

Lineáris regresszió elméleti összefoglaló

Bognár Miklós

Bevezetés az Ökonometriába

Contents

0	Matematikai összefoglaló	2
0.1	Pseudoinverzek	2
0.2	*Mátrixok szinguláris értékei, SVD	3
0.3	Valószínűségi vektorváltozók	3
0.4	Mátrixdifferenciálás nagyon röviden	4
0.5	*Pont és eloszlás Mahalanobis távolsága	5
1	A lineáris regresszió	6
1.1	A regressziós modell	6
1.2	Az Ordinary Least Squares (OLS) becslési eljárás	7
1.2.1	Az OLS-becslés geometriai értelmezése	7
1.2.2	Az OLS-becslés mint szélsőérték-feladat	8
1.2.3	Az OLS-becslés tulajdonságai	8
1.3	A Gauss-Markov feltételezések	9
1.3.1	$\hat{\beta}$ varianciája	10
1.3.2	* $\hat{\beta}$ eloszlása	11
1.3.3	Multikollinearitás	12
1.3.4	A hibavariancia becslése	12
1.4	A $p = 2$ -es egyváltozós regresszió	13
1.4.1	$\hat{\beta}$ varianciája egyváltozós regresszió esetén	14
1.5	Az R^2 mutató	15
1.6	Kihagyott változó bias - Omitted Variable Bias	16
1.7	MSE és a bias-variancia tradeoff	16
1.8	A heteroszkedaszticitás kezelése, a GLS eljárás	17
1.8.1	*A GLS becslés analitikus levezetése	18
1.8.2	A Feasible Generalized Least Squares (FGLS) eljárás	19

A *-al jelölt fejezetek/alfejezetek tudtommal nem képezik részét az anyagnak, azonban (szerintem) érdekesek, és segíthetnek jobban megérteni a lineáris regressziót.

Matematikai 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 mátrixdifferenciálás-kifejezések.

.1 Pszeudo inverzek

Legyen $\mathbf{A} \in \mathbb{R}^{n \times m}$, $n \neq m$ nem négyzetes mátrix. Ha egy $\mathbf{A}x = y$, $x \in \mathbb{R}^{m \times 1}$, $y \in \mathbb{R}^{n \times 1}$ 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 \mathbf{A}^{-1} inverzről, helyette egy általánosabb, úgynevezett *pszeudo inverz* kell.

Egy $\mathbf{A} \in \mathbb{R}^{n \times m}$, $n > m$ mátrix *bal oldali pszeudo inverze* (Más néven *Moore-Penrose pszeudo inverz*):

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

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

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

Ha jobbról szoroznánk meg:

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

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

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

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

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

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

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

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

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

.2 *Mátrixok szinguláris értékei, SVD

Legyen $\mathbf{A} \in \mathbb{C}^{n \times m}$ tetszőleges komplex mátrix. Ekkor \mathbf{A} szinguláris érték felbontása (Singular Value Decomposition - SVD):

$$\mathbf{A} = \mathbf{U} \mathbf{S} \mathbf{V}^*$$

ahol $\mathbf{U} \in \mathbb{C}^{n \times n}$ és $\mathbf{V} \in \mathbb{C}^{m \times m}$ unitér mátrixok, és $\mathbf{S} \in \mathbb{R}^{n \times m}$ kvázi-diagonális, azaz $n > m$ esetben

$$\mathbf{S} = \begin{bmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_m \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix}_{n \times m}$$

Ilyen felbontás *mindig* létezik, bármilyen is legyen \mathbf{A} dimenziója. Ha \mathbf{A} valós mátrix, akkor \mathbf{U} és \mathbf{V} ortogonálisak, és így persze a konjugált transzponálás ekvivalens lesz a transzponálással. \mathbf{S} σ elemei a szinguláris értékei \mathbf{A} -nak. Figyeljük meg, hogy

$$\mathbf{A}^T \mathbf{A} = \mathbf{V} \mathbf{S}^T \mathbf{U}^T \mathbf{U} \mathbf{S} \mathbf{V}^T = \mathbf{V} \mathbf{S}^T \mathbf{S} \mathbf{V}^T,$$

azaz $\mathbf{A}^T \mathbf{A}$ spektrálfelbontása lesz. \mathbf{V} oszlopai tehát $\mathbf{A}^T \mathbf{A}$ sajátvektorai lesznek, míg hasonlóan belátható, hogy \mathbf{U} oszlopai pedig $\mathbf{A} \mathbf{A}^T$ sajátvektorai lesznek (gondoljuk meg, hogy minden szimmetrikus valós mátrix ortogonálisan spektrálfelbontható). Mindkét esetben $\mathbf{S}^T \mathbf{S}$ négyzetes mátrix diagonális elemei a szinguláris értékek négyzetei lesznek, azaz kimondható, hogy \mathbf{A} szinguláris értékei semmi mások, mint $\mathbf{A}^T \mathbf{A}$ sajátértékeinek négyzetgyökei. Innen persze az is következik, hogy ha $\mathbf{A}^T \mathbf{A}$ szinguláris, azaz van 0 sajátértéke, akkor biztosan lesz 0 szinguláris értéke \mathbf{A} -nak. Innen következik, hogy ha még mindig az $n > m$ esetről maradva \mathbf{A} oszlopangja kisebb, mint oszlopainak száma (lineárisan összefüggő oszlopai vannak), akkor $\mathbf{A}^T \mathbf{A}$ -nak lesz 0 sajátértéke, tehát nem lesz invertálható.

Az SVD segítségével kifejezhető \mathbf{A} Moore-Penrose pseudoinverze is (az inverz a transzpozícióhoz hasonlóan megfordítja a mátrixszorzás sorrendjét):

$$\mathbf{A}^\dagger = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T = (\mathbf{V} \mathbf{S}^T \mathbf{U}^T \mathbf{U} \mathbf{S} \mathbf{V}^T)^{-1} \mathbf{V} \mathbf{S}^T \mathbf{U}^T = \mathbf{V} \mathbf{S}^{-1} \mathbf{S}^{T^{-1}} \mathbf{V}^T \mathbf{V} \mathbf{S}^T \mathbf{U}^T$$

$$\mathbf{A}^\dagger = \mathbf{V}^T \mathbf{S}^\dagger \mathbf{U}^T$$

Mivel \mathbf{S} maga sem négyzetes mátrix feltétlenül, így \mathbf{S}^{-1} és $\mathbf{S}^{T^{-1}}$ valójában \mathbf{S}^\dagger illetve \mathbf{S}^{T^\dagger} Moore-Penrose pseudoinverzeket jelent. A pseudoinverz tulajdonságai hasonlóak az egyszerű inverzéhez. Innen is látszik, hogy \mathbf{A}^\dagger csak akkor létezik, ha \mathbf{S}^\dagger létezik, ami persze \mathbf{S} kvázi-diagonalitásából következően akkor igaz, ha \mathbf{S} oszlopai között nincs csupa 0-ákból álló, azaz nincs $\sigma = 0$ szinguláris értéke \mathbf{A} -nak.

.3 Valószínűségi vektorváltozók

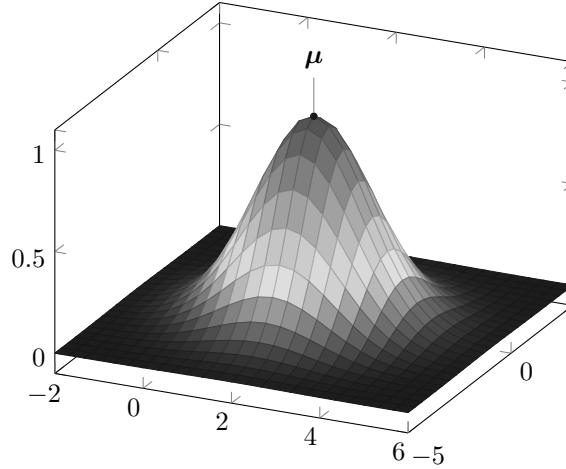
Egy $\boldsymbol{\xi} = [\xi_1, \dots, \xi_n]^T$ vektort *valószínűségi vektorváltozónak* hívunk, ha $\forall i$ -re ξ_i skalárértékű valószínűségi változó. A továbbiakban csak a vektorértékű normális eloszlást követő valószínűségi vektorváltozókkal foglalkozunk, ezek formálisan felírva:

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

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

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

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



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

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

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

$\boldsymbol{\Sigma}$ 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:

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

ahol $\sigma_1^2, \dots, \sigma_n^2$ rendre ξ_1, \dots, ξ_n varianciái.

.4 Mátrixdifferenciálás nagyon röviden

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

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

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

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

Ha \mathbf{A} szimmetrikus, akkor ezen felül

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

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

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

.5 *Pont és eloszlás Mahalanobis távolsága

Ez a rész csak érdekességként szerepel a PDF-ben, a Generalized Least Squares paraméterbecslés analitikus levezetésének bemutatásában használjuk csak, akinek nincs ideje átolvasni ezt a részt, nyugodtan ugorja át.

Legyen F egy \mathbb{R}^n -en értelmezett eloszlás $\boldsymbol{\mu} = [\mu_1, \mu_2, \dots, \mu_n]^T$ várható értékekkel és egy pozitív definit $\boldsymbol{\Sigma}$ variancia-kovariancia mátrixsal. Egy $\boldsymbol{x} = [x_1, x_2, \dots, x_n]^T$ pont Mahalanobis távolsága F -től

$$d_M(\boldsymbol{x}, F) := \sqrt{(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})}$$

Kettő $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ pont F szerinti Mahalanobis távolsága:

$$d_M(\boldsymbol{x}, \boldsymbol{y}; F) := \sqrt{(\boldsymbol{x} - \boldsymbol{y})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{y})}$$

A lineáris regresszió

.1 A regressziós modell

A regresszió kiindulópontja egy \mathcal{X} sokaság, melynek minden tagja rendelkezik \mathbf{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-k*, azaz *függetlenek és azonos eloszlásúak*, ami persze azt jelenti, hogy *minden magyarázó változó-vektor egy vektorértékű valószínűségi vektorváltozó*. Létezik egy másik konstrukció is, miszerint \mathbf{X} rögzített, és nem változik mintavételről mintavételre, ez azonban csak annyit jelent, hogy mindenhol, ahol feltételes eloszlás/várható érték van, onnan az \mathbf{X} feltételt ki kell venni. Mi \mathbf{X} -re mint valószínűségi vektorváltozók mátrixa tekintünk mostantól.

A megfigyelt magyarázó változó-vektorokat soronként egymásra rakva felépítünk egy úgynevezett *design mátrixot*, melyet mostantól \mathbf{X} -el jelölünk. Minden \mathbf{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 \mathbf{X} alakja a következő:

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

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

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

A lineáris regresszió kiindulópontja mindig egy *modell*, avagy egy elméleti feltevés arról, hogy milyen kapcsolatban áll a magyarázott \mathbf{y} változó a magyarázó \mathbf{X} regresszorokkal.

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

A lineáris kapcsolatot a $\boldsymbol{\beta}$ együtthatóvektor (avagy *paramétervektor*) írja le, míg $\boldsymbol{\epsilon}$ a regresszorok által nem magyarázott eltéréseket, avagy *hibákat* jelenti. Mostantól $\boldsymbol{\epsilon}$ -re *hibavektor* néven hivatkozunk.

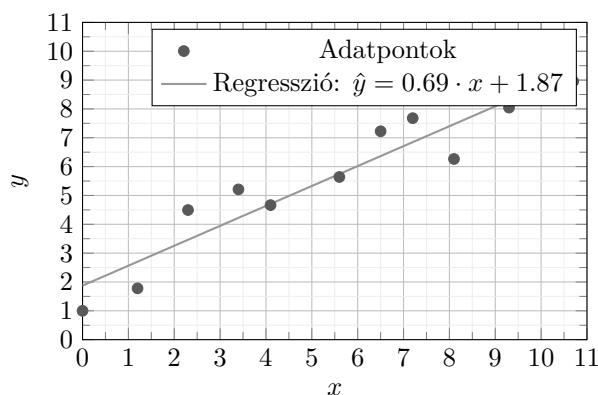
A regresszió célja, hogy megtaláljuk azt a $\hat{\boldsymbol{\beta}}$ *paraméterbecslés-vektort*, hogy az

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

úgynevezett *predikciós* egyenletből származott *becsült* $\hat{\mathbf{y}}$ vektor a lehető legközelebb legyen a valódi megfigyelt \mathbf{y} vektorhoz. Persze megfigyeletlen \mathbf{x} magyarázó változók esetén a predikciós egyenlet szintén működik,

és valójában ez is a célja a regressziónak.

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



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

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

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

.2 Az Ordinary Least Squares (OLS) becslési eljárás

A lineáris regresszió $\hat{\beta}$ -jának megtalálására az egyik lehetséges eljárás az Ordinary Least Squares, avagy legkisebb négyzetek módszere. Az eljárást kettő szemszögből is megvizsgáljuk.

.2.1 Az OLS-becslés geometriai értelmezése

Szinte mindig $n > p$, azaz több megfigyelésünk van, mint amennyi magyarázó változónk, így az

$$\mathbf{X}\beta = \mathbf{y}$$

egyenletrendszer *túlhatározott*, és nagyon specifikus esetektől eltekintve nem létezik egzakt megoldás β -ra. Az első fejezetben azonban láttuk, hogy a bal oldali pszeudoinverz pontosan ezt a problémát orvosolja. A jelölési konvenció a megoldásból nyert *paraméter-becslésre* $\hat{\beta}$, ami a mintavétel véletlenszerűségéből adódóan maga is vektorértékű valószínűségi változó ($\hat{\beta}$ pontos eloszlásáról a későbbiekben lesz szó):

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

Ebben az esetben \mathbf{y} -t az \mathbf{X} design mátrix oszlopterére vetítettük. $\mathbf{X}^T \mathbf{X}$ *Gram-mátrix* néven is ismeretes (egyébként $\mathbf{X} \mathbf{X}^T$ -ra is szoktak utalni ezen a néven, annyi különbséggel, hogy az előbbi a regresszorok közti korreláció mértékét mutatja a mintavételeken keresztülfutva, egyfajta *temporális* módon, az utóbbi pedig magukon a regresszorokon keresztülfutva egyfajta *térbeli* korrelációt mutat). Az $\mathbf{X}^T \mathbf{X}$ mátrix determinánsát *Gram-determinánsnak* is hívják.

Ha $n < p$, azaz kevesebb megfigyelésünk van, mint amennyi magyarázó változónk, az egyenletrendszer alulhatározott lesz, és nem fog létezni bal oldali pszeudoinverz, így nem lesz olyan \mathbf{X}^\dagger mátrix, amivel balról beszorozva \mathbf{X} -et az identitásmátrixot kapnánk. Ha \mathbf{X}^\dagger -el próbálkozunk, ami létezik:

$$\mathbf{X}^\dagger \mathbf{X} \beta = \mathbf{X}^\dagger \mathbf{y}$$

a bal oldalon \mathbf{X} sorterére való vetítési mátrixot kapnánk. Innen az is következik, hogy amint megtaláltuk $\hat{\beta}$ első n elemét, a maradék $p - n$ együttható az első n együttható lineáris kombinációjaként állna elő szükségszerűen. Ezért mostantól feltesszük, hogy a "normális" $n > p$ eset áll fenn.

A továbbiakban a *tényleges hibavektor* jelölése \mathbf{e} , a valós y_i -k és a $\mathbf{X}\hat{\beta} = \hat{\mathbf{y}}$ modellbecslés által prediktált \hat{y}_i -k közti eltérések vektora (sokszor \mathbf{e} -t $\hat{\mathbf{e}}$ -ként is jelölik):

$$\mathbf{e} = \begin{bmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \vdots \\ y_n - \hat{y}_n \end{bmatrix}$$

.2.2 Az OLS-becslés mint szélsőérték-feladat

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

$$\mathbf{e}^T \mathbf{e} \rightarrow \min_{\hat{\beta}}$$

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

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

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

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

Ebből megkapjuk az úgynevezett *normálegyenletet*:

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

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

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

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

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

.2.3 Az OLS-becslés tulajdonságai

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

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

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

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

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

$$\mathbf{X}^T \mathbf{e} = \mathbf{0}$$

valóban. Ez azt jelenti, hogy *minden magyarázó változó (regresszor) korrelálatlan a hibával*, pontosabban megfogalmazva *a regresszorok és a hibák mintakorrelációja zérus*. Mivel \mathbf{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:

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

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

Egy másik, ugyancsak fontos tulajdonság a predikciós formulából következik:

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

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*:

$$\bar{\mathbf{y}} = \bar{\hat{\mathbf{y}}}$$

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

.3 A Gauss-Markov feltételezések

A Gauss-Markov feltételezések biztosítják, hogy a *Gauss-Markov tétel* értelmében 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 rendelkezik, mint $\hat{\boldsymbol{\beta}}$ (az utóbbi tulajdonságra mint $\hat{\boldsymbol{\beta}}$ *hatásossága* szoktak hivatkozni).

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

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

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

A harmadik feltétel szerint

$$\mathbb{E}[\boldsymbol{\epsilon} \mid \mathbf{X}] = \mathbf{0}$$

$$\mathbb{E} \begin{bmatrix} \epsilon_1 \mid \mathbf{X} \\ \epsilon_2 \mid \mathbf{X} \\ \vdots \\ \epsilon_n \mid \mathbf{X} \end{bmatrix} = \mathbf{0}$$

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

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

A negyedik feltétel a hibák kovariancia mátrixára vonatkozik, mégpedig

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

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

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

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

Az utolsó feltétel szerint a hibatagok normális eloszlást követnek:

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

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

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

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

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

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

mindkét oldalon véve a várható értéket:

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

Mivel a Gauss-Markov feltételekből következően $\mathbb{E}[\mathbf{X}^T \boldsymbol{\epsilon}] = \mathbf{0}$, így

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

ezzel készen is vagyunk. A $\mathbb{E}[\mathbf{X}^T \boldsymbol{\epsilon}] = \mathbf{0}$ tulajdonságot *exogenitásnak* is hívjuk. Ez persze semmi mást nem jelent, mint hogy a regresszorok korrelálatlanok a hibával.

$$\text{Cov}[\mathbf{X}, \boldsymbol{\epsilon}] = \mathbb{E}[\mathbf{X}^T \boldsymbol{\epsilon}] - \mathbb{E}[\mathbf{X}]\mathbb{E}[\boldsymbol{\epsilon}] = \mathbb{E}[\mathbf{X}^T \boldsymbol{\epsilon}] = \mathbf{0}$$

.3.1 $\hat{\boldsymbol{\beta}}$ varianciája

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

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

Láttuk, hogy

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

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

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

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

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

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

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

A mátrixszorzás asszociativitásából pedig a

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

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

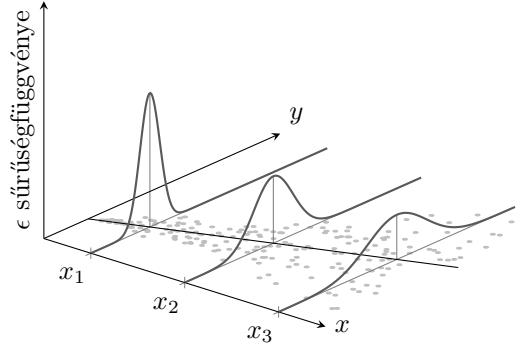
$$\mathbb{E}[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T \mid \mathbf{X}] = \begin{bmatrix} \text{Var}[\hat{\beta}_1] & \text{Cov}[\hat{\beta}_1, \hat{\beta}_2] & \dots & \text{Cov}[\hat{\beta}_1, \hat{\beta}_p] \\ \text{Cov}[\hat{\beta}_2, \hat{\beta}_1] & \text{Var}[\hat{\beta}_2] & \dots & \text{Cov}[\hat{\beta}_2,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

Ahogy $n \rightarrow \infty$, $\hat{\beta}$ eloszlása *aszimptotikusan normális lesz*, azaz

$$\hat{\beta} \mid \mathbf{X} \sim \mathcal{N}(\beta, \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1})$$

Erről a következő alfejezetben részletesebben szó 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:



.3.2 $\hat{\beta}$ eloszlása

Láttuk, hogy

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

Mivel $\hat{\beta}$ lineáris transzformációja \mathbf{y} -nek, így a normális eloszlású valószínűségi változók transzformációs tulajdonságából adódóan és $\Sigma = \sigma^2 \mathbf{I}$ -t kihasználva

$$\hat{\beta} \mid \mathbf{X} \underset{n \rightarrow \infty}{\sim} \mathcal{N}((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} \beta, \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1}) \sim \mathcal{N}(\beta, \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1})$$

azaz $\hat{\beta}$ valóban normális eloszlást követ, a már jól ismert $\sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$ varianciával és a valódi β várható értékkel. Persze ez az $\hat{\beta}$ vektorra vonatkozott, és csak *aszimptotikusan igaz*, azaz ha $n \rightarrow \infty$, hiszen ekkor a

centrális határeloszlás tétel értelmében a regresszorokból összetett mátrixkifejezés maga is normális lesz. Mi nyilván nem tudunk végtelen sok mintavétellel dolgozni, tehát azt mondjuk, hogy ha *elég nagy* n , akkor a paraméterbecslés nagyon jól megközelíti a normális eloszlást.

$\hat{\beta}$ feltétel nélküli varianciáját az alábbi módon kaphatjuk meg:

$$Var[\hat{\beta}] = \mathbb{E}[Var[\hat{\beta} \mid \mathbf{X}]] + Var[\mathbb{E}[\hat{\beta} \mid \mathbf{X}]]$$

ebből persze $\hat{\beta}$ feltétel nélküli eloszlása is számolható lesz, de erre nem térünk ki.

.3.3 Multikollinearitás

Ugyan feltettük, hogy nem létezik tökéletes multikollinearitás, de attól függetlenül valamilyen szintű multikollinearitás mindig elképzelhető a regresszorok között. Intuitíven a multikollinearitás egyfajta kapcsolatot vagy *hasonlóságot*, *korrelációt* jelent a regresszorok között.

Minél nagyobb a multikollinearitás mértéke, annál kevésbé különböznek \mathbf{X} oszlopai egymástól, azaz \mathbf{X} determinánsa annál kisebb. Emiatt $\mathbf{X}^T \mathbf{X}$ determinánsa is kisebb lesz, és mivel tetszőleges négyzetes mátrix esetén

$$\det(\mathbf{A}^{-1}) = \det(\mathbf{A})^{-1},$$

ezért a paraméterbecslés varianciája képletében $\det((\mathbf{X}^T \mathbf{X})^{-1})$ nagy lesz. Ugyan ez nem egzakt matematikai összefüggés, de intuitíven el lehet képzelni, hogy ez "agresszívebb" $\hat{\beta}$ -varianciákat eredményez. Egy másik fontos következmény inkább technikai jellegű, mégpedig hogy a numerikus algoritmus, ami kiszámolja $\mathbf{X}^T \mathbf{X}$ inverzét, jelentős multikollinearitás mellett pontatlan eredményt fog adni.

Ugyan a multikollinearitás nem sérti meg a Gauss-Markov feltételeket, azaz még mindig *BLUE* becslés lenne az *OLS*-el kapott $\hat{\beta}$, nem is a legideálisabb a lehetségesen inflálódott paraméterbecslés-variancia és a numerikus számítások nehézsége miatt. Erre jelenthet megoldást az úgynevezett *Ridge Regression* és *Lasso Regression*, avagy rendre *L2* és *L1 regularizációs regresszió*, akit érdekel utánaolvashat, de erre nem térünk ki bővebben.

.3.4 A hibavariancia becslése

A Gauss-Markov feltevések között szerepelt, hogy a hibatagok regresszorok szerinti feltételes eloszlása normális, egy bizonyos Σ variancia-kovariancia mátrixsal. Azt is feltettük, hogy Σ alakja

$$\Sigma = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix}$$

azaz a mátrix diagonális, és minden diagonálisbeli elem azonosan σ^2 . Felmerül persze a kérdés: Honnan tudjuk, hogy mi ez a σ^2 variancia? Ennek megoldásához *torzítatlan becslést kell adnunk σ^2 -ra a regresszióból*.

σ^2 torzítatlan becslése:

$$\widehat{\sigma^2} = \frac{\mathbf{e}^T \mathbf{e}}{n - p} = \frac{1}{n - p} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

ahol n a megfigyelések száma, p pedig a magyarázó változók száma (az interceptet is beleértve). Mivel $p - 1$ *valódi magyarázó változónk* van (azaz ami nem konstans, azaz nem β_0), így a hibavariancia-becslés nevezőjében - a valódi $(\beta_1 \dots \beta_{p-1})$ $p - 1$ darab magyarázó változóval - $n - (p - 1) - 1$ áll.

.4 A $p = 2$ -es egyváltozós regresszió

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 \mathbf{X} design mátrixunk most

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

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

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

A 2×2 -es mátrixok invertálása könnyen megy:

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

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* $\text{empcov}[x, y] = \frac{1}{n} \sum_i x_i y_i - (\frac{1}{n} \sum_i x_i)(\frac{1}{n} \sum_i y_i)$. x varianciája $\mathbb{E}[x^2] - \mathbb{E}[x]^2$ -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 számlálójában 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

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

Azt kaptuk tehát, hogy a legegyszerűbb egyváltozós regresszió becsült paraméterei

$$\begin{aligned} \hat{\beta}_0 &= \frac{\overline{x^2} \cdot \overline{y} - \overline{x} \cdot \overline{xy}}{\widehat{Var}[x]} \\ \hat{\beta}_1 &= \frac{\widehat{Cov}[x, y]}{\widehat{Var}[x]} \end{aligned}$$

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

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

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

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):

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

Ez nem csak intuitívan értelmezhető ("Az átlagos y semmi más, mint az y -tengellyel való metszéspont és $\hat{\beta}_1 \bar{x}$ 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:

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

valóban.

4.1 $\hat{\beta}$ varianciája egyváltozós regresszió esetén

Láttuk, hogy a paraméterbecslés varianciája az általános esetben

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

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

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

Használjuk ki az empirikus variancia képletét:

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

Innen könnyen látszik, hogy

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

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

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

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

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

A 2×2 -es mátrixszorzást elvégezve tényleg azt kaptuk, hogy az egyváltozós regresszió esetén S_{xx} semmi más, mint az $\mathbf{X}^T \mathbf{X}$ 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):

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

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 becült \hat{y} -ok és a tényleges y -ok közel vannak egymáshoz, azaz a regresszióval nagyon jól becsülhetjük a valódi y értékeket.

.5 Az R^2 mutató

Tekintsük az egyváltozós regressziós modellt. Legyen

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

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:

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

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

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

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

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

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.)

.6 Kihagyott változó bias - Omitted Variable Bias

Tekintsünk egy

$$y = \beta_0 + \beta_1 x + \beta_2 z + \epsilon$$

lineáris modellt. Ahhoz, hogy létezzen kihagyott változó bias, a kihagyott változó együtthatója nem lehet zérus, illetve a kihagyott változónak *korrelálnia kell* egy másik, regresszióban szereplő magyarázó változóval.

Tegyük fel, hogy kihagyjuk z -t a regresszióból:

$$y = \tilde{\beta}_0 + \tilde{\beta}_1 x + \tilde{\epsilon}$$

és hogy z -t x a következőképpen magyarázza:

$$z = \delta_0 + \delta_1 x + \nu$$

Helyettesítsük be a második egyenletet az eredeti teljes egyenletbe:

$$y = (\beta_0 + \beta_2 \delta_0) + (\beta_1 + \beta_2 \delta_1)x + (\epsilon + \beta_2 \nu)$$

Látható, hogy ha ezen a kihagyott változós modellen végeznénk el a regressziós paraméterbecslést, x együtthatójának nem β_1 -et, hanem $\beta_1 + \beta_2 \delta_1$ -et kapnánk, ami nyilvánvalóan az eredeti modellel konzisztensen *torzított*. Úgy is gondolhatunk erre, hogy a kihagyott z miatt x becsült együtthatója tartalmazni fogja az indirekt hatást is (z -n x hatása δ_1 , ezt megszorozva még y -n z hatásával).

Mátrixformában az Omitted Variable Bias az alábbi formában szemléltethető. Legyenek

$$\mathbf{X} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{Z} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix}$$

a regresszorokat tartalmazó vektorok. A z -t kihagyó modell design mátrixa pusztán \mathbf{X} , így az ebből nyert paraméterbecslés

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

Írjuk be \mathbf{y} helyére a tényleges, teljes modellből származó alakot:

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

Mindkét oldalon várható értéket véve, és visszaemlékezve arra, hogy az utolsó tag zérus lesz:

$$\mathbb{E}[\hat{\beta}] = \beta + (\mathbf{X}^T \mathbf{X})^{-1} \mathbb{E}[\mathbf{X}^T \mathbf{Z}] \delta$$

ahol látható, hogy a jobb oldal második tagja pontosan a kihagyott z változó miatti torzítás, avagy *bias*.

.7 MSE és a bias-variancia tradeoff

Tekintsünk egy általános

$$\mathbf{y} = \mathbf{p}(\mathbf{X}) + \epsilon$$

modellt. Csakúgy, mint eddig, \mathbf{X} a regresszorok, \mathbf{y} a magyarázott változó vektora, \mathbf{p} pedig valamilyen függvény. Az ϵ hibák regresszorok szerinti feltételes várható értéke 0. Figyeljük meg, hogy a lineáris regresszió esetében \mathbf{p} a lineáris β együtthatóvektor. A célunk, hogy megtaláljuk azt a $\hat{\mathbf{p}}$ függvényt, amivel a becsült

$$\hat{\mathbf{y}} = \hat{\mathbf{p}}(\mathbf{X})$$

$\hat{\mathbf{y}}$ -ok és a tényleges \mathbf{y} -ok négyzetes távolsága a lehető legkisebb. Nyilvánvalóan - ezt a regresszióanalízis is láttuk már - egy olyan $\hat{\mathbf{p}}$ -t találni, ami *tökéletesen* becsüli \mathbf{y} -t reális esetben lehetetlen, így fontos lesz, hogy valahogyan számszerűsíthessük a megfigyeléseken (mintán) alapuló illetve a még megfigyeletlen regresszorokon vett várható tévedésünket.

A *Mean Squared Error*, röviden *MSE* klasszikusan az átlagos avagy várható négyzetes eltérések összegét jelenti a prediktált $\hat{\mathbf{y}}$ és a tényleges \mathbf{y} -ok között. Attól függően azonban, hogy mit akarunk vele pontosan kifejezni, definiálhatjuk a *prediktorok* (a fenti "klasszikus" eltérés-négyzetösszeges definíció) és a *becslések* szemszögéből is.

A *prediktorok* szemszögéből a definíció egy n elemű mintán

$$MSE := \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Kompaktabban kifejezve a tényleges \mathbf{e} hibákkal:

$$MSE := \frac{1}{n} \mathbf{e}^T \mathbf{e}$$

Ha \mathbf{p} becsléséhez nem használtuk fel az összes n elemet, hanem csak $m < n$ -et, akkor az *MSE* definiálható úgy is, mint az átlagos négyzetes hiba a becsléshez fel nem használt adatpontokon:

$$MSE := \frac{1}{n-m} \sum_{i=m+1}^n (y_i - \hat{y}_i)^2$$

A *becslés* szemszögéből az *MSE* a $\hat{\mathbf{p}}$ becslésünkre vonatkozik, mégpedig egy teoretikus valódi \mathbf{p} függvény mellett

$$MSE(\hat{\mathbf{p}}) = \mathbb{E}_{\mathbf{p}}[(\hat{\mathbf{p}} - \mathbf{p})^2]$$

Ez semmi más, mint a *második momentuma* a $\hat{\mathbf{p}} - \mathbf{p}$ becslés-eltérésnek. Ebből a definícióból következik a *bias-variancia tradeoff*, melynek fontos következményei lesznek. Lássuk ezt be!

Tudjuk, hogy tetszőleges ξ valószínűségi változóra $\mathbb{E}[\xi^2] = Var[\xi] + \mathbb{E}^2[\xi]$. Most $\xi = \hat{\mathbf{p}} - \mathbf{p}$ -vel:

$$MSE = \mathbb{E}[(\hat{\mathbf{p}} - \mathbf{p})^2] = Var[\hat{\mathbf{p}} - \mathbf{p}] + \mathbb{E}^2[\hat{\mathbf{p}} - \mathbf{p}]$$

A $\mathbb{E}[\hat{\mathbf{p}} - \mathbf{p}]$ várható eltérést (\mathbf{p} -hez képest) hívjuk *bias*-nak, avagy *torzításnak*, ennek négyzetére $Bias^2[\mathbf{p}]$ -ként hivatkozunk mostantól. Mivel \mathbf{p} a modell szerint egy konkrét függvény, így $Var[\hat{\mathbf{p}} - \mathbf{p}] = Var[\hat{\mathbf{p}}]$, és ezzel

$$MSE = \mathbb{E}[(\hat{\mathbf{p}} - \mathbf{p})^2] = Var[\hat{\mathbf{p}}] + Bias^2[\hat{\mathbf{p}}].$$

A becslés szemszögéből tehát az *MSE* semmi más, mint a becslés varianciájának és torzítás-négyzetének összege. Ezt az összefüggést hívjuk *bias-variancia tradeoff*-nak, hiszen adott *MSE* mellett ha az egyiket csökkenteni is tudom, a másik nőni fog. Komplex $\hat{\mathbf{p}}$ becslés mellett a bias, avagy torzítottság alacsony lesz, azonban magas varianciája, avagy *bizonytalansága* lesz a becslésemnek. Egyszerű $\hat{\mathbf{p}}$ mellett pedig a bias lesz magas, alacsony varianciával.

.8 A heteroszkedaszticitás kezelése, a GLS eljárás

A Gauss-Markov feltevések egyike volt, hogy Σ hiba variancia-kovariancia mátrix diagonális, és a diagonális elemek azonosan σ^2 -ek. Láttuk azt is, hogy ezekre a σ^2 -ekre torzítatlan becslést ad a $\hat{\sigma}^2$ *hibavariancia becslés*. Azt az esetet, amikor Σ diagonális, azonban σ^2 -ek nem egyenlőek, heteroszkedaszticitásnak hívjuk, és emellett a hiba variancia-kovariancia mátrix mellett a paraméterbecslés varianciája már nem a megszokott

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

alakú, hiszen nem emelhetjük ki $\sigma^2 \mathbf{I}$ -t középről.

Tudjuk, hogy minden variancia-kovariancia mátrix szimmetrikus és pozitív szemidefinit. Ezért $\exists \mathbf{P} : \mathbf{P}\mathbf{P}^T = \mathbf{\Sigma}$ (ez a *Cholesky-felbontás*), tehát felbonthatjuk a kovariancia mátrixot kettő, $\mathbf{\Sigma}$ -val azonos dimenziójú invertálható mátrix szorzatára (Ez analóg azzal, hogy \mathbb{R} -en minden pozitív szemidefinit (nemnegatív) valós számnak létezik négyzetgyöke, és a négyzetgyök csak akkor 0, ha maga a szám 0, de most a zérus kovariancia mátrix esetétől eltekintünk).

A célunk az, hogy $\mathbf{\Sigma}$ kovariancia mátrixot $\sigma^2 \mathbf{I}$ alakúra hozzuk. Ha megszorozzuk balról \mathbf{P}^{-1} -el a ϵ hibát, a hiba varianciája:

$$\text{Var}[\mathbf{P}^{-1}\epsilon] = \mathbf{P}^{-1}\mathbf{\Sigma}\mathbf{P}^{-1T}$$

A felbontásból következően, és a $\mathbf{P}^{T-1} = \mathbf{P}^{-1T}$ összefüggést felhasználva:

$$\mathbf{P}^{-1}\mathbf{\Sigma} = \mathbf{P}^T \implies \mathbf{P}^{-1}\mathbf{\Sigma}\mathbf{P}^{-1T} = \mathbf{P}^T\mathbf{P}^{-1T} = \mathbf{I}$$

Azt kaptuk tehát, hogy ha a \mathbf{P}^{-1} -el beszorzott módosított regressziós modellegyenletet tekintjük

$$\mathbf{P}^{-1}\mathbf{X}\beta + \mathbf{P}^{-1}\epsilon = \mathbf{P}^{-1}\mathbf{y}$$

akkor ebben a modellben a hiba varianciamátrixa az identitás mátrix, így nem áll fenn heteroszkedaszticitás.

A módosított modellel való paraméterbecslés tehát

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

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

Szintén a felbontásból, most már mátrixhatványokkal kiírva adódik, hogy

$$\mathbf{P} = \mathbf{\Sigma}^{\frac{1}{2}}$$

Így

$$\mathbf{P}^{-1T}\mathbf{P}^{-1} = \mathbf{\Sigma}^{-\frac{1}{2}T}\mathbf{\Sigma}^{-\frac{1}{2}} = \mathbf{\Sigma}^{-1}$$

Ezzel a paraméterbecslés alakja:

$$\hat{\beta}_{GLS} = (\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{y}$$

Ez már konzisztens a Gauss-Markov feltételekkel, így $\hat{\beta}$ paraméterbecslés teljesíti a BLUE kritériumokat. Ez az eljárás egy speciális esete a *Generalized Least Squares (GLS)* becslési eljárásnak, ahol $\mathbf{\Sigma}$ nemdiagonális elemei mind 0-k, az angol irodalomban *Weighted Least Squares* néven szerepel. Ha $\mathbf{\Sigma}$ nem is diagonális (azaz autokorreláció van a hibák között), akkor is elvégezhető a GLS eljárás. $\mathbf{\Sigma}^{-1}$ -t, azaz az inverz variancia-kovariancia mátrixot *precíziós mátrixnak* is hívják.

.8.1 *A GLS becslés analitikus levezetése

$\hat{\beta}_{GLS}$ alakját megkaphatjuk úgy is, ha tekintjük az alábbi optimalizációs problémát:

$$(\mathbf{y} - \mathbf{X}\mathbf{b})^T\mathbf{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\mathbf{b}) \rightarrow \underset{\mathbf{b}}{\text{argmin}}$$

Ez persze semmi más, mint a Mahalanobis távolság minimalizálása \mathbf{y} és $\mathbf{X}\mathbf{b}$ között \mathbf{b} szerint. Kibontva a kifejezést és a \mathbf{b} szerinti deriváltat 0-ra állítva ($\mathbf{b} = \hat{\beta}_{GLS}$):

$$2\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X}\hat{\beta}_{GLS} - 2\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{y} = 0$$

Ebből

$$\hat{\beta}_{GLS} = (\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{y}$$

Így is megkaptuk ugyanazt az alakot.

.8.2 A Feasible Generalized Least Squares (FGLS) eljárás

Ha nem ismerjük a valódi Σ hibavariancia-kovariancia mátrixot, akkor a már bevezetett $\widehat{\sigma}^2$ becslt varianciákkal konstruálhatjuk meg a becslt $\widehat{\Sigma}$ mátrixot.

Az *FGLS* eljárás *kétlépcsős*, első lépésként először is elvégzünk a módosítatlan $\mathbf{X}\beta + \epsilon = \mathbf{y}$ modellel egy egyszerű *OLS* becslést, melyből $\hat{\beta}_{OLS}$ -t kapjuk (ez persze heteroszkedaszticitás esetén nem BLUE becslés). Az így kapott

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\beta}_{OLS}$$

hibavektorokkal megbecsüljük $\widehat{\Sigma}$ hibavariancia-kovariancia mátrixot. Persze mivel heteroszkedaszticitást feltételeztünk, így $\widehat{\sigma}_1^2, \dots, \widehat{\sigma}_n^2$ becslt hibavarianciákat csupán egyelemes mintával (rendre e_1, \dots, e_n tényleges hibákkal) becsülhetnénk, azaz a tényleges becslt varianciákhoz *valamilyen előzetes feltevés a heteroszkedaszticitással konzisztens hibavarianciákra*, de ezzel részletesebben nem foglalkozunk, és feltesszük, hogy "valahogy" meg tudjuk kapni ezen becsléseket. Így a variancia-kovariancia mátrix:

$$\widehat{\Sigma} = \begin{bmatrix} \widehat{\sigma}_1^2 & 0 & \dots & 0 \\ 0 & \widehat{\sigma}_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \widehat{\sigma}_n^2 \end{bmatrix}$$

Második lépésként az első lépésben kapott $\widehat{\Sigma}$ mátrixsal *GLS* becsléssel megkapjuk a

$$\hat{\beta}_{FGLS} = (\mathbf{X}^T \widehat{\Sigma}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \widehat{\Sigma}^{-1} \mathbf{y}$$

FGLS paraméterbecslést. Ez az eljárás *iterálható*, azaz vehetjük az *FGLS* becslésből kapott

$$\mathbf{e}_{FGLS} = \mathbf{y} - \mathbf{X}\hat{\beta}_{FGLS}$$

hibavektort, és újrabecsülhetjük $\widehat{\Sigma}$ -t:

$$\widehat{\Sigma}_{FGLS} = \begin{bmatrix} \widehat{\sigma}_{FGLS,1}^2 & 0 & \dots & 0 \\ 0 & \widehat{\sigma}_{FGLS,2}^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \widehat{\sigma}_{FGLS,n}^2 \end{bmatrix}$$

Ezzel az újrabecsült kovariancia-variancia mátrixsal az új paraméterbecslésünk

$$\hat{\beta}_{FGLS2} = (\mathbf{X}^T \widehat{\Sigma}_{FGLS}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \widehat{\Sigma}_{FGLS}^{-1} \mathbf{y}$$

Az iteráció tetszőlegesen sokáig folytatódhat, és minden iterációval egyre közelebb kerülünk a tényleges β -hoz.