Lineáris regressziószámítás 1. - kétváltozós eset

Orlovits Zsanett

2020. február 10.

Adatbázis - részlet

id	nem	kor	iskola	munka kat.	fizetes	kezdo fizetes	alk. Ideje	korabbi tapasztalat
1	m	39	15	3	57000	27000	98	144
2	m	33	16	1	40200	18750	98	36
3	f	62	12	1	21450	12000	98	381
4	f	44	8	1	21900	13200	98	190
5	m	36	15	1	45000	21000	98	138
6	m	33	15	1	32100	13500	98	67
7	m	35	15	1	36000	18750	98	114
8	f	25	12	1	21900	9750	98	0
9	f	45	15	1	27900	12750	98	115
10	f	45	12	1	24000	13500	98	244
11	f	41	16	1	30300	16500	98	143
12	m	25	8	1	28350	12000	98	26

- eredmény- és magyarázó jellegű változók
- **Cél: egy eredményváltozó alakulásának jellemzése a magyarázó változók segítségével**, avagy a változók közötti kapcsolat jellegének feltárása, matematikai függvényekkel történő leírása.

Forrás: SPSS adatfájlok

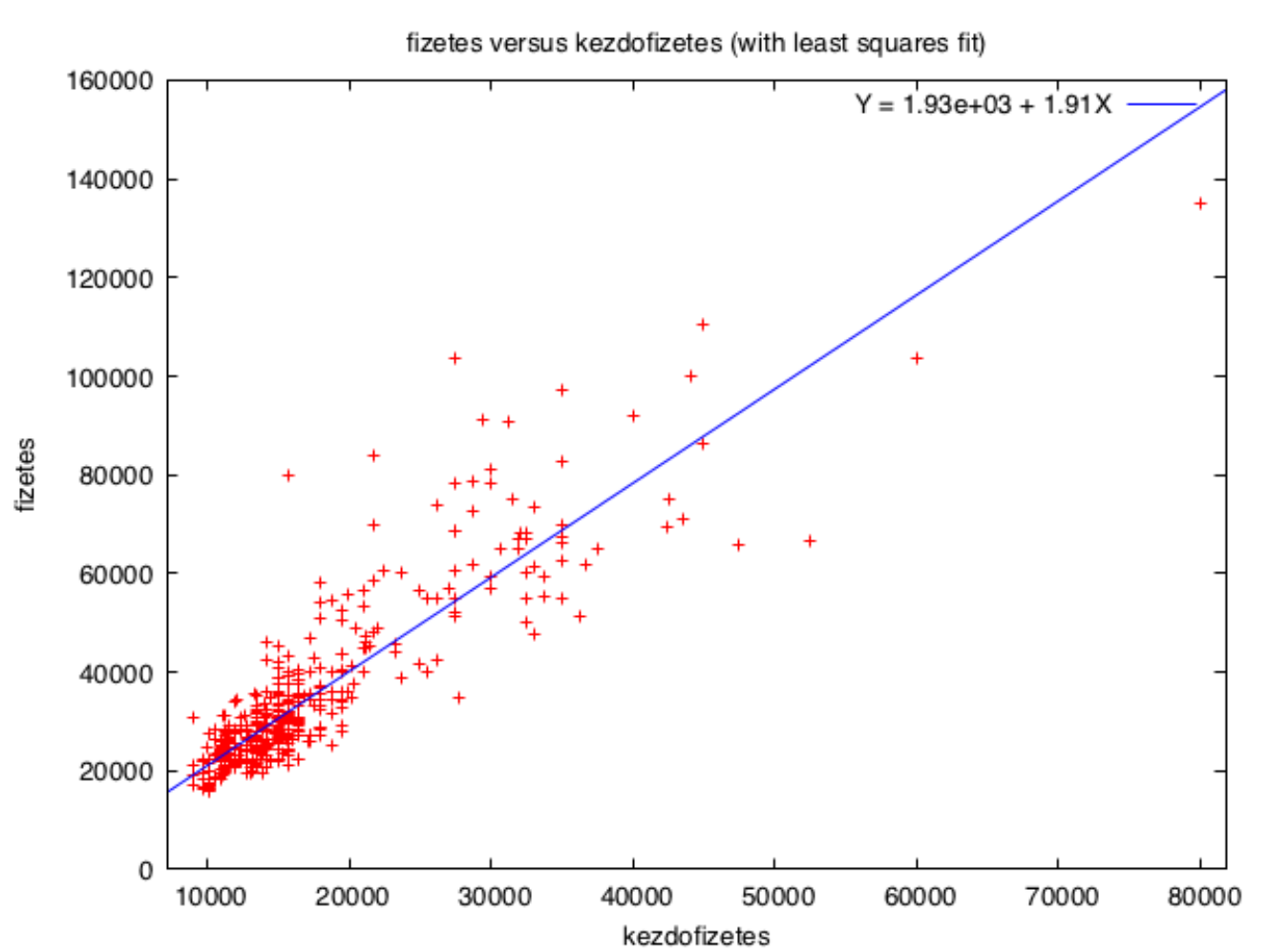
Legegyszerűbb eset - kétváltozós kapcsolat modellezése

Tekintsünk csak két változót az adathalmazunkból:
az **aktuális fizetés** alakulása a **kezdő fizetés** függvényében.

id	nem	kor	iskola	munka kat.	fizetes	kezdo fizetes	alk. Ideje	korabbi tapasztalat
1	m	39	15	3	57000	27000	98	144
2	m	33	16	1	40200	18750	98	36
3	f	62	12	1	21450	12000	98	381
4	f	44	8	1	21900	13200	98	190
5	m	36	15	1	45000	21000	98	138
6	m	33	15	1	32100	13500	98	67
7	m	35	15	1	36000	18750	98	114
8	f	25	12	1	21900	9750	98	0
9	f	45	15	1	27900	12750	98	115
10	f	45	12	1	24000	13500	98	244
11	f	41	16	1	30300	16500	98	143
12	m	25	8	1	28350	12000	98	26

Legegyszerűbb eset - kétváltozós kapcsolat modellezése

Magyarázó v. - kezdő fizetés vs. **eredmény v. - akt. fizetés**



$corr(fizetes, kezdofizetes) = 0.88$ – szoros lineáris kapcsolat.

Lineáris regresszió általában

A **korrelációszámítás** intervallum-, vagy arányskálán mért változók kapcsolatainak vizsgálatával foglalkozik, elemzi a kapcsolat meglétét, szorosságát és irányát.

A **regressziószámítás** az összefüggésekben lévő tendenciát vizsgálja, és a kapcsolat természetét valamilyen függvénnyel írja le.

- A adatokra legjobban illeszkedő egyenest keressük analitikus formában.
- Ez lesz az ún. **regressziós egyenes**, a modell pedig a lineáris regresszió kétváltozós esete.
- Torzított lesz a modell, de a lényeg kiemelésére, elemzésre és előrejelzésre kiválóan alkalmas.

Modell

Legyenek tehát

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

Ekkor a modell

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

alakú, ahol

- X_t független ε_t-től minden t esetén,
- (ε_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)

Tekintsük tehát az

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

elméleti modellt, és jelölje

$$\hat{Y}_t = \alpha + \beta X_t, \quad t = 1, \ldots, n$$

a minta alapján becsült regressziófüggvényt.

OLS módszer

Ekkor a legkisebb négyzetek (OLS) módszer lényege az, hogy azt az $\hat{\alpha}$ és $\hat{\beta}$ párt keressük, melyre a hibák négyzetösszege minimális, azaz

$$V(\alpha, \beta) = \sum_{t=1}^{n} (Y_t - \hat{Y}_t)^2 = \sum_{t=1}^{n} (\underbrace{Y_t - \alpha - \beta X_t}_{e_t})^2 = \sum_{t=1}^{n} e_t^2 \to \min!_{\alpha,\beta}$$

Paraméterbecslés – legkisebb négyzetek módszere (OLS)

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

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

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

ún. **normálegyenleteket**.

Becslések

Innen egyszerű számolással adódik, hogy

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} \quad \text{és}$$

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

Feladat

Egy 10 elemű minta alapján vizsgálták az Opel Corsa 1.2 típusú személygépkocsik életkora és eladási ára közötti kapcsolatot. A megfigyelések és a belőlük számolt regressziós részeredmények:

életkor (év)	eladási ár (ezer Ft)	számítási részeredmények
3	1720	$\sum X_i^2 = 157$
1	1800	$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = -4231$
6	1350	$\sum (\hat{Y}_i - \bar{Y})^2 = 372\,169.7$
4	1600	$\sum (Y_i - \bar{Y})^2 = 404\,610$
4	1500	$\bar{X} = 3,3,\ \bar{Y} = 1627$
5	1550	
0	2000	
1	1750	
7	1300	
2	1700	

Írja fel az életkor és az eladási ár kapcsolatát leíró becsült kétváltozós lineáris regressziós modellt! $\Rightarrow$ **GYAKORLAT**

Feladat

	életkor: X	eladási ár: Y	X-Xátlag	Y-Yátlag	S_XY	(X-Xátlag)^2	(Y-Yátlag)^2
	3	1720	-0.3	93	-27.9	0.09	8649
	1	1800	-2.3	173	-397.9	5.29	29929
	6	1350	2.7	-277	-747.9	7.29	76729
	4	1600	0.7	-27	-18.9	0.49	729
	4	1500	0.7	-127	-88.9	0.49	16129
	5	1550	1.7	-77	-130.9	2.89	5929
	0	2000	-3.3	373	-1230.9	10.89	139129
	1	1750	-2.3	123	-282.9	5.29	15129
	7	1300	3.7	-327	-1209.9	13.69	106929
	2	1700	-1.3	73	-94.9	1.69	5329
sum	33	16270	1.78E-15	0	-4231	48.1	404610
átlag	3.3	1627					

alfa= 1917.277 beta= -87.963

A becsült modell tehát

$$\hat{Y}_t = 1917.277 - 87.963 \cdot X_t$$

alakú. De hogyan is kell értelmezni ezeket a számokat?

Együtthatók értelmezése

Együtthatók értelmezése a példában

- A konstans $\hat{\alpha} = 1917$ azt mutatja meg, hogy $X_t = 0$ esetén, azaz egy új autó esetén, mennyi az átlagos eladási ár.
- A $\hat{\beta} = -87.96$ együttható azt mutatja meg, hogy a kor növekedésével évente 87 960 forinttal csökken az autók átlagos eladási ára.

Együtthatók értelmezése **általánosan**

- A konstans $\hat{\alpha}$ azt mutatja meg, hogy $X_t = 0$ esetén a modell szerint mekkora lesz az eredményváltozó átlagos értéke.
- A $\hat{\beta}$ együttható azt mutatja meg, hogy a magyarázó változó egységnyi növekedése a becsült eredményváltozóban átlagosan hány egységnyi növekedést/csökkenést okoz.

Vegyük észre, hogy $\beta = \frac{\partial Y}{\partial X}$, amit korábban marginális hatás néven tanultak! Részletek később.

Rugalmasság

- A regressziós együtthatók természetes mértékegységben jellemzik a két változó kapcsolatát.
- Előfordul azonban, hogy a kapcsolatot jobban leírhatjuk a rugalmasság fogalmának segítségével.

Rugalmasság

A **rugalmasság** azt mutatja meg, hogy a magyarázó változó 1%-os növekedése az eredményváltozó hány %-os változásával jár együtt.

- Matematikailag ez a relatív változások egymáshoz való viszonyával írható le, azaz

$$El(Y|X) = \frac{\frac{\partial Y}{Y}}{\frac{\partial X}{X}} = \frac{\partial Y}{\partial X} \cdot \frac{X}{Y}$$

Rugalmasság

Kétváltozós lineáris regresszió esetén az

$$El(\hat{Y}|X) = \frac{\hat{\beta}X}{\hat{\alpha} + \hat{\beta}X}$$

ami **mindig az éppen aktuális X függvénye!**

Módosítás

A rugalmasság azt mutatja meg, hogy **a magyarázó változó adott szintről kiinduló** 1%-os növelése a becsült eredményváltozó hány %-os változásával jár együtt.

Példa

Adott $X = 3.3$ esetén (ez pont az átlag)

$$El(\hat{Y}|\bar{X}) = \frac{-87.96 \cdot 3.3}{1917 - 87.96 \cdot 3.3} = -0.178.$$

Ha az átlagos kor 1%-kal nő, akkor a becsült átlagos eladási ár 0.178%-kal csökken.

Reziduumok

A becsült regressziós függvény segítségével a megfigyelési pontokban meghatározhatjuk a **reziduumok** értékét:

Reziduumok

$$e_t = Y_t - \hat{Y}_t, \quad t = 1, \ldots, n.$$

- Fontos szerepet játszanak a modellezésben.
- Megmutatják, hogy a modell mennyire tudott közel kerülni a valósághoz.
- A kis e_t értékek jó illeszkedést mutatnak, ez fontos kritérium lesz a modell megítélésekor.
- Mutatót képzünk belőlük: error of sum of squares (ESS), avagy **a reziduumok négyzetösszege**

ESS - reziduumok négyzetösszege

$$ESS = \sum_{t=1}^{n} e_t^2.$$

Ennek segítségével definiálható a **reziduális szórás**.

A reziduális szórás, avagy a mintán belüli reziduális variancia

$$s_e = \sqrt{\frac{ESS}{n-2}},$$

mely azt fejezi ki, hogy a regressziós becslések átlagosan mennyivel térnek el az eredményváltozó megfigyelt értékeitől.

Ezért a **regressziós becslés abszolút hibájának** is szokták nevezni.

Determinációs együttható

Azt már láttuk, hogy az OLS módszerrel becsült paraméterek mellett kapjuk meg az adatbázisra legjobban illeszkedő egyenes. De milyen értelemben a "legjobban" illeszkedő?

Válasz: **determinációs együttható**, mely az illeszkedés jóságát méri. Származtatásához a varianciafelbontásra (ANOVA) lesz szükségünk. Jelölje tehát

- **TSS**: $\sum_{t=1}^{n}(Y_t - \bar{Y})^2$ a 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, azaz ez az ANOVA belső négyzetösszege,
- **RSS**: $\sum_{t=1}^{n}(\hat{Y}_t - \bar{Y})^2$ a regressziós négyzetösszeget, mely az ANOVA külső négyzetösszege.

Determinációs együttható

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.

A külső négyzetösszegeket (RSS) magyarázott négyzetösszegnek is szokták nevezni.

Determinációs együttható

Determinációs együttható

A $TSS = ESS + RSS$ összefüggés alapján definiálható az

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

determinációs együttható, mely megmutatja, hogy a regressziós modellel az eredményváltozóban meglévő variancia (bizonytalanság) hány százalékban magyarázható meg.

Más szóval az illeszkedés jóságát méri, avagy a modell magyarázó erejének is szokás nevezni. Mértékegység független, továbbá $R = \text{Corr}(Y_t, \hat{Y})$.

Példa

Példánkban $R^2 = 0.919$, azaz egy mintabeli autónál az eladási ár varianciájának 91.9%-át magyarázza az autó kora.

Determinációs együttható - példa

	életkor: X	eladási ár: Y	(Y-Yátlag)^2	becsült Y	e_t^2
	3	1720	8649	1653.38877	4437.05551
	1	1800	29929.000	1829.314	859.306452
	6	1350	76729.000	1389.501	1560.33212
	4	1600	729.000	1565.426	1195.34796
	4	1500	16129.000	1565.426	4280.58705
	5	1550	5929.000	1477.464	5261.52679
	0	2000	139129.000	1917.277	6843.17625
	1	1750	15129.000	1829.314	6290.69938
	7	1300	106929.000	1301.538	2.36686391
	2	1700	5329.000	1741.351	1709.93426
sum	33	16270	404610.000	16270.000	32440.3326
átlag	3.3	1627	TSS		ESS

alfa= 1917.27651 beta= -87.963

rugalmasság= -0.1784121 R^2= 0.91982321
átlagos szinten

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!

Tétel 1.

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

Tétel 2.

A reziduumok szórásnégyzetének torzítatlan becslése $s_e = \sqrt{\frac{ESS}{n-2}}$.

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.

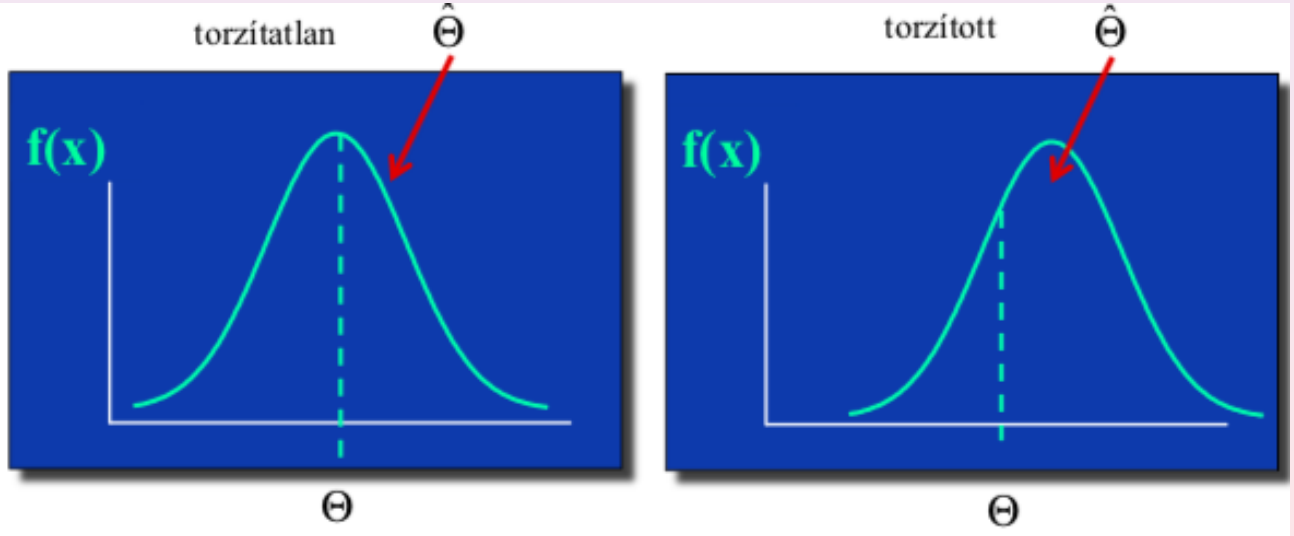
Statisztikai tulajdonságok - Torzítatlanság

- A becslés várható értéke éppen a megfelelő paraméterértékkel egyenlő, azaz

$$\mathrm{E}\hat{\theta} = \theta.$$

- Azaz bár a minta függ a véletlentől, így ezáltal a becslések is, de az eltérések középpontja az elméleti paraméter legyen.
- Amennyiben a becslés torzított, úgy a torzítás mértéke éppen

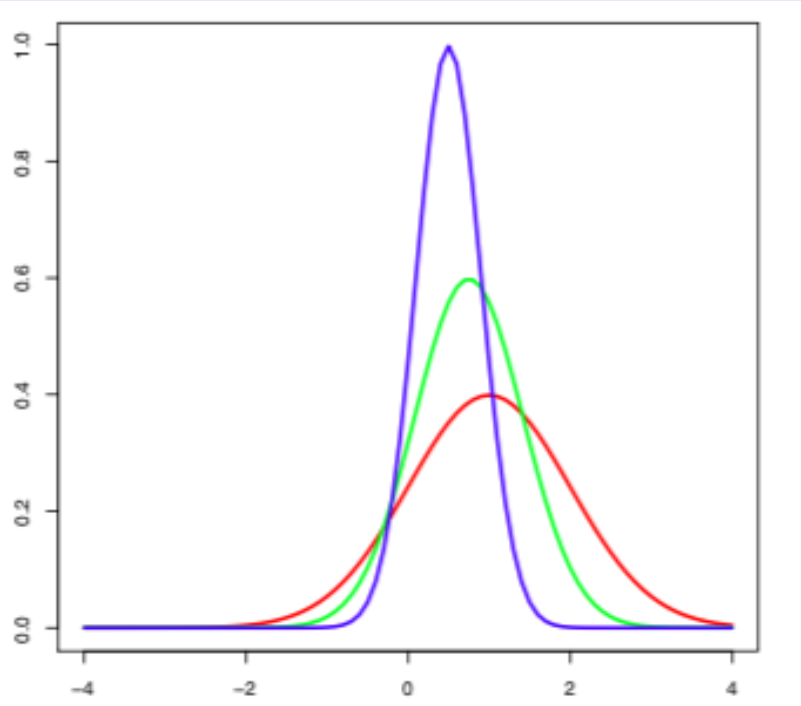
$$\mathrm{E}\hat{\theta} - \theta.$$



Forrás: Gazdaságstatisztika jegyzet, BME 2018

Statisztikai tulajdonságok - Konzisztencia

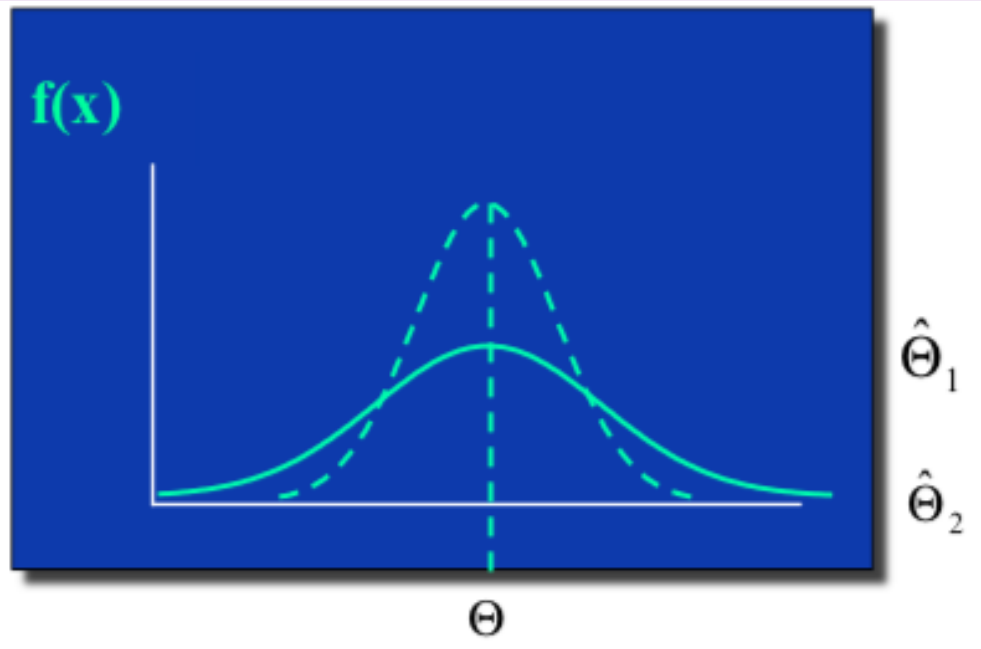
- Konzisztens a becslés, ha ingadozása a becsült paraméter körül a minta elemszámának növelésével csökken,
- azaz nagy minta esetén a becslés jól közelítse a sokasági jellemzőt.



Példa konzisztens, torzított becslésre $\theta = 0$ esetén. Piros: kis minta, zöld: közepes minta, kék: nagy minta.

Statisztikai tulajdonságok - Hatásosság

- Két becslés közül azt nevezzük hatásosabbnak, melynek kisebb a szórása.
- Egy becslés hatásos, ha minden más becslésnél hatásosabb, azaz ez a legkisebb varianciájú becslés az összes torzítatlan becslés között.



Statisztikai tulajdonságok

A mintából számított becsült paraméterek is szóródnak az elméleti értékek körül, hiszen ezek is becslések.

Az együtthatók standard hibái

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

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

Konfidencia intervallumok

- Az intervallumbecslés lényege az, hogy a pontbecslésünk ismert valószínűségi tulajdonságai segítségével adott megbízhatósági intervallumot adjunk a sokasági paraméterre.
- Ez is valószínűségi változó lesz, azaz mintáról mintára változik.
- Nagysága függ a minta elemszámától, a sokasági szórástól, és a választott megbízhatósági szinttől.

Mintavételi ingadozás

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 a paraméterekre vonatkozó konf. iv-ok

$$\hat{\alpha} \pm t_{1-\varepsilon/2}(n-2) \cdot s_{\hat{\alpha}} \quad \text{és} \quad \hat{\beta} \pm t_{1-\varepsilon/2}(n-2) \cdot s_{\hat{\beta}},$$

ahol t az $(n-2)$ szabadsági fokú t eloszlás, ε pedig a választott megbízhatósági szint.

Konfidencia intervallumok

Hasonlóan az előzőekhez az eredményváltozó $\hat{Y}$ becslésére is konstruálható konfidencia intervallum.

- Az átlagra vonatkozóan (adott X_t helyen Y várható értékére):

$$s_{\widehat{\overline{Y}}_t} = s_e \cdot \sqrt{\frac{1}{n} + \frac{(X_t - \bar{X})^2}{S_{xx}}},$$

ahonnan a konfidencia intervallum

$$\widehat{\overline{Y}}_t \pm t_{1-\varepsilon/2}(n-2) \cdot s_{\widehat{\overline{Y}}_t}$$

- A pontbecslésre (adott X_t helyen egyedi Y_t értékre)

$$s_{\widehat{Y}_t} = s_e \cdot \sqrt{1 + \frac{1}{n} + \frac{(X_t - \bar{X})^2}{S_{xx}}},$$

ahonnan a konfidencia intervallum

$$\hat{Y}_t \pm t_{1-\varepsilon/2}(n-2) \cdot s_{\widehat{Y}_t}$$

Konfidencia intervallumok - alkalmazás a példára

- A paraméterek konfidencia intervallumai:

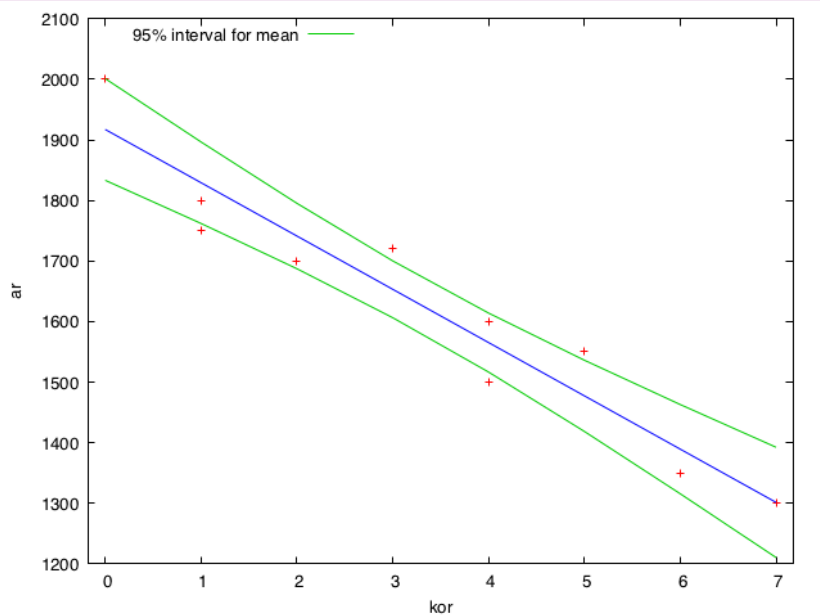
$$s_{\hat{\alpha}} = 36.38, \quad t_{0,975}(8) = 2.306,$$

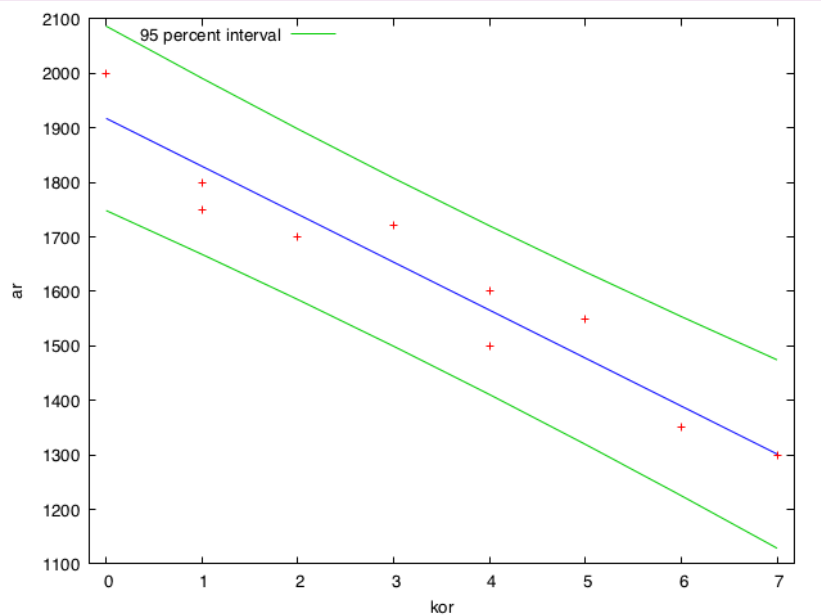
$$CI = 1917.277 \pm 36.38 \cdot 2.306 = [1833.38; 2001.17]$$

$$s_{\hat{\beta}} = 9.18, \quad t_{0,975}(8) = 2.306,$$

$$CI = -87.96 \pm 9.18 \cdot 2.306 = [-109.136; -66.789]$$

- A pontbecslések konfidencia intervallumai:





Hipotézisvizsgálatok - az eredmények ellenőrzése

- A regresszió alapvető feltétele, hogy a magyarázó- és eredményváltozó közötti korreláció nem nulla.
- Ebben az esetben a regressziós együttható (β) sem nulla.
- Előfordulhat azonban, hogy egy rossz mintavétel következtében nullától eltérő becslést kapunk olyan esetben is, amikor a két változó közt nincs semmilyen kapcsolat.
- Hipotézisvizsgálattal fogjuk ellenőrizni tehát, hogy a magyarázó- és eredményváltozó között tényleg van-e a kapcsolat.
- Azt is tudjuk ellenőrizni, hogy elegendő-e a választott magyarázó változó az eredményváltozó jellemzésére, vagy esetleg kell-e további változókat felkutatni és beépíteni a modellbe.

A paraméterek szeparált tesztelése

Azt teszteljük, hogy **a meredekségi paraméter sokasági értéke lehet-e nulla**, hiszen ekkor a nullhipotézis fennállása azt jelenti, hogy X alakulása nem befolyásolja Y alakulását, azaz a két változó közt nincs kapcsolat, tehát a modell értelmetlen.

Pontbecslések hipotézisei

A próba hipotézisei:

$$H_0: \ \beta = 0, \qquad H_1: \ \beta \neq 0,$$

melyek tesztelése adott ε szignifikancia szinten t-próbával történik:

$$t = \frac{\hat{\beta}}{s_{\hat{\beta}}} \sim t_{1-\varepsilon/2}(n-2)$$

α esete hasonlóan kezelhető, bár jelentősége kicsi, mert konstans mindig kell a modellbe! Ennek okairól később beszélünk.

Alkalmazás a példára

Példa

Alkalmazva a példára az előzőeket kapjuk, hogy

$$t = \frac{\hat{\beta}}{s_{\hat{\beta}}} = \frac{-87.96}{9.182} = -9.58.$$

A próba kétoldali, az $\varepsilon = 5\%$ szignifikancia szinthez tartozó kritikus érték $t = 2.306$. Mivel

$$|-9.58| > 2.306,$$

azaz a statiszika számított értéke a kritikus tartományba esik, tehát a **nullhipotézist elutasítjuk**.
Azaz meredekségi paraméterünk szignifikánsan nem nulla értékű, és ez egyben azt is jelenti, hogy **magyarázó változónk alakulása szignifikánsan befolyásolja az eredményváltozónk alakulását.**

Varianciaanalízis (ANOVA) a regresszióban, mely...

...**a regressziófüggvény hipotézisellenőrzésének eszköze.**

Teszteli, hogy

- vajon a modell elegendő-e, azaz minden hatást megragad-e
- kétváltozós esetben ez megegyezik a β paraméter szeparált tesztelésével
- többváltozós esetben lesz jelentősége.

Emlékezzünk a $TSS = RSS + ESS$ teljes négyzetösszeg felbontásra:

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

Világos, hogy ha

- $ESS = 0$, akkor Y teljes varianciája magyarázható az X változóval, azaz $\hat{Y}_t = Y_t \ \forall t$, tehát a kapcsolat determinisztikus, míg ha
- $ESS \neq 0$, akkor a kapcsolat sztochasztikus.

Varianciaanalízis a regresszióban

Szóródási mutatók felbontásából indulunk ki $\Rightarrow$ ANOVA teszt

Hipotézisek

$$H_0: \ \beta = 0, \qquad H_1: \ \beta \neq 0,$$

ami **ANOVA nyelven azt jelenti, hogy az X magyarázó változó szerint képzett csoportok várható értékei nem térnek el egymástól**, azaz X együtthatója a regresszióban nulla. A tesztstatisztika:

$$F = \frac{RSS/1}{ESS/(n-2)} = (n-2)\frac{R^2}{1-R^2} \sim F_{1-\varepsilon}(1, n-2).$$

- $ESS/(n-2)$ a belső szórás, azaz a szóródás azon része, melyet a csoportosítás nem magyaráz,
- $RSS/1$ pedig a csoportosítás hatása a szóródásra, azaz a külső szórás.

Varianciaanalízis a regresszióban

Szabadsági fokok:

- *TSS*: $(n-1)$, hiszen $\bar{Y}$ kiszámítását igényli (1 paraméter)
- *ESS*: $(n-2)$, hiszen $\hat{\alpha}$ és $\hat{\beta}$ becslése kell hozzá (2 paraméter)
- *RSS*: $(n-1)-(n-2)=1$, a maradék.

Variancia forrása	négyzetösszeg	szabadsági fok	átlagos négyzetösszeg	F érték
Regresszió	RSS	1	$MSR = \frac{RSS}{1}$	$F = \frac{MSR}{MSE}$
Maradék	ESS	$n-2$	$MSE = \frac{ESS}{n-2}$	
Teljes	TSS	$n-1$		

Példa

Példánkban $RSS = 372\,169.7$ és $ESS = 32\,440.3$, így

$$F = \frac{372\,169.7}{32\,440.3/8} = 91.7795 > F^*_{0.95}(1,8) = 5.31766,$$

tehát a nullhipotézis elutasítva, modellünk ezáltal releváns.

Mértékegység-váltás hatása a becslésekre

Tekintsük az

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

modellt. **Két esetet különböztetünk meg:**

- Az **eredményváltozó mértékegységének megváltozása**: legyen az új változó $Y^* = c \cdot Y$ valamely c konstans mellett (pl. ezer forint helyett forint esetén $c = 1000$). Ekkor az új modell

$$Y^* = c \cdot Y = c\alpha + c\beta X_t + c\varepsilon_t, \quad t = 1, \ldots, n.$$

- A **magyarázó változó mértékegységének megváltozatása**: legyen $X^* = c \cdot X$ valamely c konstans mellett (pl. év helyett hónapok esetén c=12). Ekkor az új modell

$$Y_t = \alpha + \beta X_t + \varepsilon_t = \alpha + \frac{\beta}{c}(cX_t) + \varepsilon_t = \alpha + \frac{\beta}{c} X_t^* + \varepsilon_t, \quad t = 1, \ldots, n$$

Mértékegység-váltás hatása a becslésekre

- Az R^2 mutató értéke nem változik, hiszen mértékegység-független mutató.
- A becslések standard hibái a fentieknek megfelelően változnak, így ezzel párhuzamosan a t-statisztikák értékei változatlanok maradnak.

Példa

A példánkban, ha az autók eladási árát (Y) ezer forint helyett forintban mérnénk, akkor

$$\hat{\alpha} = 1917.280 \cdot 1000 = 1\,917\,280,$$

$$\hat{\beta} = -87.9626 \cdot 1000 = -87\,962.6$$