

az ún. **veszteségfüggvény** (hibák négyzetösszege). Ezt kell minimalizálni az a, b paraméterek függvényében.

Megoldás: szélsőérték keresési probléma, azaz deriválunk, melyből megkapjuk a

$$\frac{\partial V(a,b)}{\partial a} = -2 \sum_{t=1}^{n} e_t = -2 \sum_{t=1}^{n} (Y_t - a - bX_t) = 0$$

$$\frac{\partial V(a,b)}{\partial b} = -2 \sum_{t=1}^{n} X_t e_t = -2 \sum_{t=1}^{n} X_t (Y_t - a - bX_t) = 0$$

ún. **normálegyenleteket**.

Paraméterbecslés – legkisebb négyzetek módszere (OLS)

Innen egyszerű számolással adódik, hogy

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} \quad \text{és} \quad \hat{\beta} = \frac{\sum_{t=1}^{n} X_t Y_t - n \cdot \bar{X}\bar{Y}}{\sum_{t=1}^{n} X_t^2 - n \cdot \bar{X}^2} = \frac{S_{xy}}{S_{xx}}$$

Példánkban (kínálati ár modellezése terület alapján) az OLS eredménye

$$\hat{\alpha} = -4,312 \quad \text{és} \quad \hat{\beta} = 0,4002.$$

Tehát ezen két együttható esetén minimális a hibák négyzetösszege, azaz ez a "legjobban" illeszkedő egyenes.

De milyen értelemben a "legjobban" illeszkedő?

Válasz: **determinációs együttható**, mely az illeszkedés jóságát méri.

Determinációs együttható

Jelölje

- **TSS**: $\sum_{t=1}^{n}(Y_t - \bar{Y})^2$ teljes négyzetösszeget, mely az átlagtól való teljes szóródás egy mérőszáma,
- **ESS**: $\sum_{t=1}^{n}(Y_t - \hat{Y}_t)^2 = \sum_{t=1}^{n} e_t^2$ a hibák négyzetösszegét, mely az eltérésváltozó szóródását méri a regressziós becslésnél,
- **RSS**: $\sum_{t=1}^{n}(\hat{Y}_t - \bar{Y})^2$ a regressziós négyzetösszeget.

Ekkor

$$TSS = \sum_{t=1}^{n}(Y_t - \bar{Y})^2 = \sum_{t=1}^{n}(Y_t - \hat{Y}_t + \hat{Y}_t - \bar{Y})^2 =$$

$$= \underbrace{\sum_{t=1}^{n}(Y_t - \hat{Y}_t)^2}_{ESS} + \underbrace{\sum_{t=1}^{n}(\hat{Y}_t - \bar{Y})^2}_{RSS} + 2\underbrace{\sum_{t=1}^{n}\underbrace{(Y_t - \hat{Y}_t)}_{e_t}(\hat{Y}_t - \bar{Y})}_{0}$$

Azaz a tökéletesen rossz modelltől a tökéletesen jó modellig vezető út (TSS) hosszából éppen RSS hosszú részt tettünk meg.

Determinációs együttható

Világos, hogy $TSS = ESS + RSS$, így definiálható az

$$R^2 = \frac{RSS}{TSS} = 1 - \frac{ESS}{TSS} \in [0, 1]$$

determinációs együttható, mely az illeszkedés jóságát méri.

- R =Corr(Y_t, $\hat{Y}$).
- mértékegység független
- ESS: meg nem magyarázott szóródás
- RSS: megmagyarázott szóródás
- értelmezhető %-ként - a magyarázó változók ismerete mennyiben csökkentette az eredményváltozó tippelésekor a hibát (ahhoz képest, mintha nem ismertünk volna egyetlen magyarázó változót sem.)

Példánkban: $R^2 = 0,7391$, azaz egy mintabeli lakásnál az ár varianciájának 73,9%-át magyarázza az alapterülete (minden más változót figyelmen kívül hagyva).

Mintavételi helyzet

Az adatbázisunk alapján tehát kaptunk egy regressziós egyenest, és azt is tudjuk, hogy ez mennyire "jól" illeszkedik az adatokra. De,

- az adatbázis csak egy **minta** az eladásra kínált lakások sokkal bővebb sokaságából, azaz a mintavétel tükröződik a paraméterek becsléseiben is;
- ezért tehát ún. **mintavételi ingadozás** lép fel;
- azaz a paraméterek ingadozását vizsgálni kell!

Statisztikai tulajdonságok

Tétel 1.

Az α és β paraméterek fenti $\hat{\alpha}$ és $\hat{\beta}$ becslései torzítatlan és konzisztens becslések.

Tétel 2.

A zaj σ^2 szórásnégyzetének torzítatlan becslése $s^2 = \frac{\sum_{t=1}^{n} e_t^2}{n-2}$.

Tétel 3 (Gauss-Markov).

A kétváltozós lineáris regressziós modell paramétereinek lineáris torzítatlan becslései közül a hagyományos legkisebb négyzetek módszerével (OLS) kapott becslések hatásosak, azaz a torzítatlan lineáris becslések közül ezek szórása a legkisebb.

Tétel 4.

A regressziós együtthatók sztenderd hibái:

$$s^2_{\hat{\beta}} = \frac{s^2}{S_{xx}}, \quad s^2_{\hat{\alpha}} = \frac{s^2 \sum_{t=1}^{n} X_t^2}{nS_{xx}}, \quad \text{és} \quad s^2_{\hat{\alpha}\hat{\beta}} = -\frac{\bar{X} \cdot s^2}{S_{xx}},$$

ahol $s^2_{\hat{\alpha}\hat{\beta}}$ a becsült paraméterértékek közti kovarianciát jelöli.

Tétel 5.

Adott X_t mellett minden u_t eloszlása $N(0, \sigma)$, akkor $\hat{\alpha}$ és $\hat{\beta}$ is normális eloszlást követnek, továbbá függetlenek s^2-től.

Konfidencia intervallumok

Könnyen igazolható, hogy

$$\frac{\hat{\alpha}-\alpha}{s_{\hat{\alpha}}} \sim t_{n-2} \quad \text{és} \quad \frac{\hat{\beta}-\beta}{s_{\hat{\beta}}} \sim t_{n-2},$$

melynek segítségével felépíthető a

$$P\left(-t^* \leq \frac{\hat{\alpha}-\alpha}{s_{\hat{\alpha}}} \leq t^*\right) = 1-\varepsilon$$

$1-\varepsilon$ szintű **konfidencia intervallum**, ahol t^* az $(n-2)$ szabadsági fokú t eloszlás $\varepsilon/2$ szinthez tartozó értéke.

$\hat{\beta}$ esetén az eljárás és a konfidencia intervallum ugyanígy írható fel.

Előrejelzés

A modell alapján adott $X = X_0$ érték mellett Y_0 becsült értéke

$$\hat{Y}_0 = \hat{\alpha} + \hat{\beta} X_0.$$

Ez az ún. **pontbecslés**.

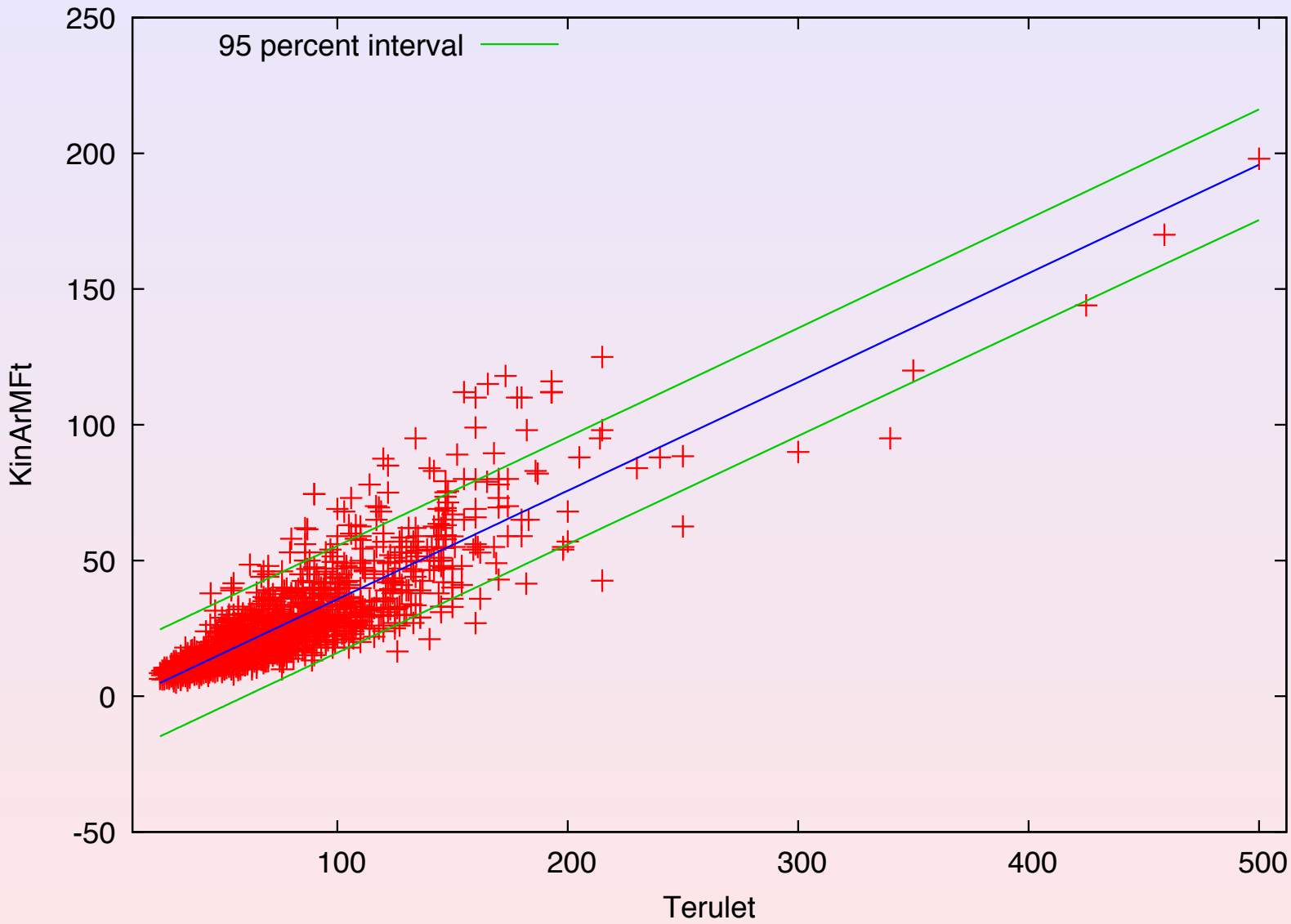
De, mivel a paraméterek becslése pontatlan, így ez is hibával terhelt, tehát a pontbecslés mellé tartozik egy konfidencia intervallum is, mely hasonlóan számolható az előzőekhez.

Alkalmazás a példára

Visszatérve a példánkra, és összefoglalva az eddigieket:

- A becsült modell: $\hat{Y} = -4,312 + 0,4002 \cdot X$
- $R^2 = 0,7391$
- A becsült paraméterek hibái: $0,5569$ és $0,0063$
- 95%-os konfidencia intervallumok: $(-5,405; -3,220)$ és $(0,388; 0,413)$
- Egy 44 m^2-es, 15,9 MFt árú lakás modell alapján becsült ára 13,3 MFt. A becslés sztenderd hibája 10,04 (!!!), tehát a konfidencia intervallum $(-6,4; 33,0)$.

Alkalmazás a példára



Többváltozós lineáris regresszió

A modell:

$$Y_t = \beta_1 + \beta_2 X_{2,t} + \ldots + \beta_k X_{k,t} + u_t, \quad t = 1, \ldots, T,$$

ahol $T \geq k$,

- $X_2, \ldots, X_k$ $(k-1)$ darab független regresszor, avagy magyarázó változó, Y a függő, avagy magyarázott változó,
- u_t azonos eloszlású, korrelálatlan sorozat, azaz $\mathrm{E}u_t = 0$, $\mathrm{E}u_t^2 = \sigma^2$ és $\mathrm{E}u_t u_s = 0$, ha $t \neq s$.

Legyen $y = (Y_1, \ldots, Y_T)^T$, $\beta = (\beta_1, \ldots, \beta_k)^T$, $u = (u_1, \ldots, u_T)^T$,

$$X = \begin{pmatrix} 1 & X_{2,1} & \ldots & X_{k,1} \\ 1 & X_{2,2} & \ldots & X_{k,2} \\ \vdots & \vdots & & \vdots \\ 1 & X_{2,T} & \ldots & X_{k,T} \end{pmatrix}.$$

Ekkor a modell az $y = X\beta + u$ kompakt formában is felírható.

Paraméterbecslés

Legyen $b \in \mathbb{R}^k$ egy tetszőleges futó paramétervektor. Ekkor a hibavektor

$$e = e(b) = y - Xb,$$

tehát a költségfüggvény, azaz a hibák négyzetösszege a

$$V(b) = e^T e = (y - Xb)^T (y - Xb) = y^T y - 2b^T X^T y + b^T X^T Xb$$

alakot ölti. Ennek a minimumát kellene megtalálni. Mivel

$$\frac{\partial V(b)}{\partial b} = -2X^T y + 2X^T Xb$$

$$\frac{\partial^2 V(b)}{\partial b^2} = 2X^T X \geq 0,$$

így a minimum

$$\hat{\beta} = (X^T X)^{-1} X^T y,$$

feltéve, hogy X teljes, azaz k rangú, mert csak ekkor létezik e fenti inverz és lesz ezáltal a minimum egyértelmű.

Paraméterbecslés - geometriai jelentés

Mit is jelent geometriailag ez a becslés?

- Az X oszlopvektorai által kifeszített altér épp azon pontokból áll, melyek előállnak eredményváltozóként valamilyen regressziós együtthatókkal.
- Általában persze a valódi eredmény változó ebben nincsen benne, ezt fejezi ki a reziduum. Azaz hogy mennyire messze van egymástól a valódi és becsült eredményváltozó.
- Az altér legközelebbi pontja kell nekünk... azaz merőlegesen vetítünk!
- A valódi eredmény változót merőlegesen levetítve az altérre kapjuk meg a legjobb becslést, a kapott együtthatók pedig az optimális regressziós koefficiensek.
- Azaz az

$$\hat{y} = X\hat{\beta} = \underbrace{X(X^T X)^{-1} X^T}_{P} y = Py$$

összefüggésben a P mátrix éppen a megfelelő projekció mátrix.

Hibatag tartalma

- Kihagyott lényeges változó(k) hatása - övezet, felújítás, fűtés típusa
- Bevont fölösleges változó(k) hatása - félszobák száma
- Helytelen függvényforma megválasztása - emelet hatása nemlineáris
- Mérési hibák
- Megmagyarázhatatlan (előre nem látható) véletlen hatások

Alkalmazás a példára

OLS, using observations 1–1406, Dependent variable: KinArMFt

	Coefficient	Std. Error	t-ratio	p-value
const	−9,01894	0,778806	−11,58	0,0000
Terulet	0,297493	0,0102045	29,15	0,0000
Terasz	0,310490	0,0206688	15,02	0,0000
Szoba	**0,689808**	0,348937	1,977	0,0483
Felszoba	−0,385095	0,405612	−0,9494	0,3426
Furdoszoba	**5,13010**	0,681629	7,526	0,0000
Emelet	0,0544509	0,132172	0,4120	0,6804
DeliTaj	1,19155	0,457579	2,604	0,0093
Buda	6,21968	0,475105	13,09	0,0000

Mean dependent var	26,49523	S.D. dependent var	19,63583
Sum squared resid	98751,24	S.E. of regression	8,407620
R^2	0,817708	Adjusted R^2	0,816664
$F(8, 1397)$	783,3158	P-value(F)	0,000000
Log-likelihood	−4984,082	Akaike criterion	9986,164
Schwarz criterion	10033,40	Hannan–Quinn	10003,82

Paraméterek értelmezése

A **paraméterek értelmezésével** elemezhetjük modellünket:

- egyszerű értelmezés - ha a vizsgált magyarázó változó egy egységnyivel megváltozna (c.p.), akkor modellünk szerint várhatóan hány egységnyit változna az eredményváltozó
- konstans értelmezése - ha valamennyi magyarázó változó nulla értékű lenne, akkor modellünk szerint várhatóan mennyi az eredményváltozónk értéke

Figyelem,

- ceteris paribus, azaz "a többi változatlanul hagyásával"
- minden változót a saját egységében mérve

A példára alkalmazva tehát

- plusz egy szoba kb. $0,68$ MFt többletet jelent, míg
- plusz egy fürdőszoba kb. 5 MFt többletet jelent! (Kicsit furcsa, de normális jelenség, később még vizsgálni fogjuk.)

Többszörös determinációs együttható

Mivel

$$X^T X\hat{\beta} = X^T y = X^T(X\hat{\beta} + e(\hat{\beta})),$$

így $X^T e(\hat{\beta}) = 0$, azaz a hibavektor mindegyik regressziós vektorral korrelálatlan! Ezt felhasználva adódik, hogy a teljes négyzetösszeg (TSS) felírható az alábbi felbontásban:

$$\sum_{i=1}^{T}(Y_i - \bar{Y})^2 = y^T y - T\bar{Y}^2 = (\hat{y} + e)^T(\hat{y} + e) - T\bar{Y}^2 =$$

$$= \hat{y}^T\hat{y} + 2\underbrace{e^T\hat{y}}_{0} + e^T e - T\bar{Y}^2 = \underbrace{(\hat{y}^T\hat{y} - T\bar{Y}^2)}_{RSS} + \underbrace{e^T e}_{ESS}$$

Így, hasonlóan a korábbi esethez, definiálható az

$$R^2 = \frac{RSS}{TSS} = 1 - \frac{ESS}{TSS} \in [0, 1]$$

többszörös determinációs együttható. Ha $\beta_1 \neq 0$, akkor $R = \text{Corr}(Y_i, \hat{Y})$.

Modellfeltevések

Ahhoz, hogy a becslésünk előnyös tulajdonságokkal rendelkezzen, modellünknek bizonyos **feltételek**et teljesítenie kell:

- lineáris modellstruktúra
- adatmátrix teljes oszloprangú, azaz nincs lineáris kapcsolat az oszlopai közt (kollinearitás)
- a hibák függetlenek a magyarázó változóktól (exogenitás)
- a hibák azonos eloszlásból származnak, azaz szórásuk megegyezik (homoszkedaszticitás)
- a hibák korrelálatlan sorozatot alkotnak (autokorrelálatlanság)

Ezek teljesülését mostantól mindig feltesszük. Feloldásukról és azok hatásairól később fogunk beszélni.

Tétel 6.

A β paraméter fenti $\hat{\beta}$ becslése torzítatlan becslés, továbbá $D^2(\hat{\beta}) = (X^T X)^{-1}\sigma^2$.

Tétel 7.

A zaj σ^2 szórásnégyzetének torzítatlan becslése $s^2 = \frac{e(\hat{\beta})^T e(\hat{\beta})}{n-k}$.

Tétel 8 (Gauss-Markov).

Legyen $c \in \mathbb{R}^k$ tetszőleges és $\mu = c^T\beta$. Ennek legkisebb szórású, lineáris, torzítatlan (BLUE) becslése az y minta alapján $\hat{\mu} = c^T\hat{\beta}$, ahol $\hat{\beta}$ a lineáris regressziós együtthatók fenti LS-becslése.

Változók relevanciája

Definíció 1.

Egy változót relevánsnak nevezünk, ha a sokasági paramétere (regressziós együtthatója) nem nulla, azaz $\beta_i \neq 0$.

A becsült regressziós együtthatók mintavételi ingadozását a

$$\frac{\hat{\beta}_i - \beta_i}{\hat{se}(\hat{\beta}_i)} \sim t_{n-k}$$

összefüggés írja le. Ennek segítségével megkonstruálható a változó relevanciájára vonatkozó (parciális) t-próba, ahol $H_0 : \beta_i = 0$.

	Coefficient	Std. Error	t-ratio	p-value
const	−9,01894	0,778806	−11,58	0,0000
Terulet	0,297493	0,0102045	29,15	0,0000
Terasz	0,310490	0,0206688	15,02	0,0000
Szoba	0,689808	0,348937	1,977	**0,0483**
Felszoba	−0,385095	0,405612	−0,9494	**0,3426**
Furdoszoba	5,13010	0,681629	7,526	**0,0000**
Emelet	0,0544509	0,132172	0,4120	**0,6804**
DeliTaj	1,19155	0,457579	2,604	0,0093
Buda	6,21968	0,475105	13,09	0,0000

Konfidencia intervallum

Ez alapján könnyen szerkeszthető **konfidencia intervallum** adott $(1-\alpha)$ megbízhatósági szintre:

$$\hat{\beta}_i \pm t_{1-\alpha/2} \cdot \hat{se}(\hat{\beta}_i)$$

A példánkban $t(1397, 0, 025) = 1, 962$, így

Variable	Coefficient	95% confidence interval	
const	−9,01894	−10,5467	−7,49118
Terulet	0,297493	0,277475	0,317511
Terasz	0,310490	0,269945	0,351035
Szoba	**0,689808**	**0,00531082**	**1,37431**
Felszoba	**−0,385095**	**−1,18077**	**0,410579**
Furdoszoba	**5,13010**	**3,79297**	**6,46723**
Emelet	**0,0544509**	**−0,204826**	**0,313728**
DeliTaj	1,19155	0,293935	2,08917
Buda	6,21968	5,28768	7,15167

A modell egészének relevanciája

- Az előző próba **csak egyenkénti vizsgálatra alkalmas**, ezért parciális.
- **A modell egészének relevanciája** is érdekes számunkra, azaz hogy a

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

hipotézis fennáll-e.

- Figyelem, most

$$H_1 : \exists i : \beta_i \neq 0$$

- Azaz azt vizsgáljuk, hogy a modell eltér-e lényegesen a nullmodelltől, vagy sem.
- **Implikálja az egyes változók irrelevanciáját külön-külön**, tehát először ezt vizsgáljuk.
- A próba

$$\frac{RSS/k}{ESS/(n-k-1)} \sim F_{k,n-k-1}$$

globális F-próba, azaz egy ANOVA-próba.

Alkalmazás a példára

Analysis of Variance

	Sum of squares	df	Mean square
Regression	442969	8	55371,1
Residual	98751,2	1397	70,6881
Total	541720	1405	385,566

$$R^2 = \frac{442969}{541720} = 0,81708$$

$$F(8, 1397) = \frac{55371,1}{70,6881} = 783,316 \text{ [p-value 0]}$$

Parciális korrelációs együttható

Cél: két kvantitatív változó kapcsolatából ki akarjuk szűrni egy vagy több kvantitatív változó hatását.

Kérdés: milyen lenne a vizsgált két változó kapcsolata, ha a kiszűrt változókat állandó szinten tartanánk?

Válasz: parciális korrelációs együttható

$$\rho_{XY,Z} = \frac{\rho_{XY} - \rho_{XZ}\rho_{YZ}}{\sqrt{1-\rho_{XZ}^2}\sqrt{1-\rho_{YZ}^2}}$$

Feltételek:

- kvantitatív (skálás) változóink vannak;
- csak lineáris összefüggés létezik köztük;
- X és Y között ugyanolyan jellegű és szintű kapcsolat van a Z változó teljes értéktartományában.

Standardizált együtthatók

- Az eddig látott $\hat{\beta}_i$ együtthatók mértékegység függőek.
- A mértékegység megváltoztatásával ezek is változhatnak. Ettől a jelenségtől szeretnénk megszabadulni.
- Azaz **standardizálunk**, ami azt jelenti, hogy 0 várható értékűvé, és 1 szórásúvá transzformáljuk az együtthatókat, mint v.v.-kat.
- **Első lehetőség**: az adatbázis standardizálása, majd a becslések lefuttatása, melynek eredményeképp megkapjuk a standardizált regressziós együtthatókat.
- **Második lehetőség**: ehelyett érvényes a

$$\tilde{\hat{\beta}}_i = \hat{\beta}_i \cdot \sqrt{\frac{Var(X_i)}{Var(Y)}}$$

összefüggés, ahol $Var(.)$ a változó szórásnégyzetét jelöli. Azaz nem kell standardizálni a teljes adatbázist a standardizált együtthatók előállításához.

Parciális korreláció és standardizált együtthatók a példában

Változó	coeff.	std.coeff.	corr	parc.corr.
konstans	-9,01894			
terület	0,297493	0,637543	**0,8597**	**0,565**
terasz	0,310490	0,196722	0,5478	0,363
szoba	0,689808	0,070112	**0,7115**	**0,085**
félszoba	-0,385095	0,008012	-0,0102	0,017
fürdőszoba	5,13010	0,109830	**0,6173**	**0,174**
emelet	0,0544509	-0,031682	-0,0800	-0,071
déli tájolás	1,19155	0,030002	0,1884	0,067

Előrejelzés