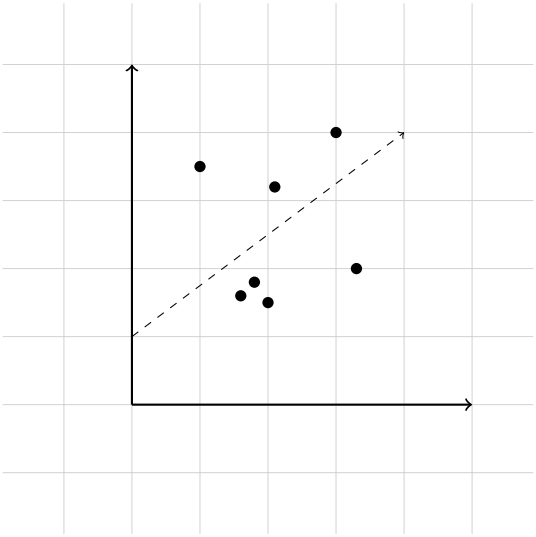
Lineáris regresszió elméleti összefoglaló

2023.03.19

Bognár Miklós



Tartalom

*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 mártixdifferenciálás-kifejezések.

.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 $n > m$ vagy $m > n$ 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. Bal oldali pszeudoinverz csakis $n \geq m$ esetben létezik, míg jobboldali az $n \leq m$ esetben.

.2 *Mátrixok szinguláris értékei, SVD

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

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

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

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

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

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

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

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

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

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

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

.3 Valószínűségi vektorváltozók

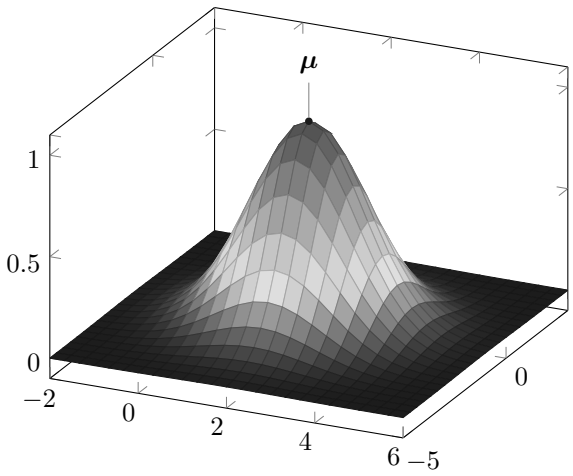
Egy $\boldsymbol{\xi} = [\xi_1, \dots, \xi_n]^T$ vektort *valószínűségi vektorváltozónak* hívunk, ha $\forall i$-re ξ_i skalárértékű valószínűségi változó. A továbbiakban csak a vektorértékű normális eloszlást követő valószínűségi vektorváltozókkal foglalkozunk, ezek formálisan felírva:

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

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

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

A sűrűségfüggvény $n = 2$ esetben $\boldsymbol{\mu} = [2, -1]^T$ és $\boldsymbol{\Sigma} = \boldsymbol{I}$ 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.

.4 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

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

.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 $\boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_n]^T$ várható értékekkel és egy pozitív definit $\boldsymbol{\Sigma}$ variancia-kovariancia mátrixxal. Egy $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]^T$ 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:

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

.6 *Particionált mátrixok és a Blockwise formula

Legyen $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ négyzetes invertálható mátrix. Particionáljuk $\boldsymbol{A}$-t az alábbi módon:

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix}$$

ahol $\boldsymbol{A}_{11} \in \mathbb{R}^{m_1 \times m_1}$, $\boldsymbol{A}_{22} \in \mathbb{R}^{m_2 \times m_2}$, $m_1 + m_2 = m$ maguk is invertálható mátrixok. Ekkor $\boldsymbol{A}^{-1}$ felírható:

$$\boldsymbol{A}^{-1} = \begin{bmatrix} (\boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21})^{-1} & -\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}(\boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12})^{-1} \\ -\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}(\boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21})^{-1} & (\boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12})^{-1} \end{bmatrix}$$

A négyzetes mátrixokról az általános $n \times m$-es mátrixokra áttérve legyen $\boldsymbol{X} \in \mathbb{R}^{n \times m}$ tetszőleges valós mátrix, $n > m$, és particionálja ezt horizontálisan $\boldsymbol{Q}$ és $\boldsymbol{R}$:

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{Q} & \boldsymbol{R} \end{bmatrix}$$

Ekkor $\boldsymbol{X}\boldsymbol{X}^\dagger$ oszloptér-vetítés mátrix a következőképpen írható fel:

$$\boldsymbol{X}\boldsymbol{X}^\dagger = \boldsymbol{Q}\boldsymbol{Q}^\dagger + ((\boldsymbol{I} - \boldsymbol{Q}\boldsymbol{Q}^\dagger)\boldsymbol{R})((\boldsymbol{I} - \boldsymbol{Q}\boldsymbol{Q}^\dagger)\boldsymbol{R})^\dagger$$

Ez az úgynevezett *Blockwise formula*, avagy a blokkonkénti vetítés formula. Még mindig ennél a partíciónál maradva $\boldsymbol{X}^\dagger$-t is megkaphatjuk a blokkokkal:

$$\boldsymbol{X}^\dagger = \begin{bmatrix} \boldsymbol{Q} & \boldsymbol{R} \end{bmatrix}^\dagger = \begin{bmatrix} \boldsymbol{P}_R\boldsymbol{Q}(\boldsymbol{Q}^T\boldsymbol{P}_R\boldsymbol{Q})^{-1} \\ \boldsymbol{P}_Q\boldsymbol{R}(\boldsymbol{R}^T\boldsymbol{P}_Q\boldsymbol{R})^{-1} \end{bmatrix} = \begin{bmatrix} (\boldsymbol{P}_R\boldsymbol{Q})^\dagger \\ (\boldsymbol{P}_Q\boldsymbol{R})^\dagger \end{bmatrix}$$

ahol $\boldsymbol{P}_Q$ és $\boldsymbol{P}_R$ rendre az ortogonális projekciómátrixok a $\boldsymbol{Q}$ és $\boldsymbol{R}$ mátrixok képterére ortogonális vektortérre:

$$\boldsymbol{P}_Q = \boldsymbol{I} - \boldsymbol{Q}(\boldsymbol{Q}^T\boldsymbol{Q})^{-1}\boldsymbol{Q}^T = \boldsymbol{I} - \boldsymbol{Q}\boldsymbol{Q}^\dagger$$

$$\boldsymbol{P}_R = \boldsymbol{I} - \boldsymbol{R}(\boldsymbol{R}^T\boldsymbol{R})^{-1}\boldsymbol{R}^T = \boldsymbol{I} - \boldsymbol{R}\boldsymbol{R}^\dagger$$

Tehát

$$\begin{bmatrix} \boldsymbol{Q} & \boldsymbol{R} \end{bmatrix}^\dagger = \begin{bmatrix} ((\boldsymbol{I} - \boldsymbol{R}\boldsymbol{R}^\dagger)\boldsymbol{Q})^\dagger \\ ((\boldsymbol{I} - \boldsymbol{Q}\boldsymbol{Q}^\dagger)\boldsymbol{R})^\dagger \end{bmatrix}$$

Ezek a formulák akkor lehetnek hasznosak, ha a lineáris regresszió *design mátrixát* particionáljuk bizonyos magyarázó változók szerint (például $\boldsymbol{Q}$ lehet a csupa 1-ekből álló intercept oszlopmátrix), de ez már nagyon túlmutat a tárgy anyagán.

A lineáris regresszió

.1 A regressziós modell

A regresszió kiindulópontja egy $\mathscr{X}$ 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-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 $\boldsymbol{X}$ rögzített, és nem változik mintavételről mintavételre, ez azonban csak annyit jelent, hogy mindenhol, ahol feltételes eloszlás/várható érték van, onnan az $\mid \boldsymbol{X}$ feltételt ki kell venni. Mi $\boldsymbol{X}$-re mint valószínűségi vektorváltozók mátrixa tekintünk mostantól.

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

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

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.

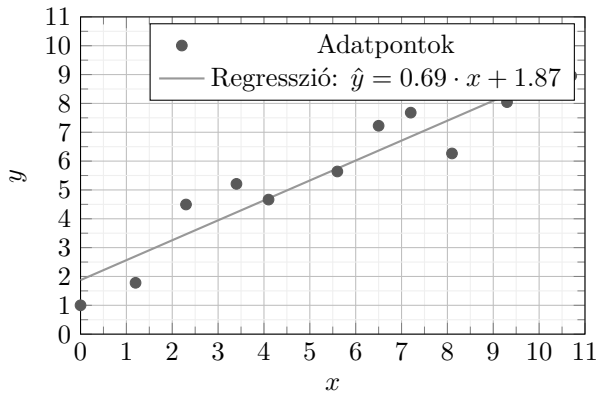
A regresszió célja, hogy megtaláljuk azt a $\hat{\boldsymbol{\beta}}$ *paraméterbecslés-vektort*, hogy az

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

úgynevezett *predikciós* egyenletből származott *becsült* $\hat{\boldsymbol{y}}$ vektor a lehető legközelebb legyen a valódi megfigyelt $\boldsymbol{y}$ vektorhoz. Persze megfigyeletlen $\boldsymbol{x}$ 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 $\hat{\boldsymbol{\beta}}$ 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).

.2 Az Ordinary Least Squares (OLS) becslési eljárás

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

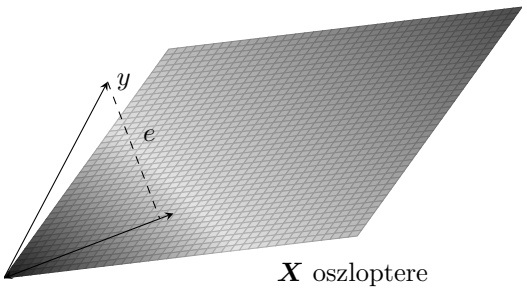
.2.1 Az OLS-becslés geometriai értelmezése

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

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

egyenletrendszer *túlhatározott*, és nagyon specifikus esetektől eltekintve nem létezik egzakt megoldás $\boldsymbol{\beta}$-ra. Az első fejezetben azonban láttuk, hogy a bal oldali pszeudoinverz pontosan ezt a problémát orvosolja. A jelölési konvenció a megoldásból nyert *paraméter-becslésre* $\hat{\boldsymbol{\beta}}$, ami a mintavétel véletlenszerűségéből adódóan maga is vektorértékű 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. $\boldsymbol{X}^T\boldsymbol{X}$ *Gram-mátrix* néven is ismeretes (egyébként $\boldsymbol{X}\boldsymbol{X}^T$-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 $\boldsymbol{X}^T\boldsymbol{X}$ 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 $\boldsymbol{X}^\dagger$ mátrix, amivel balról beszorozva $\boldsymbol{X}$-et az identitásmátrixot kapnánk. Ha $\boldsymbol{X}^\ddagger$-el próbálkozunk, ami létezik:

$$\boldsymbol{X}^\ddagger\boldsymbol{X}\beta = \boldsymbol{X}^\ddagger\boldsymbol{y}$$

a bal oldalon $\boldsymbol{X}$ sorterére való vetítési mátrixot kapnánk. Innen az is következik, hogy amint megtaláltuk $\hat{\boldsymbol{\beta}}$ 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 $\boldsymbol{e}$, 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):

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

.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.

.2.3 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:

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

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.

.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 rendelkezne, 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:

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

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

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

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

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

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

A negyedik feltétel a hibák kovariancia mátrixára vonatkozik, mégpedig

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

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

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

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 $\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T \mid \boldsymbol{X}]$ 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:

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

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

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

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

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

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

mindkét oldalon véve a várható értéket:

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

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

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

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

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

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

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

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

Láttuk, hogy

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

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

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

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

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

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

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

A mátrixszorzás asszociativitásából pedig a

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

végleges alakot kapjuk. Ugyanez megkapható az első fejezetben bemutatott $Var[\boldsymbol{A}\boldsymbol{\xi}] = \boldsymbol{A}Var[\boldsymbol{\xi}]\boldsymbol{A}^T$ 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 $Var[\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}]$ alakja

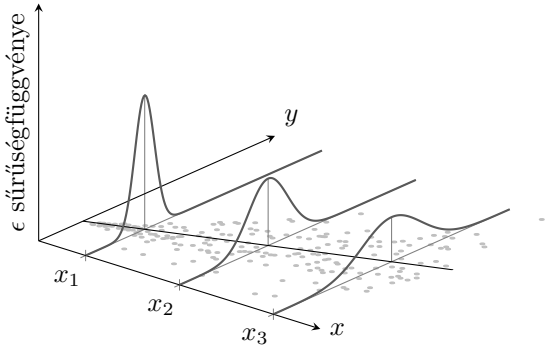
$$\mathbb{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \mid \boldsymbol{X}] = \begin{bmatrix} Var[\hat{\beta}_1] & Cov[\hat{\beta}_1, \hat{\beta}_2] & \dots & Cov[\hat{\beta}_1, \hat{\beta}_p] \\ Cov[\hat{\beta}_2, \hat{\beta}_1] & Var[\hat{\beta}_2]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

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

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

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 $\hat{\boldsymbol{\beta}}$ eloszlása

Láttuk, hogy

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

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

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

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

Gyakran $\hat{\boldsymbol{\beta}}$ aszimptotikus viselkedésére mint *konzisztencia* utalnak, ez annyit jelent, hogy az eloszlás aszimptotikusan "ráhúzódik" a $\boldsymbol{\beta}$ várható értékű normális eloszlásra. Formálisan felírva

$$\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$$

$$\mathbb{P}(|\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i| < \varepsilon) > 1 - \delta, \quad \forall i, \quad \varepsilon, \delta > 0$$

ahol $\xrightarrow{p}$ a *probability limit*. Fontos, hogy a konzisztencia *nem* vonja maga után a torzítatlanságot, hiszen gondoljuk meg, hogy ha $\frac{1}{n}$-et adunk egy torzítatlan konzisztens becsléshez, akkor a becslés továbbra is konzisztens marad, azonban már torzított lesz.

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

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

ebből persze $\hat{\boldsymbol{\beta}}$ 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 $\boldsymbol{X}$ oszlopai egymástól, azaz $\boldsymbol{X}$ determinánsa annál kisebb. Emiatt $\boldsymbol{X}^T\boldsymbol{X}$ determinánsa is kisebb lesz, és mivel tetszőleges négyzetes mátrix esetén

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

ezért a paraméterbecslés varianciája képletében $det((\boldsymbol{X}^T\boldsymbol{X})^{-1})$ nagy lesz. Ugyan ez nem egzakt matematikai összefüggés, de intuitíven el lehet képzelni, hogy ez "agresszívebb" $\hat{\boldsymbol{\beta}}$-varianciákat eredményez. Egy másik fontos következmény inkább technikai jellegű, mégpedig hogy a numerikus algoritmus, ami kiszámolja $\boldsymbol{X}^T\boldsymbol{X}$ 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 $\hat{\boldsymbol{\beta}}$, 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 $\mathbf{\Sigma}$ variancia-kovariancia mátrixxal. Azt is feltettük, hogy $\mathbf{\Sigma}$ alakja

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

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:

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

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 $\boldsymbol{X}$ 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.

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

Láttuk, hogy a paraméterbecslés varianciája az általános esetben

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

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.

.5 Az R^2 mutató

Tekintsük az egyváltozós regressziós modellt. 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.)

.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

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

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

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

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

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

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

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

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

$$\boldsymbol{y} = \mathfrak{P}(\boldsymbol{X}) + \boldsymbol{\epsilon}$$

modellt. Csakúgy, mint eddig, $\boldsymbol{X}$ a regresszorok, $\boldsymbol{y}$ a magyarázott változó vektora, $\mathfrak{P}$ pedig valamilyen függvény. Az $\boldsymbol{\epsilon}$ hibák regresszorok szerinti feltételes várható értéke 0. Figyeljük meg, hogy a lineáris regresszió esetében $\mathfrak{P}$ a lineáris $\boldsymbol{\beta}$ együtthatóvektor. A célunk, hogy megtaláljuk azt a $\widehat{\mathfrak{P}}$ függvényt, amivel a becsült

$$\hat{\boldsymbol{y}} = \widehat{\mathfrak{P}}(\boldsymbol{X})$$

$\hat{\boldsymbol{y}}$-ok és a tényleges $\boldsymbol{y}$-ok négyzetes távolsága a lehető legkisebb. Nyilvánvalóan - ezt a regressziónál is láttuk már - egy olyan $\widehat{\mathfrak{P}}$-t találni, ami *tökéletesen* becsüli $\boldsymbol{y}$-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 $\hat{\boldsymbol{y}}$ és a tényleges $\boldsymbol{y}$-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 $\boldsymbol{e}$ hibákkal:

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

Ha $\mathfrak{P}$ 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 $\widehat{\mathfrak{P}}$ becslésünkre vonatkozik, mégpedig egy teoretikus valódi $\mathfrak{P}$ függvény mellett

$$MSE(\widehat{\mathfrak{P}}) = \mathbb{E}_{\mathfrak{P}}[(\widehat{\mathfrak{P}} - \mathfrak{P})^2]$$

Ez semmi más, mint a *második momentuma* a $\widehat{\mathfrak{P}} - \mathfrak{P}$ 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 $\xi = \widehat{\mathfrak{P}} - \mathfrak{P}$-vel:

$$MSE = \mathbb{E}[(\widehat{\mathfrak{P}} - \mathfrak{P})^2] = Var[\widehat{\mathfrak{P}} - \mathfrak{P}] + \mathbb{E}^2[\widehat{\mathfrak{P}} - \mathfrak{P}]$$

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

$$MSE = \mathbb{E}[(\widehat{\mathfrak{P}} - \mathfrak{P})^2] = Var[\widehat{\mathfrak{P}}] + Bias^2[\widehat{\mathfrak{P}}].$$

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

.8 A heteroszkedaszticitás kezelése, a GLS eljárás

A Gauss-Markov feltevések egyike volt, hogy $\boldsymbol{\Sigma}$ 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 $\widehat{\sigma^2}$ *hibavariancia becslés*. Azt az esetet, amikor $\boldsymbol{\Sigma}$ diagonális, azonban σ^2-ek nem egyenlőek, heteroszkedaszticitásnak hívjuk, és emellett a hiba variancia-kovariancia mátrix mellett a paraméterbecslés varianciája már nem a megszokott

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

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

Tudjuk, hogy minden variancia-kovariancia mátrix szimmetrikus és pozitív szemidefinit. Ezért $\exists \boldsymbol{P} : \boldsymbol{P}\boldsymbol{P}^T = \boldsymbol{\Sigma}$ (ez a *Cholesky-felbontás*), tehát felbonthatjuk a kovariancia mátrixot kettő, $\boldsymbol{\Sigma}$-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 $\boldsymbol{\Sigma}$ kovariancia mátrixot $\sigma^2\boldsymbol{I}$ alakúra hozzuk. Ha megszorozzuk balról $\boldsymbol{P}^{-1}$-el a $\boldsymbol{\epsilon}$ hibatagot, a hiba varianciája:

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

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

$$\boldsymbol{P}^{-1}\boldsymbol{\Sigma} = \boldsymbol{P}^T \Longrightarrow \boldsymbol{P}^{-1}\boldsymbol{\Sigma}{\boldsymbol{P}^{-1}}^T = \boldsymbol{P}^T{\boldsymbol{P}^{-1}}^T = \boldsymbol{I}$$

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

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

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

A módosított modellel való paraméterbecslés tehát

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

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

Szintén a felbontásból, most már mátrixhatványokkal kiírva adódik, hogy

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

Így

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

Ezzel a paraméterbecslés alakja:

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

Ez már konzisztens a Gauss-Markov feltételekkel, így $\hat{\boldsymbol{\beta}}$ paraméterbecslés teljesíti a BLUE kritériumokat. Ez az eljárás egy speciális esete a *Generalized Least Squares (GLS)* becslési eljárásnak, ahol $\boldsymbol{\Sigma}$ nemdiagonális elemei mind 0-k, az angol irodalomban *Weighted Least Squares* néven szerepel. $\boldsymbol{\Sigma}^{-1}$-t, azaz az inverz variancia-kovariancia mátrixot *precíziós mátrixnak* is hívják. A *GLS* működik autokorreláció esetén is, azaz tetszőleges pozitív definit $\boldsymbol{\Sigma}$ hibavariancia mátrixxal is, és igazából ez az amit "hivatalosan" *GLS*-nek hívnak.

.8.1 *A GLS becslés analitikus levezetése

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

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

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

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

Ebből

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

Így is megkaptuk ugyanazt az alakot.

.8.2 A Feasible Generalized Least Squares (FGLS) eljárás

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

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

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

hibavektorokkal megbecsüljük $\widehat{\boldsymbol{\Sigma}}$ hibavariancia-kovariancia mátrixot. Persze mivel heteroszkedaszticitást feltételeztünk, így $\widehat{\sigma_1^2},\cdots\widehat{\sigma_n^2}$ becsült hibavarianciákat csupán egyelemes mintával (rendre $e_1,\ldots,e_n$ tényleges hibákkal) becsülhetnénk, azaz a tényleges becsült 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:

$$\widehat{\boldsymbol{\Sigma}}=\begin{bmatrix}\widehat{\sigma_1^2} & 0 & \ldots & 0\\ 0 & \widehat{\sigma_2^2} & \ldots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \ldots & \widehat{\sigma_n^2}\end{bmatrix}$$

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

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

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

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

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

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

Ezzel az újrabecsült kovariancia-variancia mátrixxal az új paraméterbecslésünk

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

Az iteráció tetszőlegesen sokáig folytatódhat, és minden iterációval egyre közelebb kerülünk a tényleges $\boldsymbol{\beta}$-hoz.

Paraméterszignifikancia-tesztek lineáris regresszió esetén

A lineáris regresszió tanulmányozása folyamán fontos kitérni a paraméterbecslések *szignifikanciájára*, azaz arra a kérdésre, hogy jelentősen csökken-e a $\hat{\boldsymbol{\beta}}$-val való predikciós/magyarázó erő, ha $\hat{\boldsymbol{\beta}}$ egy vagy több elemét 0-nak vesszük. A szignifikancia tesztelése minden esetben *hipotézisvizsgálat*, a különbség a nullhipotézisek megfogalmazása között van, és az így különböző velejáró tesztstatisztika-eloszlásokban.

.1 t-teszt egyelemes paraméterrestrikcióra

Láttuk, hogy $\hat{\boldsymbol{\beta}}$ elég közel lesz a normális eloszláshoz megfelelően sok megfigyelés mellett (mostantól mindig fix $\boldsymbol{X}$-el dolgozunk). Azt is láttuk, hogy a hibavariancia torzítatlan becslése $\widehat{\sigma^2} = \frac{\boldsymbol{e}^T\boldsymbol{e}}{n-p}$ lesz (ugyan ezt nem bizonyítottuk de akit érdekel utánanézhet ha nagyon unatkozik).

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

Jelölje $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ inverz centralizálatlan (és normálatlan) regresszor-kovariancia mátrix k-adik diagonális elemét $\boldsymbol{L}_k^2$, ez persze semmi más, mint a $0 \leq k \leq p$-adik paraméterbecslés varianciájának és a becsült magyarázatlan $\widehat{\sigma^2}$ hibavarianciának hányadosa. A null- és alternatív hipotéziseink:

H_0	H_1
$\beta_k = 0$	$\beta_k \neq 0$

azaz hogy a nullhipotézis alatt a k-adik paraméterbecslésünk igazi értéke értéke 0. A tesztstatisztikánk:

$$t = \frac{\hat{\boldsymbol{\beta}}_k - 0}{\widehat{\sigma}\boldsymbol{L}_k} = \frac{\hat{\boldsymbol{\beta}}_k}{\boldsymbol{L}_k\sqrt{\frac{1}{n-p}\boldsymbol{e}^T\boldsymbol{e}}}$$

Mivel $\boldsymbol{y}$-ok normális eloszlásúak az $\boldsymbol{X}\boldsymbol{\beta}$ várható értékeik körül, ezért $\boldsymbol{e}^T\boldsymbol{e}$ normális eloszlású valószínűségi változó négyzetösszege, azaz χ^2 eloszlású.

Még mielőtt továbbmennénk, lássuk be, hogy $\hat{\boldsymbol{\beta}}$ független $\boldsymbol{e}^T\boldsymbol{e}$-től.

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

Ha $\boldsymbol{e}$ tényleges hibát az $\boldsymbol{X}\boldsymbol{X}^\dagger$ oszloptér-vetítés mátrixxal írjuk föl, és meggondoljuk, hogy persze $\boldsymbol{I} - \boldsymbol{X}\boldsymbol{X}^\dagger$ szintén vetítés mátrix, csak az $\boldsymbol{X}$ oszlopterére ortogonális vektortérre:

$$\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{X}\boldsymbol{X}^\dagger)\boldsymbol{y} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

ebből mátrixszorzásokkal és $\boldsymbol{y}^T\boldsymbol{y} = \sigma^2$-el megkaphatjuk $\hat{\boldsymbol{\beta}}^T\boldsymbol{e}$-t:

$$\hat{\boldsymbol{\beta}}^T\boldsymbol{e} = \sigma^2(\boldsymbol{X}^\dagger - (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{X}^\dagger) = \sigma^2\boldsymbol{0} = 0$$

Tehát a paraméterbecslésünk valóban független a hibáktól, ezért a tesztstatisztikánkban a χ^2 és a $\mathcal{N}$ eloszlások függetlenek, azaz

$$t = \frac{\hat{\boldsymbol{\beta}}_k - 0}{\hat{\sigma}\boldsymbol{L}_k} = \frac{\hat{\boldsymbol{\beta}}_k}{\boldsymbol{L}_k\sqrt{\frac{1}{n-p}\boldsymbol{e}^T\boldsymbol{e}}} \sim t_{n-p}$$

t-eloszlást követ, $n-p$ szabadságfokkal. A hipotézisvizsgálat szokásos módszertana szerint definiálunk egy α szignifikanciaszintet, és megnézzük, hogy t beleesik-e az α által meghatározott elfogadási tartományba. Ha beleseik, nem tudjuk elutasítani H_0-t, azaz $\hat{\boldsymbol{\beta}}_k$-ról *nem mondhatjuk, hogy nem* 0. Ha kívül esik, akkor $\hat{\boldsymbol{\beta}}_k$ *szignifikáns (szignifikánsan eltér* 0*-tól) egy* α *szignifikancia szint mellett.*

.2 F-teszt többszörös paraméterrestrikcióra

Ha egyszerre több paraméter *közös szignifikanciáját* szeretnénk vizsgálni (például az első $m+1$ paraméterét), akkor a hipotéziseink:

H_0	H_1
$\beta_0 = \beta_1 = \cdots = \beta_m = 0$	$\beta_i \neq 0 \quad \forall i = 0\ldots, m$

Legyen u az *unrestricted*, avagy *teljes* modellünk, ahol egyik paraméterünk sem 0. Legyen r a *restricted*, avagy *korlátozott* modellünk, ahol most speciálisan az első $m+1$ paraméterünk 0 (ez a nullhipotézis melletti modell). Jelölje SSR_u és SSR_r rendre az ezen modellek melletti Sum of Squared Residualsokat, avagy hibanégyzetösszegeket. A tesztstatisztikánk

$$F = \frac{\frac{SSR_r - SSR_u}{m+1}}{\frac{SSR_u}{n-p}}$$

ahol p a teljes modell magyarázó paramétereinek száma, $m+1 < p$ pedig a teljes és korlátozott modellek paraméterszámának különbsége. Gondoljuk meg, hogy a tesztstatisztika sosem lehet negatív, hiszen a teljes modell mellett mindig kisebb lesz a hibanégyzetösszeg (több paraméterrel biztosan jobban fogjuk tudni magyarázni a magyarázott változók varianciáját, kérdés persze, hogy *nem magyarázzuk-e túl* azt). $SSR_r - SSR_u$ és SSR_u χ^2 eloszlásúak, rendre $n-(p-(m+1))-(n-p) = n-p+m+1-n+p = m+1$ és $n-p$ szabadságfokokkal, tehát a tesztstatisztikánk *F-eloszlást követ*:

$$F = \frac{\frac{SSR_r - SSR_u}{m+1}}{\frac{SSR_u}{n-p}} \sim F_{m+1,n-p}$$

A t-teszttel analóg módon itt is megkeressük az ilyen paraméterezésű F-eloszlásból α szignifikancia szint mellett a kritikus értékeket, így az elfogadási tartományt is megtaláljuk, és ha F beleesik ebbe, akkor nem tudjuk elvetni a nullhipotézist, azaz *a korlátozott modell nem magyarázza* $\boldsymbol{y}$ *varianciáját szignifikánsan rosszabban, mint a teljes modell*, így elhagyható a modellből az első $m+1$ magyarázó változó. Fontos kiemelni, hogy ebből csakis az első $m+1$ paraméter *közös szignifikanciájára* következtethetünk, egyenként semmit nem tudunk meg róluk.

.2.1 Az F-teszt és a t-teszt ekvivalenciája

Ha az F-tesztben pontosan egy β-t veszünk 0-nak a nullhipotézis alatt, ez megegyezik az adott β-ra vonatkozó t-teszttel, mégpedig a tesztstatisztikákkal felírva:

$$F_{1,n-p} = t^2_{n-p}$$

A Maximum Likelihood Estimation (MLE)

A lineáris regresszióbeli paraméterbecslésünk egy másik, elterjedt módja az úgynevezett MLE, avagy *Maximum Likelihood Estimation* eljárás. A "likelihood" a megfigyelt adatok valószínűsége valamilyen paraméterek függvényében, az MLE ezt a valószínűséget mint egy szélsőértékfeladatot maximalizálva találja meg azokat a paramétereket, amikkel ez a valószínűség a megfigyelt adatokon a lehető legnagyobb.

A lineáris regresszió modellünk továbbra is

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

Éljünk továbbra is azzal a feltétellel, hogy $\boldsymbol{\epsilon}$ normális eloszlású 0 várható értékkel és σ^2 varianciával. Az OLS-ről szóló fejezetben megnéztük, hogy a modellről alkotott feltételeink mellett $\boldsymbol{y}$ valószínűségi vektorváltozónk eloszlása

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

ahol $\boldsymbol{\Sigma}$ diagonális, és minden diagonálisbeli elem azonosan σ^2, azaz nincs heteroszkedaszticitás. Írjuk fel annak a valószínűségét, hogy valamilyen $\hat{\boldsymbol{\beta}}$ és σ^2 paraméterek mellett az i-edik megfigyelt magyarázott változót kaptuk az i-edik megfigyelt magyarázóvektorból a modell mellett:

$$\mathbb{P}(y_i \mid \boldsymbol{x}_i, \hat{\boldsymbol{\beta}}, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{1}{2\sigma^2}(y_i - \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i)^2\right)$$

Mivel minden megfigyelésünk független, annak a valószínűsége tehát, hogy mind az n darab y_i-t kaptuk:

$$\mathbb{P}(\{y_i\}_{i=1}^n \mid \{\boldsymbol{x}_i\}_{i=1}^n, \hat{\boldsymbol{\beta}}, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{1}{2\sigma^2}(y_i - \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i)^2\right)$$

Valószínűségi vektorváltozókkal kompaktabban kiírva ugyanezt:

$$\mathbb{P}(\boldsymbol{y} \mid \boldsymbol{X}, \hat{\boldsymbol{\beta}}, \sigma^2) = |\boldsymbol{\Sigma}|^{-1/2}(2\pi)^{-n/2} exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})\right)$$

Mivel $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$, így a fenti formula az alábbival egyenértékű:

$$\mathbb{P}(\boldsymbol{y} \mid \boldsymbol{X}, \hat{\boldsymbol{\beta}}, \sigma^2) = (2\sigma^2\pi)^{-n/2} exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})\right)$$

Ez a parametrizált feltételes valószínűség az $L(\hat{\boldsymbol{\beta}}, \sigma^2)$ *likelihood függvény*, melynek logaritmusát véve $l(\hat{\boldsymbol{\beta}}, \sigma^2)$ *log-likelihood függvényt* kapjuk.

$$log(\mathbb{P}(\boldsymbol{y} \mid \boldsymbol{X}, \hat{\boldsymbol{\beta}}, \sigma^2)) =: l(\hat{\boldsymbol{\beta}}, \sigma^2)$$

Mostantól az argumentumok beírásának elhagyásával egyszerűen L és l-ként fogunk hivatkozni erre.

.1 A $\hat{\boldsymbol{\beta}}_{ML}$ és σ^2_{ML} paraméterbecslések

Az MLE célja, hogy ezt az L likelihood függvényt *maximalizálja* $\hat{\boldsymbol{\beta}}$-ban és σ^2-ben. Ha feltesszük, hogy ismerjük σ^2-et, akkor természetesen a maximalizáció csak $\hat{\boldsymbol{\beta}}$-ban lesz. Mivel szorzatot nehéz differenciálni, és mivel a *log* függvény szigorúan monoton nő, így az L függvény maximalizálása helyett l-et fogjuk maximalizálni. l alakja

$$l = -\frac{n}{2} log(2\sigma^2\pi) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$$

Tegyük fel, hogy ismerjük σ^2-et. Formálisan felírva $\hat{\boldsymbol{\beta}}$-t megkaphatjuk az alábbi módon (az összeg első első tagja most konstans):

$$\hat{\boldsymbol{\beta}}_{ML} = \underset{\hat{\boldsymbol{\beta}}}{argmax} \left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \right)$$

Mivel a konstans szorzó nem befolyásolja a szélsőérték feladat optimális megoldását, így a mínusz előjelet elhagyva és minimalizációs problémára átírva:

$$\hat{\boldsymbol{\beta}}_{ML} = \underset{\hat{\boldsymbol{\beta}}}{argmin} \left((\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \right)$$

Ha összehasonlítjuk ezt az OLS becslés szélsőérték feladatával, látjuk, hogy *pontosan ugyanazt kaptuk* (csak ott $\boldsymbol{e}^T\boldsymbol{e}$-el volt felírva). Kimondhatjuk tehát, hogy *ha normális eloszlású hibát feltételezünk a lineáris regressziós modellben, és ezen kívül minden egyéb feltétel szintén teljesül, akkor az MLE becslés eredményének ugyanazt kapjuk, mint az OLS becslés eredménye, $\hat{\boldsymbol{\beta}}_{ML} = \hat{\boldsymbol{\beta}}_{OLS}$.*

A Maximum Likelihood becslés azonban nem csak $\hat{\boldsymbol{\beta}}$ becslését tudja megadni, hanem σ^2-ét is! Ha most megfordítjuk a helyzetet, és feltesszük, hogy már ismerjük $\hat{\boldsymbol{\beta}}_{ML}$-t, akkor a fenti szélsőértékfeladat már σ^2-ről szól:

$$\sigma^2_{ML} = \underset{\sigma}{argmax} \left(-\frac{n}{2} log(2\sigma^2\pi) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{ML})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{ML}) \right)$$

Ezt megoldva kapjuk σ^2 MLE becslését:

$$\widehat{\sigma^2}_{ML} = \frac{1}{n}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{ML})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{ML})$$

Ez persze semmi más, mint

$$\widehat{\sigma^2}_{ML} = \frac{1}{n}\boldsymbol{e}^T\boldsymbol{e}$$

ahol $\boldsymbol{e}$-k a $\hat{\boldsymbol{\beta}}_{ML}$-el becsült $\hat{y}_i$-k és a tényleges y_i-k hibavektora. Felmerülhet a kérdés, hogy miért n-el, és nem $n-1$-el osztjuk le $\boldsymbol{e}^T\boldsymbol{e}$-t a varianciabecslésnél, mint eddig. Az n-el való leosztás valóban nem eredményez torzítatlan varianciabecslést, azonban emlékezzünk vissza a *Mean Squared Error (MSE)* definíciójára, most $\widehat{\sigma^2}_{ML}$-re nézve:

$$MSE = \frac{\sum_{i=1}^{n} (\widehat{\sigma^2}_{ML} - \sigma^2)^2}{n}$$

Ha $n-1$-el osztottunk volna le, ugyan torzítatlan becslésünk lenne, de *az MSE nagyobb lenne.* Mivel $MSE = Bias^2 + Var$, így ugyan csökkentettük $Bias^2$-et (így persze $Bias$-t is), de ezzel *növeltük* a becslés varianciáját. Az, hogy melyik a jobb vagy rosszabb, nagyon függ attól a becslési feladat kontextusától, és általánosan nem is könnyű az, hogy "melyik rossz a kevésbé rossz".

A hibanormalitási és az egyéb feltételek mellett tehát az MLE becslés az OLS-el együtt $BLUE$. Fontos, hogy a Maximum Likelihood Estimation technika nem csak regressziós modellek esetén alkalmazható, ezért is különösen fontos nekünk a hibanormalitás.

Ha nem tesszük fel $\boldsymbol{y}$ normális eloszlását, akkor az MLE becslés *hatásosabb* az OLS-nél, és numerikus kiszámítása is stabilabb (nem kell magas dimenziójú mátrixokat invertálgatnunk - multikollinearitás esetén láttuk, hogy az OLS becslés kiszámítása ott problémákba tud ütközni. Összességében tehát a feltételeink mellett (a hibanormalitást és $\boldsymbol{y}$ normális eloszlását is hozzávéve persze) kijelenthető, hogy a kettő megegyezik.

.1.1 Fisher-információ

Az MLE-becslés *bizonytalansága* a *Fisher információval* fejezhető ki. Formálisan a Fisher információ azt az információmennyiséget adja meg, amennyit a megfigyelt $\boldsymbol{X}$ adat az ismeretlen $\theta \in \Theta$ paraméterről *elárul* a megfigyelést követően. A következőkben jelölje $f(\boldsymbol{X};\theta)$ $\boldsymbol{X}$ megfigyelésének valószínűségét θ mellett. Definiáljuk a *log-likelihood function* θ szerinti parciális deriváltját

$$s(\boldsymbol{X};\theta) = \frac{\partial}{\partial\theta} log f(\boldsymbol{X};\theta),$$

ezt *score-function*nek hívjuk. Könnyen belátható, hogy ha feltesszük, hogy $\boldsymbol{X}$ valóban $f(\boldsymbol{X};\theta)$ eloszlást követ, akkor az igazi θ paraméter mellett vett várható értéke a scorenak pontosan 0:

$$\mathbb{E}\Big[s(\boldsymbol{X};\theta)|\theta\Big] = \int_{\mathbb{R}} \frac{\frac{\partial}{\partial\theta} f(x;\theta)}{f(x;\theta)} f(x;\theta)dx = \frac{\partial}{\partial\theta}\int_{\mathbb{R}} f(x;\theta)dx = \frac{\partial}{\partial\theta} 1 = 0$$

A Fisher információ a score function varianciája, formálisan

$$\mathcal{I}(\theta) = Var[s(\boldsymbol{X};\theta)]$$

Előbb beláttuk, hogy $\mathbb{E}[s(\boldsymbol{X};\theta)|\theta] = 0$, így - mostantól egyszerűsített jelölésekkel - $\mathbb{E}[s_\theta]^2 = 0$ is fennáll. Ezek szerint tehát

$$\mathcal{I}_\theta = \mathbb{E}[s_\theta^2] - \mathbb{E}[s_\theta]^2 = \mathbb{E}[s_\theta^2]$$

Tudjuk, hogy

$$\frac{\partial^2}{\partial\theta^2} log f_\theta = \frac{\frac{\partial^2}{\partial\theta^2} f_\theta}{f_\theta} - \Big(\frac{\frac{\partial}{\partial\theta} f_\theta}{f_\theta}\Big) = \frac{\frac{\partial^2}{\partial\theta^2} f_\theta}{f_\theta} - \Big(\frac{\partial}{\partial\theta} log f_\theta\Big)^2$$

Mindkét oldalon várhatóértéket véve, és megfigyelve, hogy $\mathbb{E}\Big[\frac{\frac{\partial^2}{\partial\theta^2} f_\theta}{f_\theta}|\theta\Big] = \frac{\partial^2}{\partial\theta^2}\int_{\mathbb{R}} f_\theta(x)dx = 0$,

$$\mathbb{E}\Big[\frac{\partial^2}{\partial\theta^2} log f_\theta\Big] = -\mathbb{E}\Big[\Big(\frac{\partial}{\partial\theta} log f_\theta\Big)^2\Big]$$

Megfordítva az előjeleket, és az elsőrendű parciális derivált helyére s_θ-t írva adódik, hogy

$$\mathcal{I}_\theta = \mathbb{E}[s_\theta^2] = -\mathbb{E}\Big[\frac{\partial^2}{\partial\theta^2} log f_\theta|\theta\Big]$$

Intuitíven ez azt jelenti, hogy *minél gyorsabban változik a meredeksége a log-likelihood functionnek* egy adott θ mellett, annál nagyobb bizonyossággal írja le a megfigyelt adatokat θ. Ha ez az érték kicsi, akkor a log-likelihood function alakja lapos. Fontos megjegyezni, hogy mivel $\mathbb{E}[s|\theta] = 0$, így a log-likelihood maximuma nyilván a valódi θ paraméter mellett lesz, a görbület e körül a valódi érték körül értendő. A negatív előjel a likelihood függvény konkavitása miatt kell, hiszen a második derivált negativitása miatt nem akarunk negatív információt definiálni. Természetesen itt $\theta \in \Theta$-ra mint tetszőleges elemszámú (lehetőleg véges) paramétervektor tekinthetünk , így az aszerinti másodrendű parciális derivált egy *Hesse-mátrix*ot fog adni. Intuitíven az is következik tehát, hogy ez a mátrix pontosan az MLE-becslés *variancia-kovariancia mátrixát* fogja megadni.

Általánosabb lineáris modellek és kvalitatív változók

Az előző fejezetekben csupán a lineáris $y_i = \boldsymbol{\beta}^T \boldsymbol{x}_i$ modellekkel foglalkoztunk, azonban a gyakorlatban nagyon sokszor nem elég ez a szimplisztikus megközelítés. Ebben a fejezetben bevezetjük a *polinomiális regresszió* fogalmát, majd ennek segítségével a bonyolultabb modellek reprezentációját.

.1 A "feature transform" függvény - ϕ

Ahhoz, hogy elrugaszkodjunk a lineáris modellek egyszerűségétől, először is be kell vezetni az úgynevezett *feature transform* (magyarázó változó-transzformáló) függvényt, melyre ezentúl ϕ néven hivatkozunk. Legyen $p \in \mathbb{N}$ az eredeti magyarázó változók száma. Azt mondjuk, hogy a

$$\phi : \mathbb{R}^p \mapsto \mathbb{R}^D, \quad D \in \overline{\mathbb{R}}, D > p$$

függvény minden magyarázó változó-vektort egy *magasabb dimenziójú térbe ágyaz be.* Nézzünk egy konkrét példát:

Legyen $\boldsymbol{x}_i \in \mathbb{R}^2$ magyarázó változó-vektora az i-edik megfigyelésnek. Jelöljük a vektor elemeit az alábbi módon:

$$\boldsymbol{x}_i = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Most vegyünk egy ϕ transzformációt, melyet definiáljunk a következőképpen:

$$\phi\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) := \begin{bmatrix} x_1 \\ x_2 \\ x_1^2 \\ x_2^2 \\ x_1 x_2 \end{bmatrix}$$

Ebben az esetben ϕ kettő dimenziós vektorokból 5 dimenziójúakat csinál, mégpedig úgy, hogy az új, magasabb dimenziós vektor elemei *lineárisan függetlenek* (Gondoljuk meg, hogy $x_1 + x_2 + x_1^2 + x_2^2 + x_1 x_2 = 0$ akkor és csak akkor, ha minden elem pontosan 0. Ellenpéldaként például ha $[x_1, x_2]^T \mapsto [x_1, x_2, 2x_1, 2x_2]^T$ lenne ϕ, akkor ez lineárisan összefüggő elemeket eredményez, *nem kaptunk igazi új magyarázó változókat.* Fontos megjegyezni, hogy elméletben semmi nem akadályoz meg minket végtelen dimenziós featurevektorokat kapjunk ϕ által, persze a gyakorlatban ez nem megvalósítható.

$\boldsymbol{X}$ design-mátrix soraiban most $\boldsymbol{x}_i$-k helyett tehát $\phi(\boldsymbol{x}_i)$-k lesznek, így ha az új design mátrixot $\boldsymbol{\Phi}$-vel jelöljük:

$$\boldsymbol{\Phi} = \begin{bmatrix} - & \phi(\boldsymbol{x}_1) & - \\ - & \phi(\boldsymbol{x}_2) & - \\ \vdots & \vdots & \vdots \\ - & \phi(\boldsymbol{x}_n) & - \end{bmatrix}$$

$\mathbf{\Phi}$ a gyakorlatban sokszor *lényegesen* nagyobb dimenziójú, mint $\boldsymbol{X}$, így az ezzel való mátrixoperációk is szignifikánsan lassabbak lesznek. A probléma megoldásához úgynevezett *kernelizáció* szükséges, de ezzel nem fogunk foglalkozni részletesen.

.1.1 A polinomiális regresszió

A ϕ transzformáció egy speciális esetében polinomiális regresszióról beszélünk. Legyen $m \in \mathbb{R}$ maximális polinomfokszám és az eredeti magyarázó változó-vektor dimenziója 1. $\mathbf{\Phi}$ alakja ekkor

$$\mathbf{\Phi} = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^m \\ 1 & x_2 & x_2^2 & \dots & x_2^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^m \end{bmatrix}$$

Az ilyen $\mathbf{\Phi}$ transzformált design-mátrixot *polinomiális design-mátrixnak* nevezünk, és ez szolgál alapul a polinomiális regresszióhoz. A regressziós modellünk tehát az alábbi formában írható fel ilyenkor:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \dots + \beta_m x_i^m$$

.2 Interakciós változók

Emlékezzünk vissza a design-mátrix egy ϕ feature-transformmal kibővített általános formájára, ahol is egy adott $\boldsymbol{x} \in \mathbb{R}^m$ featurevektor egy magasabb $D > m$ dimenziós featurevektorrá válik. A design mátrixunkat ilyenkor $\mathbf{\Phi}$-al jelöljük. A különbség a polinomális regresszió és az interakciós változós regresszió között csupán annyi, hogy míg a polinomiális esetben $\mathbf{\Phi}$ nem tartalmazott különböző featureök kombinációjából álló új feature-t, az interakciós változóknak ellenben pontosan ez a lényege, a szorzat-tagok a különböző featureök közti interakciót jelentik. Legyen példának okáért az eredeti $\boldsymbol{x}$ featurevektorunk 2-dimenziós, rendre x_1 és x_2 featureökkel. Ekkor egy interakciós változókat tartalmazó design-mátrix lehetséges alakja:

$$\mathbf{\Phi} = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{11}x_{12} \\ 1 & x_{21} & x_{22} & x_{21}x_{22} \\ 1 & x_{31} & x_{32} & x_{31}x_{32} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{N1} & x_{N2} & x_{N1}x_{N2} \end{bmatrix}$$

Az interakciós változók lehetővé teszik lineáris kapcsolatok modellezését különböző intercept és meredekségekkel. Ezt legkönnyebben egy konkrét példával érthetjük meg. Fontos kihangsúlyozni, hogy csakúgy mint eddig, semmi nem akadályoz meg minket kvalitatív (diszkrét értékeket felvevő) magyarázó változók (featureök) használatában, mint azt látni fogjuk, a velük történő interakcióknak is hasznos értelmezése lesz.

Tekintsük az alábbi modellezési feladatot: Legyen LE a *life expectancy at birth*, amit $COUNTRY$, $RADIATION$, GDP és SUN (human-developement index) magyaráznak (a példa relatíve morbid, ettől most tekintsünk el). GDP és SUN mint folytonos változók, a többi mint diszkrétek vannak kezelve, a $RADIATION$ bináris (2-értékű) változó, míg a SUN a napsütés erősségét jelenti. Feltesszük, hogy a magyarázó változók lineárisan függetlenek egymástól, valamennyi korreláció persze előfordulhat, de most ettől eltekintünk. Az interakciók nélküli egyszerű lineáris modellünket a következőképpen írhatjuk fel:

$$LE_i = \beta_0 + \beta_1 COUNTRY_i + \beta_2 RADIATION_i + \beta_3 GDP_i + \beta_4 SUN_i$$

Most vezessük be a $RADIATION$ és a SUN interakcióját, melyet $RADIATION * SUN$-al jelölünk:

$$LE_i = \beta_0 + \beta_1 COUNTRY_i + \beta_2 RADIATION_i + \beta_3 GDP_i + \beta_4 SUN_i + \beta_5 RADIATION_i * SUN_i$$

Ez annyit jelent, hogy azoknál az előrejelzéseknél, ahol $RADIATION = 1$, ott a SUN együtthatója nem csupán β_4, hanem $\beta_4 + \beta_5$ lesz, azaz *az interakciós változókkal bővített modellben előfordulhat különböző meredekség kvalitatív és kvantitatív interakciós változók bevezetése mellett.* Bővítsük a modellünket a

$COUNTRY * RADIATION$ interakcióval, ekkor a modell

$$LE_i = \beta_0 + \beta_1 COUNTRY_i + \beta_2 RADIATION_i + \beta_3 GDP_i + \beta_4 SUN_i + \beta_5 RADIATION_i * SUN_i +$$

$$+\beta_6 COUNTRY_i * RADIATION_i$$

Mivel az új interakció mindkét tagja kvalitatív (diszkrét), így ha egy előrejelzésnél $RADIATION = 1$, $COUNTRY$ együtthatója $\beta_1+\beta_6$ lesz, azaz ha az x-tengelyen a SUN-t mint magyarázó változót képzeljük el, akkor az *intercept* $\beta_0+\beta_1 COUNTRY$-ról $\beta_0+(\beta_1+\beta_6)COUNTRY$-ra változott. Kimondhatjuk tehát, hogy ebben az egyszerű esetben míg egy diszkrét és egy folytonos magyarázó változó interakciója a meredekséget, kettő diszkrét változó interakciója az interceptet engedi változtatni, ami a modellben különböző interceptű és meredekségű lineáris kapcsolatokat eredményez, ezzel egyfajta flexibilitást nyújtva a modellezőnek (nekünk).

Természetesen semmi nem akadályoz meg minket abban, hogy egyszerre több változó interakcióját is beletegyük a modellben, habár ennek intuitív értelmezése egy kicsit nehezebb, ezért ezzel a továbbiakban nem fogunk mélyebben foglalkozni. Természetesen - mint azt a feature-transformoknál láttuk - különböző változók interakciói szükségképpen lineárisan függetlenek lesznek egymástól (gondoljunk itt például az $x_1 * x_2$, $x_1 * x_3$, $x_2 * x_3$, $x_1 * x_2 * x_3$ esetére, aholis ezek mind lineárisan függetlenek egymástól).