Lineáris regresszió elméleti összefoglaló

Bognár Miklós

Bevezetés az Ökonometriába

0.1 Lineáris algebra, valószínűségi vektorváltozók és mátrixdifferenciálás összefoglaló

A lineáris regresszió megértéséhez elengedhetetlen, hogy tisztában legyünk néhány, lineáris algebrából ismeretes fogalommal és összefüggéssel. Ezen felül nagyon hasznos, ha ismerjük, hogy hogyan kezelendőek a valószínűségi vektorváltozók illetve a mártixdifferenciálás-kifejezések.

0.1.1 Pszeudoinverzek

Legyen $\boldsymbol{A} \in \mathbb{R}^{n \times m}, n \neq m$ nem négyzetes mátrix. Ha egy $\boldsymbol{A}x = y, x \in \mathbb{R}^{m \times 1}, y \in \mathbb{R}^{n \times 1}$ lineáris egyenletrendszer együtthatómátrixaként gondolunk rá, akkor $n \geq m$ vagy $m \geq n$ esetén rendre a *túlhatározottság* vagy *alulhatározottság* esete állna fent, az első esetben általánosságban nem lenne megoldásunk, a második esetben pedig végtelen sok megoldásunk lenne rá. Látszik, hogy az $n \neq m$ esetben nem beszélhetünk $\boldsymbol{A}^{-1}$ inverzről, helyette egy általánosabb, úgynevezett *pszeudoinverz* kell.

Egy $\boldsymbol{A} \in \mathbb{R}^{n \times m}, n > m$ mátrix *bal oldali pszeudoinverze* (Más néven *Moore-Penrose pszeudoinverz*):

$$\boldsymbol{A}^{\dagger} := (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T \in \mathbb{R}^{m \times n}$$

Figyeljük meg, hogy ha $\boldsymbol{A}^{\dagger}$-el balról megszorozzuk $\boldsymbol{A}$-t, az identitás mátrixot kapjuk, tehát bal oldalról valóban identitásként működik:

$$\boldsymbol{A}^{\dagger}\boldsymbol{A} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{I}$$

Ha jobbról szoroznánk meg:

$$\boldsymbol{A}\boldsymbol{A}^{\dagger} = \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T$$

Ez semmi más, mint a *projekció-mátrix* $\boldsymbol{A}$ oszlopvektorai által kifeszített vektortérre. Ha egy vektor ebben az oszloptérben van, rá persze identitásként hat $\boldsymbol{A}\boldsymbol{A}^{\dagger}$, ha viszont ezen kívül esik, akkor rávetíti az oszloptérre a vektort. Egy túlhatározott $\boldsymbol{A}x = y$ egyenletrendszert tehát "meg lehet oldani", ha y-t rávetítjük $\boldsymbol{A}$ oszlopterére, és megoldjuk az $\boldsymbol{A}x = \tilde{y}$ egyenletrendszert:

$$\tilde{y} = \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T y = \boldsymbol{A}\boldsymbol{A}^{\dagger}y$$

$$\boldsymbol{A}x = \tilde{y} = \boldsymbol{A}\boldsymbol{A}^{\dagger}y$$

$$\boldsymbol{A}^{\dagger}\boldsymbol{A}x = \boldsymbol{A}^{\dagger}\boldsymbol{A}\boldsymbol{A}^{\dagger}y$$

$$x = \boldsymbol{A}^{\dagger}y$$

Az $n < m$ esetben alulhatározottság áll fenn, itt *jobb oldali pszeudoinverzről* beszélhetünk:

$$\boldsymbol{A}^{\ddagger} := \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{A}^T)^{-1} \in \mathbb{R}^{n \times m}$$

Bár ezt nem fogjuk a későbbiekben használni, érdemes lehet megjegyezni, hogy a jobb oldali pszeudoinverzzel való balról szorzás esetén - hasonlóan a bal oldali pszeudoinverzhez - projekciómátrixot kapunk, csak most $\boldsymbol{A}$ sorvektorai által kifeszített vektortérre.

0.1.2 Valószínűségi vektorváltozók

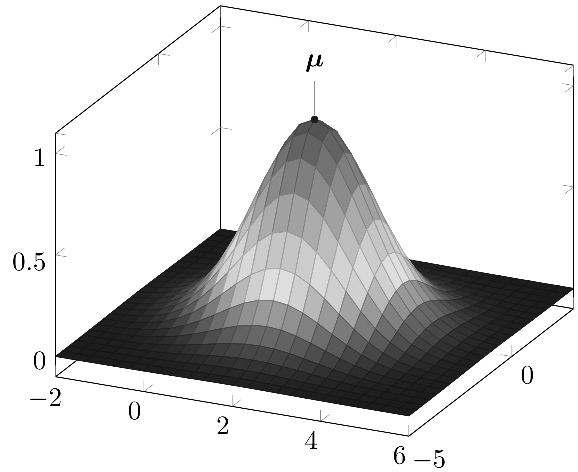
Egy $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_n]^T$ vektort *valószínűségi vektorváltozónak* hívunk, ha $\forall i$-re ξ_i skalárértékű valószínűségi változó. A továbbiakban csak a vektorértékű normális eloszlást követő valószínűségi vektorváltozókkal foglalkozunk, ezek formálisan felírva:

$$\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

ahol $\boldsymbol{\mu} \in \mathbb{R}^{n \times 1}$ a várható értékek vektora, $\boldsymbol{\Sigma}$ pedig a *variancia-kovarianca mátrix.* Természetesen $Var[\boldsymbol{\xi}] = \boldsymbol{\Sigma}$. Természetesen $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$ pozitív szemidefinit és szimmetrikus mártix. Az $n = 1$ esettel analóg módon $\boldsymbol{\xi}$ sűrűségfüggvénye

$$f_{\boldsymbol{\xi}}(\xi_1, \ldots, \xi_n) = \frac{e^{-\frac{1}{2}(\boldsymbol{\xi}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\xi}-\boldsymbol{\mu})}}{\sqrt{(2\pi)^n|\boldsymbol{\Sigma}|}}$$

A sűrűségfüggvény $n = 2$ esetben $\boldsymbol{\mu} = [2, -1]^T$ és $\boldsymbol{\Sigma} = \boldsymbol{I}$ várhatóérték és kovariancia mátrix mellett:



Egy $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ mátrix mellett a skaláresethez hasonlóan

$$Var[\boldsymbol{A}\boldsymbol{\xi}] = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T$$

$$\mathbb{E}[\boldsymbol{A}\boldsymbol{\xi}] = \boldsymbol{A}\mathbb{E}[\boldsymbol{\xi}]$$

$\boldsymbol{\Sigma}$ kovariancia mátrixot kifejezhetjük várható értékekkel is:

$$\boldsymbol{\Sigma} = \mathbb{E}[(\boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}])(\boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}])^T] = \mathbb{E}[\boldsymbol{\xi}\boldsymbol{\xi}^T] - \mathbb{E}[\boldsymbol{\xi}]\mathbb{E}[\boldsymbol{\xi}^T]$$

$\boldsymbol{\Sigma}$ alakja:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & Cov[\xi_1, \xi_2] & \dots & Cov[\xi_1, \xi_n] \\ Cov[\xi_2, \xi_1] & \sigma_2^2 & \dots & Cov[\xi_2, \xi_n] \\ \vdots & \vdots & \ddots & \vdots \\ Cov[\xi_n, \xi_1] & Cov[\xi_n, \xi_2] & \dots & \sigma_n^2 \end{bmatrix}$$

ahol $\sigma_1^2, \dots, \sigma_n^2$ rendre $\xi_1, \dots, \xi_n$ varianciái.

0.1.3 Mátrixdifferenciálás nagyon röviden

Legyenek $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^{k \times 1}$ vektorok. Ekkor

$$\frac{\partial \boldsymbol{a}^T \boldsymbol{b}}{\partial \boldsymbol{b}} = \frac{\partial \boldsymbol{b}^T \boldsymbol{a}}{\partial \boldsymbol{b}} = \boldsymbol{a}$$

Ha $\boldsymbol{A} \in \mathbb{R}^{k \times k}$ mátrix, akkor

$$\frac{\partial \boldsymbol{b}^T \boldsymbol{A} \boldsymbol{b}}{\partial \boldsymbol{b}} = 2\boldsymbol{A}\boldsymbol{b}$$

Ha $\boldsymbol{A}$ szimmetrikus, akkor ezen felül

$$2\boldsymbol{A}\boldsymbol{b} = 2\boldsymbol{b}^T\boldsymbol{A}$$

Legyen $\boldsymbol{\beta} \in \mathbb{R}^{k \times 1}$, $\boldsymbol{A} \in \mathbb{R}^{n \times k}$ és $\boldsymbol{y} \in \mathbb{R}^{n \times 1}$. Ekkor

$$\frac{\partial 2\boldsymbol{\beta}^T \boldsymbol{A}^T \boldsymbol{y}}{\partial \boldsymbol{\beta}} = \frac{\partial 2\boldsymbol{\beta}^T (\boldsymbol{A}^T \boldsymbol{y})}{\partial \boldsymbol{\beta}} = 2\boldsymbol{A}^T\boldsymbol{y}$$

0.2 A lineáris regresszió és az OLS eljárás

A regresszió kiindulópontja egy $\mathscr{X}$ normális eloszlású sokaság, melynek minden tagja rendelkezik $\boldsymbol{x}_i$ *featurevektor*-ral, avagy magyarázó változó-vektorral (ezek a *regresszorok*), illetve egy-egy skalár y_i *label*-lel, avagy magyarázott változóval (amiket a regresszorok magyaráznak egy lineáris modell alapján, ezt később jobban kifejtjük). A sokaságból n darab mintát veszünk (megfigyelést végzünk), *a minták iid. normális eloszlásúak*, ami persze azt jelenti, hogy *minden magyarázó változó-vektor egy vektorértékű normális eloszlású valószínűségi vektorváltozó.*

A megfigyelt magyarázó változó-vektorokat soronként egymásra rakva felépítünk egy úgynevezett *design mátrix*ot, melyet mostantól $\boldsymbol{X}$-el jelölünk. Minden $\boldsymbol{x}_i$ magyarázó változó-vektor első eleme konstans 1, ez tölti be az intercept, avagy kétdimenziós esetben az y-tengellyel való metszéspont szerepét. n darab megfigyelés és p elemszámú magyarázó változó-vektorral $\boldsymbol{X}$ alakja a következő:

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{1,1} & \dots & x_{1,p-1} \\ 1 & x_{2,1} & \dots & x_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,p-1} \end{bmatrix}_{n \times p}$$

A megfigyelt magyarázott változókat szintén sorokba tömörítjük, így mivel mindegyik skalár, egy vektort kapunk:

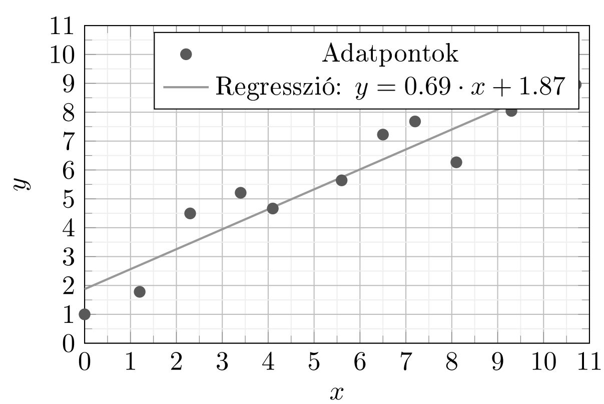
$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

A lineáris regresszió feladata, hogy a lineáris

$$\boldsymbol{X\beta} + \boldsymbol{\epsilon} = \boldsymbol{y}$$

modell mellett megtalálja azt a $\boldsymbol{\beta} \in \mathbb{R}^{p \times 1}$ *együttható-vektort*, amelyre az $\boldsymbol{X\beta} = \boldsymbol{y}$ egyenletrendszer " legjobban teljesül". $\boldsymbol{\epsilon}$ az úgynevezett *hibavektor*. Az OLS, avagy *Ordinary Least Squares* becslési eljárást pontosan ezt a $\boldsymbol{\beta}$ paramétervektort becsüli, a paraméterbecslést kétféleképpen is levezetjük. Az OLS elnevezés valójában az analitikus levezetésből nyer értelmet a legkönnyebben, de először a - szerintem intuitívabb - projekciós módszert nézzük meg.

A lineáris regresszió egy darab regresszor (magyarázó változó) esetén az alábbi ábrával szemléltethető:



Itt $\hat{\boldsymbol{\beta}}$ paraméterbecslés vektor alakja

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} 1,87 \\ 0,69 \end{bmatrix}$$

Azt, hogy hogyan kaptuk meg $\hat{\boldsymbol{\beta}}$ paraméterbecslést, a következő fejezetek tárgyalják részletesen. Ezen kívül külön foglalkozunk majd a fenti egyváltozós regresszióval is (a $p = 2$-es eset).

0.2.1 Az OLS-becslés geometriai értelmezése

Szinte mindig $n > p$, így az egyenletrenszer *túlhatározott*, és nagyon specifikus esetektől eltekintve nem létezik egzakt megoldása. Az első fejezetben azonban láttuk, hogy a bal oldali pszeudoinverz pontosan ezt a problémát orvosolja. A jelölési konvenció a megoldásból nyert *paraméter-becslésre* $\hat{\boldsymbol{\beta}}$, ami a mintavétel véletlenszerűségéből adódóan maga is vektorértérkű valószínűségi változó ($\hat{\boldsymbol{\beta}}$ pontos eloszlásáról a későbbiekben lesz szó):

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

Ebben az esetben $\boldsymbol{y}$-t az $\boldsymbol{X}$ design mátrix oszlopterére vetítettük. Legyen $\boldsymbol{e}$ a valós y_i-k és a $\boldsymbol{X}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{y}}$ modellbecslés által prediktált $\hat{y}_i$-k közti eltérések vektora (sokszor $\boldsymbol{e}$-t $\widehat{\boldsymbol{\epsilon}}$-ként is jelölik):

$$\boldsymbol{e} = \begin{bmatrix} y_1 - \hat{y_1} \\ y_2 - \hat{y_2} \\ \vdots \\ y_n - \hat{y_n} \end{bmatrix}$$

0.2.2 Az OLS-becslés mint szélsőérték-feladat

$\hat{\boldsymbol{\beta}}$ paraméterbecslés-vektort megkaphatjuk úgy is, ha tekintjük az alábbi minimalizálási feladatot:

$$\boldsymbol{e}^T\boldsymbol{e} \to \min_{\hat{\boldsymbol{\beta}}}$$

azaz minimalizáljuk a becsült $\hat{y}_i$ és tényleges y_i magyarázott változók közötti négyzetösszeget. $\boldsymbol{e}^T\boldsymbol{e}$-t RSS, azaz *sum of squared residuals* néven is emlegetik. Írjuk ki a hiba-négyzetösszeg teljes alakját:

$$\boldsymbol{e}^T\boldsymbol{e} = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{y}^T\boldsymbol{y} - \hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{y}^T\boldsymbol{y} - 2\hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{y} + \hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}}$$

Itt felhasználtuk, hogy a transzponálás "megfordítja a szorzatot", illetve hogy skalár transzponáltja önmaga, így $\boldsymbol{y}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = (\boldsymbol{y}^T\boldsymbol{X}\hat{\boldsymbol{\beta}})^T = \hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{y}$. A minimalizációhoz vennünk kell a kifejezés $\hat{\boldsymbol{\beta}}$ szerinti deriváltját, majd 0-val egyenlővé tenni:

$$\frac{\partial \boldsymbol{e}^T\boldsymbol{e}}{\partial \hat{\boldsymbol{\beta}}} = -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = 0$$

Ebből megkapjuk az úgynevezett *normálegyenletet*:

$$(\boldsymbol{X}^T\boldsymbol{X})\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{y}$$

$(\boldsymbol{X}^T\boldsymbol{X})$ szimmetrikus, és ha feltesszük, hogy létezik inverze, akkor balról beszorozva mindét oldalt:

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

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

Látható, hogy pontosan ugyanaz jött ki, mint a pszeudoinverzes levezetésben. Míg ez utóbbi pusztán analitikus úton jutott el $\hat{\boldsymbol{\beta}}$-hoz, a pszeudoinverzes módszert geometrikus úton is el lehet képzelni.

0.3 Az OLS-becslés tulajdonságai

Vegyük az OLS paraméterbecslés normálegyenletét, és figyeljük meg, hogy $\boldsymbol{X}^T\boldsymbol{e} = \boldsymbol{0}$:

$$(\boldsymbol{X}^T\boldsymbol{X})\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{y}$$

A modellből adódóan $\boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{e}$ behelyettesítéssel:

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

$$(\boldsymbol{X}^T\boldsymbol{X})\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})\hat{\boldsymbol{\beta}} + \boldsymbol{X}^T\boldsymbol{e}$$
$$\boldsymbol{X}^T\boldsymbol{e} = \boldsymbol{0}$$

valóban. Ez azt jelenti, hogy *minden magyarázó változó (regresszor) korrelálatlan a hibával*, pontosabban megfogalmazva *a regresszorok és a hibák mintakorrelációja zérus.* Mivel $\boldsymbol{X}$ mátrix első oszlopa konstans 1-eket tartalmaz, így $\hat{\beta}_0$ maga az intercept lesz, és emiatt

$$\sum_{i=1}^{n} e_i = 0$$

azaz a hibák összege 0. Ha leosztunk n-nel:

$$\frac{1}{n}\sum_{i=1}^{n} e_i = \overline{\boldsymbol{e}}$$

azaz a hibatagok (*rezidiumok*) mintaátlaga - ami persze torzítatlan becslése a várható értéknek - 0, tehát $\mathbb{E}[\boldsymbol{e}] = \boldsymbol{0}$.

Egy másik, ugyancsak fontos tulajdonság a predikciós formulából következik:

$$\hat{\boldsymbol{y}}^T\boldsymbol{e} = (\boldsymbol{X}\hat{\boldsymbol{\beta}})^T\boldsymbol{e} = \hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{e} = 0$$

azaz *a becsült $\hat{y}_i$-ok korrelálatlanok a rezidiumokkal.* Így azt is beláthatjuk, hogy *a modell által prediktált és a tényleges magyarázott változók mintaátlagai megegyeznek*:

$$\overline{\boldsymbol{y}} = \overline{\hat{\boldsymbol{y}}}$$

Felmerülhet a kérdés, hogy mindig létezik-e $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Abban az esetben, ha $\boldsymbol{X}$ oszloprangja kisebb, mint p, tehát *tökéletes multikollinearitás* áll fenn, akkor $\boldsymbol{X}$ szinguláris értékei között lesz 0, így $\boldsymbol{X}^T\boldsymbol{X}$ sajátértékei között is, azaz $\boldsymbol{X}^T\boldsymbol{X}$ nem lesz invertálható. Ezentúl tehát feltételezzük, hogy nem áll fenn tökéletes multikollinearitás.

0.3.1 A Gauss-Markov feltételezések

A Gauss-Markov feltételezések biztosítják, hogy az OLS eljárással kapott $\hat{\boldsymbol{\beta}}$ paraméterbecslésünk *BLUE*, azaz *Best Linear Unbiased Estimator* lesz. Ez azt jelenti, hogy nem fogunk tudni találni olyan - nem az OLS eljárással kapott - paraméterbecslést $\boldsymbol{\beta}$-ra, ami lineáris, torzítatlan, és kisebb mintavarianciával rendelkezne, mint $\hat{\boldsymbol{\beta}}$.

Formálisan kimondva az első Gauss-Markov feltétel a már látott modellegyenlet:

$$\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \boldsymbol{y}$$

A második Gauss-Markov feltétel szerint $\boldsymbol{X}$ oszloprangja megegyezik oszlopainak számával, az oszlopok mind lineárisan függetlenek, azaz nincs zérus szinguláris értéke. Ezt $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ létezésénél már feltételeztük, formálisan ez is egyike a feltételeknek.

A harmadik feltétel szerint

$$\mathbb{E}[\boldsymbol{\epsilon} \mid \boldsymbol{X}] = \boldsymbol{0}$$
$$\mathbb{E}\begin{bmatrix}\epsilon_1 \mid \boldsymbol{X} \\ \epsilon_2 \mid \boldsymbol{X} \\ \vdots \\ \epsilon_n \mid \boldsymbol{X}\end{bmatrix} = \begin{bmatrix}\mathbb{E}[\epsilon_1] \\ \mathbb{E}[\epsilon_2] \\ \vdots \\ \mathbb{E}[\epsilon_n]\end{bmatrix} = \boldsymbol{0}$$

Ez azt jelenti, hogy a modell szerinti hibatag várható értékét nem befolyásolja egyik magyarázó változó sem. Ebből következőleg

$$\mathbb{E}[\boldsymbol{y} \mid \boldsymbol{X}] = \mathbb{E}[\boldsymbol{X\beta} + \boldsymbol{\epsilon} \mid \boldsymbol{X}] = \boldsymbol{X\beta}$$

A negyedik feltétel a hibák kovariancia mátrixára vonatkozik, mégpedig

$$\mathbb{E}[\boldsymbol{\epsilon\epsilon}^T \mid \boldsymbol{X}] = \sigma^2 \boldsymbol{I}$$

A hibatagok *homoszkedasztikusak és korrelálatlanok*, azaz azonosan σ^2 varianciájúak és $\forall i \neq j : Cov[\epsilon_i, \epsilon_j] = 0$. Ha kiírjuk $\boldsymbol{\epsilon\epsilon}^T$ mátrixformáját:

$$\mathbb{E}[\boldsymbol{\epsilon\epsilon}^T \mid \boldsymbol{X}] = \mathbb{E}\begin{bmatrix} \epsilon_1^2 \mid \boldsymbol{X} & \epsilon_1\epsilon_2 \mid \boldsymbol{X} & \dots & \epsilon_1\epsilon_n \mid \boldsymbol{X} \\ \epsilon_2^1 \mid \boldsymbol{X} & \epsilon_2\epsilon_2 \mid \boldsymbol{X} & \dots & \epsilon_2\epsilon_n \mid \boldsymbol{X} \\ \vdots & \vdots & \ddots & \vdots \\ \epsilon_n^1 \mid \boldsymbol{X} & \epsilon_n\epsilon_2 \mid \boldsymbol{X} & \dots & \epsilon_n^2 \mid \boldsymbol{X} \end{bmatrix}$$

és persze $\forall i : \mathbb{E}[\epsilon_i \mid \boldsymbol{X}] = 0$ miatt a fenti mátrix diagonálisában ϵ_i-k varianciái, a többi helyen pedig a kovarianciák, amik a feltétel szerint 0-k, így $\mathbb{E}[\boldsymbol{\epsilon e}^T \mid \boldsymbol{X}]$ kovarianciamátrix valóban diagonális, a homoszkedaszticitás feltétele mellett pedig minden diagonális elem σ^2. Mostantól a hibatagok varianciáját $\boldsymbol{\Sigma}$ fogja jelölni, $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$.

Az utolsó feltétel szerint a hibatagok normális eloszlást követnek:

$$\boldsymbol{\epsilon} \mid \boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$$

Kijelenthetjük tehát, hogy y_i-k varianciáját nem csak $\boldsymbol{x}_i$-ek magyarázzák, hanem σ^2 *magyarázatlan variancia* is. Úgy is megfogalmazhatjuk, hogy a modell szerint minden $\boldsymbol{y}$ magyarázott változó-vektor regresszorok szerinti feltételes eloszlása

$$\boldsymbol{y} \mid \boldsymbol{X} \sim \mathcal{N}(\boldsymbol{X\beta}, \boldsymbol{\Sigma})$$

Lássuk be, hogy a feltételek teljesülése mellett $\hat{\boldsymbol{\beta}}$ valóban torzítatlan becslést ad $\boldsymbol{\beta}$-ra! Láttuk, hogy $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$, és a modell szerinti $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ behelyettesítéssel

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{X\beta} + \boldsymbol{\epsilon})$$

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon},$$

mindkét oldalon véve a várható értéket:

$$\mathbb{E}[\hat{\boldsymbol{\beta}}] = \mathbb{E}[\boldsymbol{\beta}] + \mathbb{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon}] = \boldsymbol{\beta} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\mathbb{E}[\boldsymbol{X}^T\boldsymbol{\epsilon}]$$

Mivel a Gauss-Markov feltételek egyike, hogy $\mathbb{E}[\boldsymbol{X}^T\boldsymbol{\epsilon}] = \boldsymbol{0}$, így

$$\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$$

ezzel készen is vagyunk.

0.3.2 $\hat{\boldsymbol{\beta}}$ varianciája

A hibavektor variancia-kovariancia mátrixához hasonlóan képezhetjük $\hat{\boldsymbol{\beta}}$ valószínűségi vektorváltozó variancia-kovariancia mátrixát:

$$Var[\hat{\boldsymbol{\beta}}] = \mathbb{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T]$$

Láttuk, hogy

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon} \Longrightarrow \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon}$$

$$\mathbb{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T] = \mathbb{E}\left[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon}((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon})^T\right]$$

A transzponálás "szorzatmegfordító" tulajdonságából következően, illetve $\boldsymbol{X}^T\boldsymbol{X}$ szimmetrikus voltából

$$Var[\hat{\boldsymbol{\beta}}] = \mathbb{E}\left[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right]$$

$$Var[\hat{\boldsymbol{\beta}}] = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T]\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

Itt válik igazán fontossá, hogy $\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T]$ variancia-kovariancia mátrix alakja $\sigma^2\boldsymbol{I}$, így σ^2 kiemelhető a mátrixszorzások elé, az identitást pedig triviálisan nem szükséges kiírni:

$$Var[\hat{\boldsymbol{\beta}}] = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

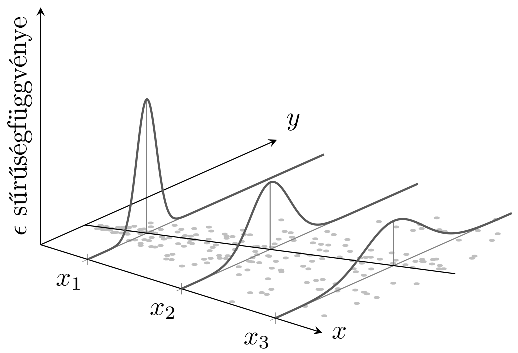
A mátrixszorzás asszociativitásából pedig a

$$Var[\hat{\boldsymbol{\beta}}] = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

végleges alakot kapjuk. Ugyanez megkapható az első fejezetben bemutatott $Var[\boldsymbol{A\xi}] = \boldsymbol{A}Var[\boldsymbol{\xi}]\boldsymbol{A}^T$ transzformált variancia képlettel is, $\boldsymbol{\xi}$ helyett $\boldsymbol{y}$, $\boldsymbol{A}$ helyett pedig $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ transzformáció mátrixxal (már ha $\boldsymbol{X}$-eket fixnek tekintjük). A várható értékes felírásból látszik, hogy persze $Var[\hat{\boldsymbol{\beta}}]$ alakja

$$\mathbb{E}[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T] = \begin{bmatrix} Var[\hat{\beta}_1] & Cov[\hat{\beta}_1,\hat{\beta}_2] & \dots & Cov[\hat{\beta}_1,\hat{\beta}_p] \\ Cov[\hat{\beta}_2,\hat{\beta}_1] & Var[\hat{\beta}_2] & \dots & Cov[\hat{\beta}_2,\hat{\beta}_p] \\ \vdots & \vdots & \ddots & \vdots \\ Cov[\hat{\beta}_p,\hat{\beta}_1] & Cov[\hat{\beta}_p,\hat{\beta}_2] & \dots & Var[\hat{\beta}_p] \end{bmatrix}$$

Ha n elég nagy, akkor $\hat{\boldsymbol{\beta}}$ eloszlása *megközelítőleg normális lesz.* Csupán érdekesség, de el lehet képzelni, hogy heteroszkedaszticitás ($\exists i,j : \sigma_i^2 \neq \sigma_j^2$) és $p = 2$ mellett a modell az alábbi ábrával szemléltethető:



A $\hat{\boldsymbol{\beta}}$ varianciája formulában szereplő σ^2 hibavariancia maga is becslésre szorul, ennek $\hat{\sigma}^2$ torzítatlan becslése a tényleges $\boldsymbol{e}$ hibatagokkal számolható:

$$\hat{\sigma}^2 = \frac{\boldsymbol{e}^T\boldsymbol{e}}{n-p}$$

ezt azonban nem fogjuk belátni.

0.4 A $p = 2$-es egyszerű modell

Nézzük meg, hogy eddig látott paraméterbecslés és becslés-variancia hogy néz ki a legegyszerűbb, egy darab konstans interceptet és egy darab magyarázó változót tartalmazó OLS-el becsült modellben. A modell egyenlete minden $i = 1 \dots n$ megfigyelésre

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Az $\boldsymbol{X}$ design mátrixunk most

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \in \mathbb{R}^{n \times 2}$$

lesz, $\hat{\boldsymbol{\beta}}$ paraméterbecslés pedig

$$\hat{\boldsymbol{\beta}} = \left(\begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{bmatrix}$$

A 2×2-es mátrixok invertálása könnyen megy:

$$\hat{\boldsymbol{\beta}} = \frac{1}{n \sum_i x_i^2 - (\sum_i x_i)^2} \begin{bmatrix} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & n \end{bmatrix} \begin{bmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{bmatrix} = \frac{1}{n \sum_i x_i^2 - (\sum_i x_i)^2} \begin{bmatrix} \sum_i x_i^2 \sum_i y_i - \sum_i x_i \sum_i x_i y_i \\ -\sum_i x_i \sum_i y_i + n \sum_i x_i y_i \end{bmatrix} =$$

$$= \begin{bmatrix} \frac{n(\frac{1}{n}\sum_i x_i^2) \cdot n(\frac{1}{n}\sum_i y_i) - n(\frac{1}{n}\sum_i x_i) \cdot n(\frac{1}{n}\sum_i x_i y_i)}{n^2(\frac{1}{n}\sum_i x_i^2) - n^2(\frac{1}{n}\sum_i x_i)^2} \\ \frac{n^2 \frac{1}{n}\sum_i x_i y_i - n(\frac{1}{n}\sum_i x_i) \cdot n(\frac{1}{n}\sum_i y_i)}{n^2(\frac{1}{n}\sum_i x_i^2) - n^2(\frac{1}{n}\sum_i x_i)^2} \end{bmatrix}$$

Az n elemű mintából képzett *mintaátlag* semmi más, mint $\frac{1}{n}\sum_i x_i$ illetve $\frac{1}{n}\sum_i y_i$, a kovariancia x és y között pedig $\mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y]$, n elemű - a várható értéket torzítatlanul becsülő - mintaátlagokkal ez persze semmi más, mint az *empirikus kovariancia* $empcov[x,y] = \frac{1}{n}\sum_i x_i y_i - (\frac{1}{n}\sum_i x_i)(\frac{1}{n}\sum_i y_i)$. x varianciája $\mathbb{E}[x^2] - \mathbb{E}^2[x]$-ként áll elő, $\mathbb{E}[x^2]$ empirikus becslése pedig $\frac{1}{n}\sum_i x_i^2$. A vektor mindkét elemében n^2-el leosztva látható, hogy a nevezőkben pontosan x mintából számolt varianciája ($empvar$) van, míg a vektor második elemének számlálója pontosan x és y mintából számolt kovarianciája. A vektor első elemének számlálójában $\overline{x^2} \cdot \overline{y} - \overline{x} \cdot \overline{xy}$ áll. Jelölje mostantól a mintából számolt varianciát és kovarianciát $\widehat{Var}$ és $\widehat{Cov}$, ezzel a paraméterbecslés alakja

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \frac{\overline{x^2} \cdot \overline{y} - \overline{x} \cdot \overline{xy}}{\widehat{Var}[x]} \\ \frac{\widehat{Cov}[x,y]}{\widehat{Var}[x]} \end{bmatrix}$$

Azt kaptuk tehát, hogy a legegyszerűbb egyváltozós regresszió becsült paraméterei

$$\hat{\beta}_0 = \frac{\overline{x^2} \cdot \overline{y} - \overline{x} \cdot \overline{xy}}{\widehat{Var}[x]}$$

$$\hat{\beta}_1 = \frac{\widehat{Cov}[x,y]}{\widehat{Var}[x]}$$

Sokszor a mintaszámmal normálatlan empirikus kovarianciát és varianciát S_{xy} és S_{xx} jelöléssel látják el:

$$S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2,$$

ezekkel felírva β_1 becslését:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

β_0 becslésének alakja β_1 ismeretében is kiszámolható, és sokszor ez a módszer sokkal kényelmesebb (már ha ismerjük $\hat{\beta}_1$ értékét):

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

Ez nem csak intuitívan értelmezhető ("Az átlagos y semmi más, mint az y-tengellyel való metszéspont és $\hat{\beta}_1\overline{x}$ összege"), hanem formálisan is levezethető a modell egyenletéből (meg abból, hogy beláttuk, hogy a paraméterbecslés torzítatlan a feltevéseink mellett, illetve hogy a hibatagok várható értéke 0):

$$y = \beta_0 + \beta_1 x + \epsilon$$

$$\mathbb{E}[y] = \beta_0 + \beta_1 \mathbb{E}[x]$$
$$\beta_0 = \mathbb{E}[y] - \beta_1 \mathbb{E}[x]$$

A várhatóérték-operátor helyett persze a mintaátlagokkal dolgozva:

$$\beta_0 = \overline{y} - \beta_1 \overline{x}$$

valóban.

0.4.1 $\hat{\boldsymbol{\beta}}$ varianciája és az R^2 mutató

Láttuk, hogy a paraméterbecslés varianciája

$$Var[\hat{\boldsymbol{\beta}}] = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

A már levezetett $p = 2$-es design mátrixxal dolgozva:

$$Var[\hat{\boldsymbol{\beta}}] = \sigma^2 \left(\begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \right)^{-1} = \sigma^2 \frac{1}{n\sum_i x_i^2 - (\sum_i x_i)^2} \begin{bmatrix} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & n \end{bmatrix}$$

Használjuk ki az empirikus variancia képletét:

$$n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2 = n\sum_{i=1}^n (x_i - \overline{x})^2$$

Innen könnyen látszik, hogy

$$Var[\hat{\beta}_0] = \sigma^2 \frac{\sum_{i=1}^n x_i^2}{n\sum_{i=1}^n (x_i - \overline{x})^2}$$
$$Var[\hat{\beta}_1] = \sigma^2 \frac{1}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

Kimondhatjuk tehát, hogy ahogy σ^2 nő, úgy nő a paraméterbecslésünk varianciája, avagy *bizonytalansága* is. Hasonlítsuk össze az általános esetben kapott $\hat{\boldsymbol{\beta}}$ variancia képletét β_1 varianciáéval:

$$\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

$$\sigma^2(\sum_{i=1}^n (x_i - \overline{x})^2)^{-1}$$

A 2×2-es mátrixszorzást elvégezve tényleg azt kaptuk, hogy az egyváltozós regresszió esetén S_{xx} semmi más, mint az $\boldsymbol{X}^T\boldsymbol{X}$ centralizálatlan regresszor-kovariancia mátrix.

Nagyon fontos - és ezért itt is kihangsúlyozandó - hogy *y varianciája kettő forrásból jön: a regresszorok varianciájából és a regresszorok által nem magyarázott hibavarianciából.* Írjuk ezt az összefüggést fel a mi esetünkben a modellegyenlet segítségével (persze a regresszorok és a hibák korrelálatlansága mellett):

$$Var[\boldsymbol{y}] = \beta_1^2 Var[\boldsymbol{x}] + Var[\boldsymbol{\epsilon}]$$

Itt kihasználtuk, hogy a modell szerint β_0 konstans, így zérus varianciája van. $Var[\boldsymbol{\epsilon}]$ hibavariancia az a része y varianciájának, amit nem magyaráznak a regresszorok. Ha $Var[\boldsymbol{\epsilon}]$ kicsi, ez annyit jelent, hogy a becsült $\hat{y}$-ok és a tényleges y-ok közel vannak egymáshoz, azaz a regresszióval nagyon jól becsülhetjük a valódi y értékeket.

Legyen

$$R^2 := \frac{\beta_1^2 Var[\boldsymbol{x}]}{Var[\boldsymbol{y}]}$$

az arány, amiben a regresszorok varianciája magyarázza a magyarázott változó teljes varianciáját. R^2 0 és 1 közötti szám, minél közelebb van 1-hez, annál jobban becsülhető y a regresszorokkal. β_1 becslését beírva adódik:

$$R^2 = \frac{|Cov[\boldsymbol{x}, \boldsymbol{y}]|^2}{Var[\boldsymbol{x}]Var[\boldsymbol{y}]}$$

R^2 a regresszió "erősségét" mutatja, így a normálatlan empirikus kovarianciákkal és varianciákkal (S_{xy}, S_{xx}, S_{yy}):

$$R^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}}$$

Itt persze $S_{yy} = \sum_i (y_i - \overline{y})^2$ Vezessük be az alábbi jelöléseket:

$$SST = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

$$SSE = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$$

$$SSR = \sum_{i=1}^{n} e_i^2$$

SST a *Sum of Squares Total*, SSE a *Sum of Squares Explained*, SSR pedig a *Sum of Squares Residual.* Az előbbi varianciafelbontásból könnyen látszik, hogy mivel SSE a regresszorok által magyarázott variancia, SSR pedig a magyarázatlan variancia:

$$SST = SSE + SSR$$

R^2-et az előbbihez hasonlóan, csak most az új jelölésekkel felírva:

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

(Az irodalomban néha - zavaró módon - Az SSE a hibák négyzetösszegét jelenti, mint Sum of Squares Error, és az SSR jelenti a magyarázott varianciát, mint Sum of Squares Regression.)