Eötvös Loránd Tudományegyetem

Természettudományi Kar

Kiri Csaba

Regresszió és alkalmazása a meteorológiában

BSc Szakdolgozat

Témavezető:

Kovács Ágnes

Alkalmazott Analízis és Számításmatematikai Tanszék

Budapest, 2019

Tartalomjegyzék

1. Bevezetés **5**

2. Egyszerű lineáris regresszió **6**

2.1. Modell leírása 6

2.2. Regressziós együtthatók becslése 8

2.2.1. A legkisebb négyzetek módszere 8

2.2.2. Maximum likelihood becslés 9

2.2.3. Hibabecslés 10

2.3. Szóródás és hipotézisvizsgálat 11

2.3.1. Teljes szóródás felbontása 11

2.3.2. Hipotézisvizsgálat a regresszió együtthatóira 12

2.4. Az illesztett modell minősége 13

3. Többszörös lineáris regresszió **14**

3.1. Modell leírása 14

3.1.1. A modell leírása 14

3.1.2. Linearizáló transzformációk 15

3.2. Regressziós együtthatók becslése 16

3.2.1. Legkisebb négyzetek módszere 16

3.2.2. Hibabecslés 17

3.2.3. Maximum likelihood becslés 17

3.2.4. A regressziós együtthatók értelmezése 18

3.3. Szóródás és hipotézisvizsgálat 18

3.3.1. Teljes szóródás felbontása 18

3.3.2. Regressziós együtthatók hipotézisvizsgálata 18

3.3.3. Konfidenciaintervallum illesztés az együtthatókra 19

3.4. Az illesztett modell minősége . . . 20
3.5. Feltételek . . . 21
3.5.1. Normalitás . . . 21
3.5.2. Multikollinearitás . . . 22
3.5.3. Heteroszkedaszticitás . . . 23
3.5.4. Autokorreláció . . . 23

4. Logisztikus regresszió **25**
4.1. Modell leírása . . . 25
4.2. Modell értelmezése . . . 27
4.3. Az illesztett modell minősége . . . 28
4.3.1. Paraméterek szignifikanciája . . . 28
4.3.2. Akaike Információs Kritérium (AIC) . . . 29
4.3.3. Változók kiválasztása: stepwise szelekció . . . 30
4.4. Feltételek . . . 30
4.4.1. Linearitás . . . 30
4.4.2. Multikollinearitás . . . 31
4.4.3. Mintanagyság . . . 31

5. Alkalmazás a meteorológiában **32**
5.1. Feladat leírása . . . 32
5.2. Adatok bemutatása . . . 33
5.3. Előzetes vizsgálatok . . . 37
5.4. Modell felépítése . . . 39
5.4.1. Tanuló és teszt adatokra bontás . . . 39
5.4.2. Lineáris regresszió felépítése . . . 40
5.4.3. Előrejelzés a lineáris modell alapján . . . 42
5.4.4. Logisztikus modell felépítése . . . 44
5.4.5. Előrejelzés a logisztikus modell alapján . . . 46

6. Összegzés, további vizsgálatok **47**

A. Részlet az R programkódból **50**

Köszönetnyilvánítás

Szeretném megköszönni témavezetőmnek Kovács Ágnesnek a dolgozatom készítése közben tanúsított mérhetetlen türelmét és segítőkészségét. Valamint köszönettel tartozom Kolláth Kornélnak az Országos Meteorológiai Szolgálat munkatársának, aki rendelkezésemre bocsátotta a statisztikai vizsgálatokhoz szükséges adatokat és segített azok kivizsgálásában, illetve a meteorológiai részek lektorálásaban.

1. fejezet

Bevezetés

A regresszió a tudomány számos területén alkalmazott statisztikai módszer. Elnevezése az angol regression to the mean kifejezésből ered (jelentése: visszatérés az átlaghoz) Francis Galtontól[15], aki a 19. században élő angol polihisztor volt és azt figyelte meg, hogy a magasabb apáknak magasabb fiaik születnek. A módszer nagyon leegyszerűsítve arról szól, hogy függvényszerű kapcsolatot keressünk egy általunk kiválasztott (függő) változó és egy vagy több (magyarázó) változó értékei között. Fontos feltétel a magyarázó, illetve a függő változókra nézve, hogy legalább intervallum skálán legyenek mérhetőek. Ez azt jelenti, hogy a változók által felvett számértékek egymáshoz viszonyított nagysága, illetve a különbségük mértéke is értelmezhető. Erre jó példa a hőmérséklet (például Celsius fokban mérve), hisz el lehet dönteni, hogy a 20°C és 10°C közül melyik jelent magasabb hőmérsékletet, illetve a kettőjük közötti különbség is értelmezhető. Galton a meteorológia fejlődéséhez is hozzájárult (anticiklon felfedezése, első népszerű időjárási térképek).

Dolgozatom első felében az egyszerű, illetve többszörös lineáris regresszió és a logisztikus regresszió elméleti hátteréről írtam. A második felében a regresszió egy gyakorlati alkalmazásának megvalósítását írtam le a meteorológiában. Az elméleti rész leírásában a regresszió azon részeivel foglalkoztam részletesebben, amelyeket az alkalmazás során is felhasználtam. Részleteztem a különböző regressziós modellek leírását, illetve feltételeit, majd a regressziós paraméterek becsléséről írtam. Bemutatom az illesztett modellek mérőszámait, hipotézisvizsgálatát és jelentésüket. A gyakorlati alkalmazás leírásában az Országos Meteorológiai Szolgálat által részemre bocsátott zivatarokkal kapcsolatos előrejelzési paraméterek és az adott napon detektált villámok számának kapcsolatáról, illetve a regressziós modellek alapján történő villámkisülésekre vonatkozó előrejelzésekről lesz szó. A megvalósításhoz az ingyenesen hozzáférhető R programcsomagot használtam.

2. fejezet

Egyszerű lineáris regresszió

2.1. Modell leírása

A regresszió különböző változók közötti összefüggés leírására szolgál. A legegyszerűbb eset, amikor két változó közötti lineáris összefüggést vizsgálunk. Ebben az esetben az y (függő, magyarázandó, output vagy célváltozó) változót szeretnénk az x (magyarázó, input vagy független) változó segítségével felírt lineáris függvénnyel becsülni. Feltesszük, hogy az x magyarázó változó konstans vagy elhanyagolható hibával rendelkezik. A cél: adott n pár (y_i, x_i) megfigyelés alapján becsülni a lineáris függvény együtthatóit. Az egyszerű lineáris regresszió alkalmazása előtt érdemes egy táblázatba rendezni az adatokat, például a következő formában:

Függő változók értékei	Magyarázó változó értékei
y_1	x_1
y_2	x_2
y_3	x_3
$\vdots$	$\vdots$
y_n	x_n

Már a táblázat felírásakor észrevehetőek egyszerűbb összefüggések (például nagyobb x_n értékhez nagyobb y_n tartozik stb.), melyek könnyíthetnek a további vizsgálatok irányán. Ezen felül érdemes egy koordináta-rendszerben ábrázolni az adatokat, hiszen így lesz szembetűnő, hogy megfelelő lehet-e számunkra a lineáris regresszió. Ha a vízszintes

koordinátatengelyen felvesszük az x, illetve a függőleges tengelyen az y változó értékeit és az ezek által ábrázolt ponthalmaz közel egy egyenesen fekszik, akkor feltételezhetünk lineáris kapcsolatot a változók között. Ha pontosan egy egyenesre esnének az ábrázolt adatok, akkor az egyenes felírható lenne a következő egyenlettel:

$$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x} \tag{2.1}$$

A β_0 jelöli az egyenes metszéspontját az y tengellyel, a β_1 pedig az egyenes meredekségével egyenlő. Azonban szinte soha nincs ilyen tökéletes lineáris kapcsolat a változók között. Egy hibataggal (általában ε) jelöljük a pontok eltérését a regressziós egyenestől.

Az egyszerű lineáris regresszió statisztikai modellje:

Tegyük fel, hogy a megfigyelt adatokra a következő lineáris modellt szeretnénk illeszteni

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n \tag{2.2}$$

ahol $\varepsilon_1, \ldots, \varepsilon_n$ a fent említett hibatagok, melyekről feltesszük, hogy függetlenek, 0 várható értékű és azonos, de ismeretlen σ^2 szórásnégyzetű valószínűségi változók. A későbbiekben szintén fel fogjuk tenni, hogy a hibák normális eloszlásúak (utóbbi feltételt a paraméterek hipotézisvizsgálatakor fogjuk felhasználni). A cél a β_0 és β_1 regressziós együtthatók becslése. Ugyanez felírva vektoralakban a következőképp néz ki:

$$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x} + \boldsymbol{\varepsilon} \tag{2.3}$$

Ennek a modellnek egy alternatív formája a következőképpen írható le:

$$y_i = \beta_0^* + \beta_1 (x_i - \overline{x}) + \varepsilon_i \quad i = 1, 2, \ldots, n \tag{2.4}$$

ahol $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ az x_i magyarázó változó értékeinek az átlaga, illetve $\beta_0^* = \beta_0 + \beta_1 \overline{x}$.

A két modell ekvivalens, de az eltolási paraméterek (β_0 ill. β_0^*) értelmezése megváltozik. Az előző modell esetén az y_i függő változónak a várható értéke $E(y_i|x_i) = \beta_0 + \beta_1 x_i$, míg az utóbbi modellben ugyanez a várható érték éppen β_0^*, azaz $E(y_i|x_i) = \beta_0^*$. A β_0 paraméter jelentése az $x = 0$ értékhez tartozó becsült $\widehat{y}$ érték (bár ez nem minden esetben értelmezhető), míg β_0^* paraméter jelentése a becsült $\widehat{y}$ értékek átlaga. A két modellben a β_1 értelmezése azonos, azt mutatja, hogy x egy egységgel növelt értékéhez y-nak átlagosan mennyivel nagyobb értéke tartozik.

Az alternatív modell felírást a regressziós együtthatók becslésekor fogjuk felhasználni.

2.2. Regressziós együtthatók becslése

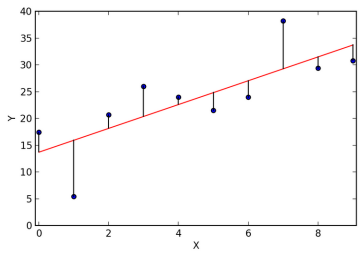
Az ismeretlen β_0 és β_1 paramétereket az n pár $(x_1, y_1), \ldots, (x_n, y_n)$ adatokból fogjuk becsülni. Először a legkisebb négyzetek módszerét mutatjuk be, ezt követően a maximum likelihood becslést, majd ezek kapcsolatát fogjuk tárgyalni.

2.2.1. A legkisebb négyzetek módszere

A legkisebb négyzetek módszerének célja, hogy minimalizálja a megfigyelt adatok és az illesztett modellből számított értékek közti különbségek négyzetösszegét, vagyis a reziduális négyzetösszeget. Pontosabban olyan b_0 és $b_1$becslést szeretnénk kapni a β_0 és β_1 együtthatókra, amelyekkel felírva az

$$\widehat{y_i} = b_0 + b_1 x_i \tag{2.5}$$

regressziós egyenletet teljesül, hogy a reziduálisok négyzetösszege, azaz a $\sum_{i=1}^{n}(y_i - \widehat{y_i})^2$ érték minimális lesz.



2.1. ábra. Reziduálisok a $\widehat{y} = b_0 + b_1 x$ regressziós modellben

Ehhez a becsült b_0 és b_1 változók szerinti parciális deriválással keressük meg a megfelelő szélsőértékeket:

$$\frac{\partial}{\partial b_0}\left[\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2\right] = 0 \tag{2.6}$$

$$\frac{\partial}{\partial b_1}\left[\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2\right] = 0 \tag{2.7}$$

Egyszerűbb a megoldás levezetése, ha az alternatív modell felírásból indulunk ki:

$$\frac{\partial}{\partial b_0^*}\sum_{i=1}^{n}\left[y_i - b_0^* - b_1\left(x_i - \overline{x}\right)\right]^2 = 0 \tag{2.8}$$

$$\frac{\partial}{\partial b_1}\sum_{i=1}^{n}\left[y_i - b_0^* - b_1\left(x_i - \overline{x}\right)\right]^2 = 0 \tag{2.9}$$

A deriválás elvégzése után kapjuk a következő egyenletrendszert:

$$nb_0^* + b_1\sum_{i=1}^{n}\left(x_i - \overline{x}\right) = \sum_{i=1}^{n} y_i \tag{2.10}$$

$$b_0^*\sum_{i=1}^{n}\left(x_i - \overline{x}\right) + b_1\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 = \sum_{i=1}^{n} y_i\left(x_i - \overline{x}\right) \tag{2.11}$$

Tekintve, hogy $\sum_{i=1}^{n}\left(x_i - \overline{x}\right) = 0$, az együtthatók legkisebb négyzetek módszerével kapott becslése:

$$b_1 = \frac{S_{xy}}{S_{xx}} \tag{2.12}$$

$$b_0^* = \overline{y} \tag{2.13}$$

ahol $S_{xy} = \sum_{i=1}^{n}\left(x_i - \overline{x}\right)y_i$ és $S_{xx} = \sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$. Innen az eredeti modell felírásban szereplő β_0 becslése:

$$b_0 = \overline{y} - b_1\overline{x} \tag{2.14}$$

Ezekkel a becsült együtthatókkal írható fel a regressziós egyenes:

$$\widehat{y} = b_0 + b_1 x \tag{2.15}$$

2.2.2. Maximum likelihood becslés

Az eddig leírt legkisebb négyzetek módszerével történő becslésnél feltettük az ε_i hibatagokról, hogy függetlenek, 0 a várható értékük, illetve egy közös σ^2 varianciával rendelkeznek, azonban nem állítottunk semmit az eloszlásukról. Ebben a fejezetben tegyük fel, hogy az eddigi tulajdonságaik mellett normális eloszlásból származnak.

A regressziós együtthatókat most úgy keressük, hogy a mellettük bekövetkezett eredmény maximális valószínűségű legyen, azaz a likelihood függvény maximumát keressük

β_0-ban és β_1-ben. Tegyük fel, hogy az ε_i hibatagok függetlenek és azonos $N(0, \sigma^2)$ eloszlásból származnak. Ekkor a likelihood függvény a következőképp néz ki:

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} f(\varepsilon_i) = \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n} e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}\varepsilon_i^2} = \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n} e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2} \tag{2.16}$$

ahol $f(\varepsilon_i)$ az ε_i i. hibatag sűrűségfüggvénye.

Jelen esetben a megfelelő β_0 és β_1 együtthatók megtalálása - amelyekkel L maximális lesz - éppen azt jelenti, hogy $-\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$ minimumát keressük. Ez ekvivalens a legkisebb négyzetek módszerével történő becsléssel, azaz a reziduális négyzetösszeg minimalizálásával. Következésképp ha feltesszük, hogy a hibatagok normális eloszlásból származnak, akkor a β_0 és β_1 együtthatók becslése ezzel a módszerrel azonos lesz a már fentebb taglalt legkisebb négyzetes becsléssel. Fontos megjegyezni, hogy amennyiben nincs lehetőségünk ezt feltételezni, akkor a két becslési eljárással kapott együtthatók nem fognak megegyezni.

2.2.3. Hibabecslés

A további vizsgálatokhoz szükségünk van az együtthatók becslésén túl a hiba varianciájának becslésére is. Először a legkisebb négyzetek módszerével történő becslésből indulunk ki (nem feltételezzük a normális eloszlást a hibatagokra). Az eredeti σ^2 varianciát az illesztett modellből kapott varianciával becsüljük az alábbi módon:

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{n-2} \tag{2.17}$$

A reziduális négyzetösszeget azért $(n-2)$-vel osztjuk, mert az együtthatókra vonatkozó becslésnél elveszítünk 2 szabadsági fokot. Fontos megjegyezni, hogy ez a becslés torzítatlan.

Maximum likelihood módszerénél az alábbi módon alakul σ^2 becslése. Tekintsük a megfelelő likelihood függvényt, ahol a becsült b_0 és b_1 együtthatókat behelyettesítettük:

$$L\left(\sigma^2\right) = \frac{1}{(2\pi)^{\frac{n}{2}}\left(\sigma^2\right)^{\frac{n}{2}}} e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2} \tag{2.18}$$

Vegyük a logaritmusát (loglikelihood) és deriváljuk parciálisan a függvényt σ^2 szerint, tekintve, hogy ebben a változóban keressük a szélsőértéket:

$$\frac{\partial \log L\left(\sigma^2\right)}{\partial \sigma^2} = \frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 \tag{2.19}$$

Ezt a deriváltat 0-val egyenlővé téve kapjuk a következő becslést a varianciára:

$$\widehat{\sigma}^2 = \sum_{i=1}^{n}\frac{(y_i - \widehat{y}_i)^2}{n} \tag{2.20}$$

Tehát a maximum likelihood módszerrel egy torzított becslést kapunk a varianciára, ami a következő alakban is felírható:

$$\widehat{\sigma}^2 = s^2\left(\frac{n-2}{n}\right) \tag{2.21}$$

2.3. Szóródás és hipotézisvizsgálat

2.3.1. Teljes szóródás felbontása

Ebben a fejezetben szeretnénk kapcsolatot találni a teljes négyzetösszeg és a regressziós négyzetösszeg között. Elemzéskor a célunk, hogy az illesztett $\widehat{y}_i$ értékek az eredeti y_i értékek közelében legyenek, ekkor teljesülni fog, hogy az $\widehat{y}_i$ értékek varianciája $\overline{y}$ körül megfelelően közel lesz az y_i értékek varianciájához $\overline{y}$ körül. Érdemes észrevenni, hogy

$$\overline{\widehat{y}} = \sum_{i=1}^{n}\frac{\widehat{y}_i}{n} = \frac{1}{n}\sum_{i=1}^{n}\left[b_0^* + b_1\left(x_i - \overline{x}\right)\right] = \frac{1}{n}\sum_{i=1}^{n}\left[\overline{y} + b_1\left(x_i - \overline{x}\right)\right] = \overline{y} \tag{2.22}$$

Természetes tehát, hogy úgy tekintsünk a teljes négyzetösszegre ($\sum_{i=1}^{n}(y_i - \overline{y})^2$) és a regressziós négyzetösszegre ($\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2$), mint a megfelelő varianciák forrása. Az elsőből számítható ki a megfigyelt függő változó varianciája, míg a másodikból a regresszió által megmagyarázott variancia. A közöttük levő összefüggés pedig a következő:

2.3.1. Tétel.

$$\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2 = \sum_{i=1}^{n}\left(\widehat{y}_i - \overline{y}\right)^2 + \sum_{i=1}^{n}\left(y_i - \widehat{y}_i\right)^2 \tag{2.23}$$

Bizonyítás. Induljunk ki a következő egyenletből:

$$y_i - \overline{y} = \widehat{y}_i - \overline{y} + y_i - \widehat{y}_i \tag{2.24}$$

Összegezzük és emeljük négyzetre mindkét oldalt:

$$\sum_{i=1}^{n}(y_i-\overline{y})^2=\sum_{i=1}^{n}(\widehat{y_i}-\overline{y})^2+\sum_{i=1}^{n}(y_i-\widehat{y_i})^2+2\sum_{i=1}^{n}(\widehat{y_i}-\overline{y})(y_i-\widehat{y_i}) \tag{2.25}$$

Amennyiben a jobb oldal utolsó tagja 0, úgy végeztünk. Ez pedig fennáll, mert:

$$\sum_{i=1}^{n}(\widehat{y_i}-\overline{y})(y_i-\widehat{y_i})=\sum_{i=1}^{n}\widehat{y_i}(y_i-\widehat{y_i})= \tag{2.26}$$

$$=\sum_{i=1}^{n}\left[b_0+b_1(x_i-\overline{x})\right]\left[y_i-\widehat{y_i}\right]=b_1\sum_{i=1}^{n}x_i(y_i-\widehat{y_i})= \tag{2.27}$$

$$=b_1\sum_{i=1}^{n}x_i\left[(y_i-\overline{y_i})-b_1(x_i-\overline{x})\right]=b_1\left[S_{xy}-b_1S_{xx}\right] \tag{2.28}$$

Tekintve, hogy $b_1=\frac{S_{xy}}{S_{xx}}$, így $b_1\left[S_{xy}-b_1S_{xx}\right]=0$. Ezzel beláttuk az eredeti egyenlőséget.

□

Másképp felírva az egyenlőség:

$$SS_{Total}=SS_{Reg}+SS_{Res} \tag{2.29}$$

A két komponens közül SS_{Reg} jelöli a regressziós négyzetösszeget, míg SS_{Res} a reziduális négyzetösszeget. Elemzéskor az ideális, ha SS_{Reg} jóval nagyobb, mint SS_{Res}.

2.3.2. Hipotézisvizsgálat a regresszió együtthatóira

Fontos megállapítanunk, hogy a magyarázó változó valóban szignifikánsan befolyásolja-e a függő változó értékét. Ezt hipotézisvizsgálattal tehetjük meg. Tekintsük az $y_i=\beta_0+\beta_1x_i+\varepsilon_i$ regressziós modellt, ahol $\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_n$ függetlenek, 0 a várható értékük és normális eloszlásúak. Ekkor vizsgáljuk a következő hipotézist:

$$H_0:\beta_1=0$$

$$H_1:\beta_1\neq 0$$

Amennyiben a nullhipotézis állna fenn, a modell redukálódna és $E(y)=\beta_0$ lenne igaz. Ez egyértelműen azt jelentené, hogy a magyarázó változó nem magyarázza az y változót. A vizsgálatot a lentebb található táblázat alapján elvégzett egyszerű F-próbával teszteljük.

Tekintve, hogy a $\frac{SS_{Reg}/1}{SS_{Res}/(n-2)}$ statisztika a nullhipotézis fennállása esetén $\frac{\chi_1^2/1}{\chi_{n-2}^2/(n-2)}$ eloszlást követ, és igaz a következő összefüggés:

$$\frac{SS_{Reg}/1}{SS_{Res}/(n-2)} = \frac{MS_{Reg}}{s^2} \tag{2.30}$$

így $\frac{MS_{Reg}}{s^2}$ statisztika $F_{1,n-2}$ eloszlást követ a nullhipotézis fennállása esetén és megfelelő lesz a próba teszt statisztikájának. Jelen esetben tekinthetünk az F-statisztikára úgy, mint a modell által megmagyarázott variancia hányadosa a hibák által magyarázott varianciával. Ebből következően, ha a statisztika értéke legalább akkora, mint az F-eloszlásból származó megfelelő kritikus érték, akkor a választott szignifikanciaszinten elvetjük a nullhipotézist.

Varianciaanalízis				
	Négyzetösszegek (SS)	Szabadsági fok (df)	Mean Square	F
Regresszió	SS_{Reg}	1	MS_{Reg}	$F = \frac{MS_{Reg}}{s^2}$
Reziduális	SS_{Res}	$n-2$	s^2	
Teljes	SS_{Total}	$n-1$		

2.4. Az illesztett modell minősége

Induljunk ki újra az eredeti $y = \beta_0 + \beta_1 x + \varepsilon$ regressziós modellből. Ekkor a determinációs együttható:

$$R^2 = \frac{SS_{Reg}}{SS_{Total}} = 1 - \frac{SS_{Res}}{SS_{Total}} \tag{2.31}$$

a korábbiakban felírt $SS_{Total} = SS_{Reg} + SS_{Res}$ egyenlőség miatt.

A felírásból könnyen észrevehető, hogy az együttható a regressziós modell által megmagyarázott variancia aránya a teljes varianciához képest. Ebből következően $0 < R^2 < 1$, illetve minél közelebb van 1-hez az értéke annál nagyobb részét magyarázza a modell az eredeti varianciának.

3. fejezet

Többszörös lineáris regresszió

Sokszor nem elég egyetlen magyarázó változó ahhoz, hogy megfelelő mértékben magyarázni tudjuk a függő változót. Ezért szükséges kiterjesztenünk az eddig taglalt módszert az egynél több magyarázó változót megengedő többszörös lineáris regressziós modellre.

3.1. Modell leírása

A magyarázó és függő változók értékeit a következőképp rendezhetjük táblázatba:

$$\begin{array}{ccccc} y & x_1 & x_2 & \dots & x_k \\ y_1 & x_{11} & x_{21} & \dots & x_{k1} \\ y_2 & x_{12} & x_{22} & \dots & x_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ y_n & x_{1n} & x_{2n} & \dots & x_{kn} \end{array}$$

Ezen táblázat sorai reprezentálják az összetartozó mérési eredményeket.

3.1.1. A modell leírása

Amennyiben a fenti táblázatban szereplő adatokban feltételezhető az egyes y_i értékek lineáris függése a hozzá tartozó x_i értékektől, úgy a modell felírható az

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i \quad (i = 1, 2, \ldots, n; n \geq k + 1) \tag{3.1}$$

egyenlettel.

A többszörös lineáris modell felírása mátrixos alakban:

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{3.2}$$

ahol

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{k1} \\ 1 & x_{12} & x_{22} & \dots & x_{k2} \\ \vdots & \vdots & & \vdots & \\ 1 & x_{1n} & x_{2n} & \dots & x_{kn} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{3.3}$$

Az ε_i hibák az egyszerű lineáris regresszióhoz hasonlóan függetlenek, 0 a várható értékük és egy konstans σ^2 szórásnégyzettel rendelkeznek, vagyis $\boldsymbol{\varepsilon}$ variancia-kovarianciamátrixa $Var(\boldsymbol{\varepsilon}) = \sigma^2 I_n$. Az x_{ij} értékek elhanyagolható hibával rendelkező vagy konstans adatok. Továbbá X oszlopai lineárisan függetlenek, azaz X rangja $k+1$.

3.1.1. Definíció. *Egy többdimenziós regressziós modellt akkor nevezünk lineárisnak, ha a magyarázó változók β együtthatóira nézve lineáris.*

Például az x-re nézve kvadratikus modellt lineárisnak nevezzük:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon \tag{3.4}$$

3.1.2. Linearizáló transzformációk

Gyakran előfordul, hogy nem teljesül minden feltétel a lineáris regressziós modell felépítéséhez. Sok esetben segít, ha különböző transzformációkat hajtunk végre a magyarázó, illetve a függő változókon.

A magyarázó változók transzformációjával lehetséges, hogy jelentősen javítható a függő változóval való lineáris kapcsolat. A feltételek további sérülésére a függő változó transzformációja is megoldás lehet (logaritmikus, exponenciális vagy valamilyen hatványtranszformáció).

Természetesen végezhetünk transzformációkat egyszerre a magyarázó és függő változókon is, például:

$$\ln \boldsymbol{y} = \beta_0 + \beta_1 \frac{1}{x_1} + \beta_2 e^{x_2} + \beta_3 \sqrt{x_3} + \boldsymbol{\varepsilon} \tag{3.5}$$

3.2. Regressziós együtthatók becslése

3.2.1. Legkisebb négyzetek módszere

Ebben az esetben a $(\boldsymbol{y} - X\boldsymbol{b})'(\boldsymbol{y} - X\boldsymbol{b})$ felel meg a reziduális négyzetösszegnek. Ennek keressük a minimumát $\boldsymbol{b}$-ben. A szélsőérték meghatározásához deriválást alkalmazunk. Tehát azt a $\boldsymbol{b}$ vektort keressük, amely kielégíti a következő egyenletet:

$$\frac{\delta}{\delta \boldsymbol{b}}\left[(\boldsymbol{y} - X\boldsymbol{b})'(\boldsymbol{y} - X\boldsymbol{b})\right] = 0 \tag{3.6}$$

Elvégezve a deriválást:

$$-2X'\boldsymbol{y} + 2\left(X'X\right)\boldsymbol{b} = 0 \tag{3.7}$$

Elemi átalakítások után adódik a következő egyenlet:

$$(X'X)\boldsymbol{b} = X'\boldsymbol{y} \tag{3.8}$$

A becsült paraméterek vektorát a normálegyenletek megoldásával kapjuk. Tekintve, hogy X teljes rangú, kifejezhetjük a keresett $\boldsymbol{b}$-t:

$$\boldsymbol{b} = (X'X)^{-1}X'\boldsymbol{y} \tag{3.9}$$

A becslőfüggvény várható értéke:

$$E(\boldsymbol{b}) = E((X'X)^{-1}X'\boldsymbol{y}) = E((X'X)^{-1}X'(X\boldsymbol{\beta} + \boldsymbol{\varepsilon})) = E(\boldsymbol{\beta} + (X'X)^{-1}X\boldsymbol{\varepsilon}) = \boldsymbol{\beta} \tag{3.10}$$

Illetve a variancia-kovariancia mátrixa:

$$D^2(\boldsymbol{b}) = E(\boldsymbol{b} - E\boldsymbol{b})(\boldsymbol{b} - E\boldsymbol{b})' = E(\boldsymbol{b} - \boldsymbol{\beta})(\boldsymbol{b} - \boldsymbol{\beta})' = E((X'X)X'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'X(X'X)^{-1}) =$$

$$= (X'X)X'E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')X(X'X)^{-1} = (X'X)X'\sigma^2 IX(X'X)^{-1} = \sigma^2(X'X)^{-1}$$

3.2.1. Tétel. *(Gauss-Markov) Többszörös lineáris regresszió esetén a fenti feltételek mellett a legkisebb négyzetekkel számolt* $\boldsymbol{b} = (X'X)^{-1}X'\boldsymbol{y}$ *a legjobb lineáris torzítatlan becslése* $\boldsymbol{\beta}$*-nak (BLUE, azaz Best Linear Unbiased Estimate). A "legjobb" a minimális szórás értelemben értendő, azaz bármely* $\boldsymbol{a} \in R^{k+1}$ *esetén*

$$D^2(\boldsymbol{a}'\boldsymbol{b}) \leq D^2(\boldsymbol{a}'\tilde{\boldsymbol{b}}),$$

ahol $\tilde{\boldsymbol{b}}$ egy tetszőleges lineáris torzítatlan becslése $\boldsymbol{\beta}$-nak.

Bizonyítás. Mivel $\tilde{\boldsymbol{b}}$ a $\boldsymbol{\beta}$ egy tetszőleges lineáris és torzítatlan becslése, így $\tilde{\boldsymbol{b}} = B\boldsymbol{y}$ alakú és $E(\tilde{\boldsymbol{b}}) = E(B\boldsymbol{y}) = BE(\boldsymbol{y}) = BX\boldsymbol{\beta} = \boldsymbol{\beta}$. Ez utóbbi csak akkor teljesül, ha $BX = I_{k+1}$. Legyen most $Q = B - (X'X)^{-1}X'$, vagyis $QX = BX - (X'X)^{-1}X'X = \mathbf{0}$. Ekkor

$$D^2(\tilde{\boldsymbol{b}}) = D^2(B\boldsymbol{y}) = BD^2(\boldsymbol{y})B' = ((X'X)^{-1}X' + Q)\sigma^2 I_n(Q + X(X'X)^{-1})$$

$$= \sigma^2((X'X)^{-1}X' + Q)(Q + X(X'X)^{-1}) = \sigma^2(X'X)^{-1} + \sigma^2 QQ',$$

azaz $D^2(\tilde{\boldsymbol{b}}) - D^2(\boldsymbol{b})$ mátrix pozitív szemidefinit. Ez utóbbi éppen azt jelenti, hogy bármely $\boldsymbol{a} \in R^{k+1}$ vektorra $D^2(\boldsymbol{a}'\tilde{\boldsymbol{b}}) - D^2(\boldsymbol{a}'\boldsymbol{b}) \geq 0$. $\square$

3.2.2. Hibabecslés

Az egyszerű lineáris regresszióhoz hasonlóan itt is szükség van a hibatagok σ^2 varianciájának becslésére. A regressziós négyzetösszeg a k darab magyarázó változó által magyarázott varianciát határozza meg, így a teljes $(n-1)$ szabadsági fok felbontása így néz ki:

$$n - 1 = k + (n - k - 1) \tag{3.11}$$

Tehát a megfelelő s^2 torzítatlan becsléshez a reziduális négyzetösszeget a maradék $(n-k-1)$ szabadsági fokkal való osztással jutunk, ami a következőképp írható fel:

$$s^2 = \frac{\|y - Xb\|}{n-k-1} = \frac{(\boldsymbol{y} - X\boldsymbol{b})'(\boldsymbol{y} - X\boldsymbol{b})}{n-k-1} = \sum_{i=1}^{n} \frac{(y_i - \widehat{y}_i)^2}{n-k-1} \tag{3.12}$$

ahol k a magyarázó változók száma, n pedig a mérési adatok száma.

3.2.3. Maximum likelihood becslés

Az egyszerű lineáris modellhez hasonlóan itt is becsülhetjük a paramétereket maximum-likelihood becsléssel, ha feltesszük, hogy a hibák normális eloszlást követnek. A likelihood-függvény maximalizálásával belátható, hogy a paraméterek becslése ekkor megegyezik a legkisebb négyzetek módszerével kapott becsléssel, azonban a varianciára torzított becslést kapunk:

$$\widehat{\sigma}^2 = \frac{1}{n}||\boldsymbol{y} - X\boldsymbol{b}||^2 = (n-k-1)\frac{s^2}{n} \tag{3.13}$$

3.2.4. A regressziós együtthatók értelmezése

A modell paramétereinek az elemzésnél hasznos algebrai jelentése is van. Amennyiben az x_j magyarázó változót növeljük egy adott Δx_j-vel a többi változó értékének fixen tartása mellett, akkor az $\boldsymbol{y}$ függő változó változása éppen az x_j-hez tartozó b_j regressziós együttható és Δx_j szorzataként számolható ki:

$$\Delta \boldsymbol{y} = b_j \Delta x_j \tag{3.14}$$

Következésképp a b_j becsült paraméter az x_j egységnyi változásának hatását fejezi ki az $\boldsymbol{y}$ függő változóra, a többi változó értékének változatlansága mellett.

3.3. Szóródás és hipotézisvizsgálat

3.3.1. Teljes szóródás felbontása

Ebben az esetben is felírható a teljes négyzetösszeg felbontása a regressziós és reziduális négyzetösszeg összegére:

$$\sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (\widehat{y_i} - \overline{y})^2 + \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 \tag{3.15}$$

ahol az $\widehat{y_i}$ értékeket a $\widehat{y_i} = b_0 + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_n x_{ki}$ regressziós modellből kapjuk.

Másképp:

$$SS_{Total} = SS_{Reg} + SS_{Res} \tag{3.16}$$

3.3.2. Regressziós együtthatók hipotézisvizsgálata

A modellépítés elsődleges funkciója annak meghatározása, hogy a magyarázó változók milyen mértékben befolyásolják a függő változót, azaz mely magyarázó változók felelősek a függő változó szignifikáns szóródásában. Tegyük fel, hogy β_m a kérdéses paraméter. Ekkor a következő hipotézist szeretnénk tesztelni:

$$H_0 : \beta_m = 0$$

$$H_1 : \beta_m \neq 0$$

A nullhipotézis ellenőrzésére használhatunk t-próbát. A β_m paraméter tesztelésének próbastatisztikája:

$$t = \frac{\widehat{\beta_m}}{s_{\widehat{\beta_m}}} \tag{3.17}$$

ahol $s_{\widehat{\beta_m}}$ a β_m paraméter becslésének szórása. A nullhipotézis teljesülése esetén a próbastatiszika megközelítőleg $n-2$ szabadsági fokú t-eloszlást követ. Ha a kiszámított t próbastatisztika abszolút értéke nagyobb, mint a t-eloszlásból származó kritikus érték, akkor elutasítjuk a nullhipotézist, vagyis mondhatjuk, hogy β_m nem 0, tehát az x_m változó fontos a függő változó magyarázásában.

Az együtthatók globális tesztelésére F-próbát használhatunk. Ekkor a hipotézisek a következőek:

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

$$H_1 : \exists \beta_i \neq 0 \quad i = 1 \ldots n$$

A nullhipotézis fennállása esetén a próbastatisztika F eloszlást követ. A próbastatisztika:

$$F = \frac{\frac{SS_{Reg}}{k}}{\frac{SS_{Res}}{n-k-1}} \tag{3.18}$$

Amennyiben fennáll, hogy $F > F_{k-1,n-k}$, akkor a nullhipotézist elutasítjuk, azaz a modell szignifikánsan jobb a csak konstans tagot tartalmazó modellnél.

3.3.3. Konfidenciaintervallum illesztés az együtthatókra

A β paramétereket eddig egyetlen statisztikával becsültük, ami nem elég informatív, ugyanis nem tudni mennyi bizonytalanság van a becslésben. Így minden β paramétert az adatoktól függő intervallum belsejébe akarjuk szorítani egy előírt valószínűséggel, azaz konfidenciaintervallumot számítunk. A β_m elméleti paraméter konfidenciaintervalluma α szignifikanciaszinten:

$$\widehat{\beta_m} \pm t_{n-k-1,1-\frac{\alpha}{2}} s_{\widehat{\beta_m}} \tag{3.19}$$

$E(\boldsymbol{y}|\boldsymbol{x_0})$ feltételes várható értékre is számíthatunk konfidenciaintervallumot. Ha egy adott $\boldsymbol{x} = \boldsymbol{x_0}$ pontban keressük a függvényértéket, akkor $\widehat{\boldsymbol{y_0}} = \boldsymbol{x_0'b}$ torzítatlan becs-

lést ad a megfelelő pontban vett függvényértékre. A becsléshez tartozó variancia pedig a következőképp számolható ki:

$$Var(\widehat{\boldsymbol{y}}|\boldsymbol{x} = \boldsymbol{x_0}) = s^2 \boldsymbol{x_0'} \left(X'X\right)^{-1} \boldsymbol{x_0} \tag{3.20}$$

A függő változó $\boldsymbol{x_0}$-ra vonatkozó várható értékének konfidenciaintervallumát α szignifikanciaszinten az alábbi módon számolhatjuk ki:

$$\boldsymbol{x_0'b} \pm t_{n-k-1,1-\frac{\alpha}{2}} s \sqrt{\boldsymbol{x_0'} \left(X'X\right)^{-1} \boldsymbol{x_0}} \tag{3.21}$$

Azonban előfordulhat, hogy nem a függő változó $\boldsymbol{x_0}$-hoz tartozó várható értékére szeretnénk konfidenciaintervallumot, hanem egy $\boldsymbol{x_0}$-hoz tartozó új megfigyelésre:

$$\boldsymbol{x_0'b} \pm t_{n-k-1,1-\frac{\alpha}{2}} s \sqrt{1 + \boldsymbol{x_0'} \left(X'X\right)^{-1} \boldsymbol{x_0}} \tag{3.22}$$

Ezt nevezzük előrejelzési intervallumnak.

3.4. Az illesztett modell minősége

A determinációs együtthatónak (R^2) ugyanaz a jelentése, mint az egyszerű lineáris regresszióban. Azt mutatja meg, hogy a modellben figyelembe vett magyarázó változók mekkora részét magyarázzák meg az $\boldsymbol{y}$ változó varianciájának:

$$R^2 = \sum_{i=1}^{n} \frac{\left(\widehat{y}_i - \overline{\boldsymbol{y}}\right)^2}{\left(y_i - \overline{\boldsymbol{y}}\right)^2} = \frac{SS_{Reg}}{SS_{Total}} \tag{3.23}$$

Új változó hozzáadása a modellhez a legtöbbször növeli (néha változatlanul hagyja) ezt a determinációs együtthatót, ami túl sok változó bevonásakor nem a megfelelő információt hordozza magában a megfelelő modell kiválasztásához. Ezért van szükség a korrigált determinációs együtthatóra, amely figyelembe veszi a modellbe választott változók számát.

A korrigált determinációs együttható:

$$R^2_{adj} = 1 - \frac{\frac{SS_{Res}}{n-k-1}}{\frac{SS_{Total}}{n-1}} = 1 - \frac{n-1}{n-k-1} \frac{SS_{Res}}{SS_{Total}} = 1 - \frac{n-1}{n-k-1} \left(1 - R^2\right) \tag{3.24}$$

A korrigált determinációs együttható csökkenhet is, ha a reziduális négyetösszeg csökkenését meghaladja a szabadságfok módosulása a nevezőben. Általában ezt az együtthatót vesszük figyelembe modellválasztáskor.

3.5. Feltételek

3.5.1. Normalitás

A modell leírásnál feltettük, hogy a hibatagok normális eloszlásúak. Ennek tesztelésére több módszer is használható. Az egyik legelterjedtebb grafikus módszer a Q-Q plot(kvantilis-kvantilis ábra), amely a legtöbb statisztikai programcsomagban is megtalálható (pl. R). Ez a módszer az $x_1, \ldots, x_n$ minta tapasztalati kvantiliseit veti össze a standard normális eloszlás kvantiliseivel. Először sorba rendezi a mintát ($x_1^* \leq \ldots \leq x_n^*$), így megkapjuk a tapasztalati kvantiliseket. Az elméleti kvantilis értékek a standard normális eloszlás kvantilisei az $\frac{i}{n+1} \quad (i = 1, \ldots, n)$ pontokban. Ezeket felhasználva ábrázoljuk az $(\Phi^{-1}\left(\frac{i}{n+1}\right), x_i^*)$ pontokat, ahol Φ a normális eloszlás eloszlásfüggvénye. Ha a minta normális eloszlást követ, a pontok megközelítőleg lineárisan helyezkednek el. Egyéb módszerekkel is tesztelhető a normalitás (pl. Shapiro-Wilk teszt). Amennyiben nem teljesül, alkalmazhatunk bootstrap eljárást.

Bootstrap

A bootstrap egy újramintavételezési eljárás, mely többek közt a regressziós paraméterek becslései szórásának vizsgálatára, illetve a modell illeszkedésének ellenőrzésére használható, például amikor a modell eloszlására vonatkozó feltétel nem teljesül. Három főbb bootstrap eljárással találkozhatunk: 1) paraméteres bootstrap, amikor az illesztett modellből szimulálunk bootstrap mintákat, 2) reziduális újramintavételezés (*residual resampling* vagy *fixed X*) eljárás, ahol az illesztett regressziós modell reziduálisaiból veszünk mintát visszatevéssel, majd ezt adjuk hozzá az illesztett értékhez, 3) eset mintavételezés (*case-resampling* ill. *random X*) eljárás, ahol az eredeti adatokból visszatevéssel generálunk bootstrap mintákat (vagyis nem használjuk fel a regressziós illesztést). A paraméteres bootstrap esetében támaszkodunk arra, hogy korrekt regressziós modellt használunk, megfelelő hibaeloszlással, míg a reziduális újramintavételezésnél nem teszünk fel hibákra vonatkozó eloszlást. Az eset mintavételezés a legáltalánosabb (az egyik előbbi feltételt sem teszi fel), így egyben a "legbiztonságosabb" eljárás. Mi a modell megbízhatóságát vizsgáló reziduális újramintavételezési eljárással fogunk foglalkozni. Először írjuk fel a becsült regressziós egyenletet a következőképpen:

$$\boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \hat{\boldsymbol{y}} + \boldsymbol{\varepsilon},$$

ahol $\boldsymbol{y}$ az illesztett értékek, és $\boldsymbol{\varepsilon}$ pedig a reziduálisok.
Ekkor az $(\boldsymbol{y}^*, \boldsymbol{x}^*)$ bootsrap mintát a következőképpen kapjuk:

(1) legyen $\boldsymbol{x}^* = \boldsymbol{x}$,

(2) visszatevéssel vegyünk mintát az $\boldsymbol{\varepsilon}$ residuálisokból: $\boldsymbol{\varepsilon}^*$ és adjuk hozzá az illesztett $\hat{\boldsymbol{y}}$ értékhez: $\boldsymbol{y}^* = \hat{\boldsymbol{y}} + \boldsymbol{\varepsilon}^*$

Az eljárást R-szer megismételjük. Ennek az eljárásnak egy szintén használatos változata, amikor a reziduálisok egy transzformáltjából mintavételezünk:

$$\varepsilon_i^* = \frac{\varepsilon_i}{\sqrt{1 - h_{ii}}} - \overline{\varepsilon}$$

ahol h_{ii} az i-edik "hat" érték vagy leverage ($h_{ii} = [X(X'X)^{-1}X']_{ii}$). Az R program *Boot* függvénye (*car* csomagban) szintén ezt az eljárást alkalmazza (method="residual").

3.5.2. Multikollinearitás

A többszörös lineáris regresszió egyik feltétele szintén, hogy a magyarázó változók egymástól lineárisan függetlenek legyenek. Multikollinearitás akkor fordul elő, amikor lineáris ill. közel lineáris kapcsolat van a magyarázó változók közt. Tökéletes lineáris kapcsolat esetén az X mátrix nem teljes rangú, így $X'X$ nem invertálható, ami pedig szükséges lenne a β paraméterek legkisebb négyzetekkel történő becsléséhez. Valós adatokra inkább csak közel lineáris kapcsolat, mint tökéletes lineáris kapcsolat jellemző. Közel lineáris kapcsolat esetén ugyan $X'X$ invertálható, de a becslés standard hibái megnőnek, a becslések bizonytalanná válnak. Multikollinearitás detektálására több módszer alkalmazható, ezek közül a variancia inflációs együtthatót (variance inflation factor) fogjuk bemutatni.

A VIF annak a mértéke, hogy a multikollinearitás jelenléte milyen mértékben növeli a becsült paraméterek varianciáját.

Először számoljuk ki a magyarázó változók közötti determinációs együtthatót az esetben, mikor a függő változó X_j és a magyarázó változók a többi $X_i (j \neq i)$ változók, jelölje ezt R_j^2. Ez meghatározza, hogy a multikollinearitás jelenléte milyen mértékben növeli a becsült paraméterek varianciáját. A j. változó VIF értéke:

$$VIF_j = \frac{1}{1 - R_j^2} \tag{3.25}$$

Ennek határai: $1 \leq VIF_j < \infty$

Ha a j. változó lineárisan független a többitől, akkor a VIF értéke 1. Minél nagyobb a multikollinearitás, a VIF értéke annál nagyobb. Értékét minden változóra kiszámítjuk.

Multikollinearitás kiküszöbölésére a legegyszerűbb eljárás, hogy elhagyunk néhány magyarázó változót, míg szofisztikáltabb módszerek közé beletartozik a Ridge-regresszió ill. főkomponens regresszió alkalmazása.

3.5.3. Heteroszkedaszticitás

A heteroszkedaszticitás a modellbeli szórások különbözőségére utal, vagyis amikor a regressziós modellben a hibatagok szórása nem egyezik meg. Fentebb feltettük, hogy a hibatagok függetlenek és a variancia-kovariancia mátrixuk $Var(\boldsymbol{\varepsilon}) = \sigma^2 I_n$ alakú. Feltéve, hogy a szórások nem korreláltak, de megfigyelésenként változnak, a mátrix általánosabb alakja $Diag[\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2]$.

A reziduálisokat az illesztett értékek ellenében ábrázolva a heteroszkedaszticitás grafikusan is látható: amennyiben az illesztett értékek növekedésével a hibák szórása is növekszik. Egyik legnépszerűbb módszer heteroszkedaszticitás kiküszöbölésére a súlyozott legkisebb négyzetek módszerének alkalmazása.

3.5.4. Autokorreláció

Időtől függő adatok esetén gyakran nem feltételezhetjük, hogy a hibák korrelálatlanok. Inkább az feltételezhető, hogy a hibák sorozatosan korreláltak, más szóval $E(\varepsilon_i \cdot \varepsilon_j) \neq 0$, ekkor azt mondjuk, hogy a hibák autokorreláltak. Tehát az autokorreláció egy adatsor különböző (pl. időben eltolt) megfigyelései értékei közti kapcsolatot fejezi ki. Az időbeli elsőrendű autokorreláció azt jelenti, hogy minden i. időponthoz tartozó adat korrelált az előző időpontnál $(i-1)$ felvett értékével. Rendszerint nagyszámú, azonos előjelű reziduálisok csoportosulása a hibák pozitív autokorrelációjára, míg az előjelek gyors váltakozása a hibák negatív autokorrelációjára utal. Tegyük fel most, hogy egyenlő időközönként

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t,$$

ahol ρ az autokorreláció és u_t egy véletlen $N(0, \sigma_u)$ eloszlású zaj. Ekkor az előbbi hibastruktúrát elsőrendű autoregresszív hibastruktúrának mondjuk. Gyakran a reziduális ábrákból látható az autokorreláció, de többféle statisztikai módszer is ismert ennek detektálására. Ezek közül a legismertebb a pozitív autokorreláció detektálására alkalmazható

Durbin-Watson teszt, amelynek a null- és ellenhipotézise a következő:

$$H_0: \ \rho = 0$$

$$H_1: \ \rho > 0$$

Pozitív autokorreláció esetén azt várnánk, hogy a szomszédos reziduálisok numerikusan hasonló értékűek, így természetes, hogy a próbastatisztika a szomszédos reziduálisok különbségén alapuljon. A Durbin-Watson próbastatisztika:

$$d = \frac{\sum_{t=2}^{n}(\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{i=1}^{n}\hat{\varepsilon}_t^2}$$

melynek kis értékeire ($d < d_L$) elutasítjuk, míg nagy értékeire ($d > d_U$) nem utasítjuk el H_0-t (d_L és d_U a d eloszlás megfelelő értékei). A d próbastatisztika egyéb értékeire ($d_l < d < d_u$) a teszt inkonkluzív.

4. fejezet

Logisztikus regresszió

4.1. Modell leírása

A logisztikus regresszió egy klasszifikációs eljárás, amelyben a magyarázó változó(k)ból nyert információ alapján a megfigyeléseket előre definiált, egymást kölcsönösen kizáró csoportok egyikébe soroljuk be. Kétféle kimenetelű függő változó - például lesz-e villámkisülés egy adott napon vagy sem - esetén annak változékonyságát szeretnénk leírni a magyarázó változókkal. Logisztikus regresszió esetén a függő változó diszkrét: lehet bináris (kétféle kimenetű) vagy polichotom (többféle kimenetű). A továbbiakban csak a bináris esettel foglalkozunk. A kimenetet 0-val, illetve 1-gyel kódoljuk. Annak a valószínűségét szeretnénk magyarázni, illetve prediktálni a magyarázó változókkal, hogy 1 lesz a kimenet. A valószínűségre közvetlenül nincs adatunk, becsülni sem tudjuk, hiszen nyilván értelmetlen, hogy egy adott villám milyen gyakran sült ki. Míg a közönséges regresszió esetén a választ a magyarázó változók ellenében megjelenítve a kapcsolat jellegéről képet kaphatunk, bináris válasz esetén a válasz értékei helyett a

$$E(Y|X = x_i) = P(Y = 1|X = x_i) = \pi(x_i)$$

függvényt vizsgáljuk. Vagyis az X ismeretében mennyi a feltételes valószínűsége a vizsgált esemény bekövetkezésének az i-edik megfigyelés esetén ($i = 1, \ldots, n$). Azonban ez a függvény nem lineáris, inkább S alakú görbe, így egy szigmoid függvénnyel (melynek értékei 0 és 1 közt vannak) modellezhetjük a feltételes valószínűség és a magyarázó változók kapcsolatát. Most csak egy magyarázó változóra írjuk fel a képleteket, de több magyarázó változó esetén is hasonló, amelyre a későbbiekben visszatérünk. Tehát a bekövetkezés valószínűsége:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}} \tag{4.1}$$

Majd ún. logit transzformációval azt kapjuk, hogy

$$g(x) = logit(\pi(x)) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x \tag{4.2}$$

vagyis a $g(x)$ függvény lineáris a paraméterekben és értékeit $-\infty$ és ∞ közt veheti fel. Ezeket a β együtthatókat kell meghatároznunk. Mivel a $\pi(x)$-et nem ismerjük, így a legkisebb négyzetes módszer helyett a maximum likelihood módszert alkalmazhatjuk. Először írjuk fel az $Y_1, \dots Y_n$ minta sűrűségfüggvényét (jelölje $\boldsymbol{\beta} = (\beta_0, \beta_1)$):

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i} \tag{4.3}$$

melynek logaritmusa

$$l(\boldsymbol{\beta}) = \ln L(\boldsymbol{\beta}) = \ln\left(\prod_{i=1}^{n} \pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i}\right) = \sum_{i=1}^{n} y_i \ln(\pi(x_i)) + \sum_{i=1}^{n}(1-y_i)\ln(1-\pi(x_i)) =$$

$$= \sum_{i=1}^{n} y_i \ln\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) - \sum_{i=1}^{n} \ln\left(1+e^{\beta_0+\beta_1 x_i}\right) = \sum_{i=1}^{n} y_i(\beta_0+\beta_1 x_i) - \sum_{i=1}^{n} \ln\left(1+e^{\beta_0+\beta_1 x_i}\right)$$

amit maximalizálni kell a paraméterekben. Ehhez parciálisan deriváljunk β_0 és β_1 szerint és a deriváltakat tegyük 0-val egyenlővé:

$$\sum_{i=1}^{n}(y_i - \pi(x_i)) = 0 \quad \text{és} \quad \sum_{i=1}^{n} x_i(y_i - \pi(x_i)) = 0,$$

Ez azonban nem lineáris egyenletrendszer, így csak iteratív megoldása létezik (feltéve, hogy konvergens), melyet számítógépes programmal (pl. R-rel) számolhatunk ki.

Legyen most $\boldsymbol{x}' = (x_1, \dots, x_p)$ a p magyarázó változó, és jelölje $\boldsymbol{\beta}' = (\beta_0, \beta_1, \dots, \beta_p)$ a paraméter vektort. Ekkor

$$\pi(\boldsymbol{x}) = \frac{e^{\beta_0+\beta_1 x_1+\dots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\dots+\beta_p x_p}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\dots+\beta_p x_p)}} = \frac{1}{1+e^{-\boldsymbol{x}'\boldsymbol{\beta}}}$$

és

$$g(\boldsymbol{x}) = logit(\pi(\boldsymbol{x})) = \ln\left(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p.$$

Tegyük most fel, hogy a minta n független $(y_i, x_{1i}, \ldots, x_{pi}) = (y_i, \boldsymbol{x_i})$ $(i = 1, \ldots, n)$ megfigyelésből áll, ekkor hasonlóan a fentihez, a likelihood függvény deriválásával $p+1$ likelihood egyenletet kapunk

$$\sum_{i=1}^{n}(y_i - \pi(\boldsymbol{x_i})) = 0 \quad \text{és} \quad \sum_{i=1}^{n} x_{ji}(y_i - \pi(\boldsymbol{x_i})) = 0, \quad j = 1, \ldots, p$$

amely nem lineáris egyenletrendszer iteratív megoldása adja a keresett paraméterbecsléseket, $\hat{\boldsymbol{\beta}}$-et. Ennek a felhasználásával pedig a $\pi(\boldsymbol{x})$-re kapunk becslést.

Előfordulhat, hogy az egyik X magyarázó változó diszkrét, azaz r különböző értéket vehet fel: $a_1, \ldots, a_r$. Ekkor $r-1$ dummy változót kell bevezetnünk, $D_k = \begin{cases} 1 & a_k \text{ esetén} \\ 0 & \text{különben} \end{cases}$ $(k = 1, \ldots, r-1)$, hogy megkülönböztessük az r kategóriát.

4.2. Modell értelmezése

A fenti képlet alapján következőképp definiáljuk az 1 *oddsz*-át, azaz az 1 és 0 bekövetkezési valószínűségének arányát feltételesen a magyarázó változóktól:

$$oddsz = \frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})} = e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}. \tag{4.4}$$

Az *oddsz* mindig 0-nál nagyobb értéket vesz fel és 1-nél nagyobb értéke azt jelenti, hogy az 1 bekövetkezésének a valószínűsége nagyobb, mint a 0 bekövetkezésének a valószínűsége. Például a fent említett példában $oddsz = 3$ azt jelenti, hogy 3-szor olyan valószínű, hogy villámlani fog, mint az, hogy nem fog villámlani. Az *oddsz* logaritmusa:

$$log(oddsz) = logit(\pi(\boldsymbol{x})) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p. \tag{4.5}$$

Pozitív β_j növeli, míg negatív β_j csökkenti a logit, így úgyszintén az *oddsz* értékét. A β_j becsült paraméter az x_j változó egy egységnyi változásának a logitra gyakorolt hatását mutatja, míg az e^{β_j} viszont az x_j egy egységnyi abszolút növekedésének hatása az *oddsz*-ra.

4.3. Az illesztett modell minősége

A logisztikus regressziós modell alkalmazásakor az egyik legnehezebb feladat a magyarázó változók kiválasztása a modellbe, különösen akkor, ha a változók száma meglehetősen nagy. A modellnek egyrészt elég bonyolultnak kell lennie ahhoz, hogy illeszkedjen az adatokhoz, másrészt könnyen értelmezhetőnek is kell lennie. Fontos, hogy a modellt ne csak matematikai módszerek alapján válasszuk ki, hanem az aktuális jelenség vagy kísérlet értelmezése is szerepet játsszon.

A célunk az adatokra legjobban illeszkedő modell megtalálása minél kevesebb paraméter felhasználásával. Először a fontosnak ígérkező változókat tartalmazó egyszerű modellt illesztünk, például stepwise eljárással, majd a kapott modellt a likelihood-hányados próba segítségével összehasonlítjuk a teljes modellel.

4.3.1. Paraméterek szignifikanciája

Likelihood-hányados χ^2

A többváltozós lineáris regressziónál használatos globális F-próba mintájára itt likelihood-hányados próbát alkalmazhatunk a paraméterek együttes vizsgálatára. A vizsgálandó hipotézis:

$$H_0 : \beta_1 = \cdots = \beta_p = 0$$

$$H_1 : \text{legalább az egyik } \beta_j \neq 0$$

Ekkor a likelihood-hányados statisztika

$$G = -2(l(\hat{\boldsymbol{\beta}}^0) - l(\hat{\boldsymbol{\beta}})) \tag{4.6}$$

ahol $l(\hat{\boldsymbol{\beta}})$ a loglikelihood függvénye a teljes modellnek, míg $l(\hat{\boldsymbol{\beta}}^0)$ a loglikelihood függvénye a csak konstanst tartalmazó modellnek. Bizonyos regularitási feltételek mellett nagy n mintanagyság esetén a likelihood-hányados statisztika p szabadsági fokú χ^2 eloszlással közelíthető. Gyakran "Residual deviance" jelöli a $-2l(\hat{\boldsymbol{\beta}})$-et, melyre úgy is lehet tekinteni, mint annak a mértékére, hogy mennyire magyarázza az adatokat a változókat tartalmazó modell. (Hasonlóan "Null deviance" jelöli a $-2l(\hat{\boldsymbol{\beta}}^0)$-t.)

Megjegyezzük, hogy az előbbi tesztet a $\boldsymbol{\beta}$ paraméterek bármely részhalmazának 0-val való egyenlőségének tesztelésére is alkalmazhatjuk, azzal a változtatással, hogy $l(\hat{\boldsymbol{\beta}}^0)$ a

loglikelihood függvénye a leszűkített modellnek, a próbastatisztika szabadsági foka pedig $p - r$.

Wald-statisztika

Programcsomagok szintén megadják a paraméterekre vonatkozó Wald-statisztikát is. Ez azt teszteli, hogy egy adott β paraméter egyenlő-e 0-val:

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

A próbastatisztika

$$W_j = \frac{\hat{\beta}_j}{\widehat{SE}(\hat{\beta}_j)} \tag{4.7}$$

a nullhipotézis teljesülése esetén megközelítőleg standard normális eloszlású. Ugyan a Wald-statisztikát könnyű kiszámolni, a megbízhatósága kérdéses, különösen kis mintanagyság esetén. Nagy paraméter becslés esetén a szórás gyakran túlbecsült, a Wald-statisztika kisebb lesz, így kevésbé utasítjuk el a nullhipotézist, pedig a változó lehet fontos. Így általában a likelihood-hányados próba javasolt.

4.3.2. Akaike Információs Kritérium (AIC)

Különböző paraméterű modellek összehasonlítására használatos. A modell minőségét mérő érték, hasonlóan az adjusted R^2-hez. (Csak ugyanazon az adathalmazon futtatott logisztikus regresszió összehasonlítására használható.) Akaike információs kritérium a következőképpen definiáljuk:

$$AIC = -2(\ln(\hat{L}) - k) \tag{4.8}$$

ahol k a modellben szereplő paraméterek száma és $\hat{L}$ a loglikelihood függvény becslése a minta alapján. Több magyarázó változó esetén növekszik, viszont nagyobb mintanagyság esetén pedig csökken az értéke, így alacsonyabb AIC érték jobb illeszkedésre utal. Ezt a kritériumot használhatjuk a stepwise szelekció alkalmazásakor is.

4.3.3. Változók kiválasztása: stepwise szelekció

A stepwise eljárások lényege, hogy egyesével vizsgáljuk a magyarázó változókat és döntjük el, hogy tartalmazza-e az adott változót a modell vagy sem, vagyis megállapíthatjuk, hogy mely magyarázó változók fontosak a függő változó leírásában. Bármely stepwise eljárásban a változók kiválasztása illetőleg elhagyása a modellből likelihood-hányados χ^2 segítségével történik. Hátrafelé (backwards) és előrefelé (forwards) eljárásokat különböztetünk meg:

Backwards szelekció: Teljes modellel kezd, a nem szignifikáns változókat likelihood-hányados χ^2 segítségével kiválogatva megkeresi a matematikailag legjobban illeszkedő modellt.

Forwards szelekció: Az üres modellből kiindulva egyesével hozzáadva a likelihood-hányados χ^2 kritériummal szignifikáns változókat, építi fel a modellt.

A két módszer nem biztos, hogy ugyanazokat a változókat adja, bár a legszignifikánsabb változókban egyezés lesz. A modellt érdemes a jelenség vagy kísérlet háttere alapján választani, így tartalmazhat még olyan változót ill. változókat, amelyet a kísérletező fontosnak tart a függő változó megmagyarázásában, de matematikailag nem feltétlen jön ki szignifikánsnak.

4.4. Feltételek

4.4.1. Linearitás

Feltétel, hogy a magyarázó változók lineárisak legyenek a magyarázó változó logit transzformáltjának függvényében. Ennek grafikus vizsgálata mellett gyakran használatos a Hosmer & Lemeshow [2] által javasolt ún. Box-Tidwell módszert, melynek lényege, hogy a függő változón egy olyan logisztikus regressziót futtatunk, melyben minden folytonos magyarázó változó és annak a magyarázó változó logaritmusával való kölcsönhatása szerepel (természetesen ez nempozitív értékekre nem használható). Amennyiben egy változó is szignifikáns, akkor a linearitás feltétel nem teljesül. Nemlinearitás esetén a magyarázó változók transzformáltja, illetve hatványra emelése segíthet, továbbá dummy változók használata is tekintetbe vehető.

4.4.2. Multikollinearitás

A logisztikus regresszió érzékeny a lineáris függőséget mutató magyarázó változókra. Ilyen változók esetén a szórások magasak lehetnek, így a paraméterbecslések nem megbízhatóak. A kollinearitást a regressziónál említett VIF módszerrel lehet ellenőrizni.

4.4.3. Mintanagyság

Amennyiben túl kevés eset van a magyarázó változók számához viszonyítva, a paraméterbecslések magasak lehetnek nagy szórással, így lehetséges, hogy a modell nem konvergál. Ökölszabályként legyen legalább $5 - 10$-szer annyi esemény a legkisebb csoportban, mint a vizsgált magyarázó változók száma.

A logisztikus regresszió előnye viszont, hogy a magyarázó változók eloszlására nincs feltétel.

5. fejezet

Alkalmazás a meteorológiában

5.1. Feladat leírása

A meteorológiában nagyon fontos probléma a zivatarok előrejelzése, mert ezek olyan tényezőkkel (jégeső, villámkisülések, hatalmas széllökések) járnak együtt, amelyek nagy károkat képesek okozni a környezetünkben. Tudjuk, hogy a villámcsapások tüzeket okozhatnak, károkat okozhatnak az elektromos hálózatban, befolyásolhatják a légi közlekedést, illetve az emberi életre is veszélyesek lehetnek. Dolgozatomban a zivatarokkal együtt járó villámkisülések számának, illetve a valószínűségük előrejelzésével foglalkoztam.

A villámkisülések számáról a LINET-rendszerből kapott adatok alapján volt információm. Ez egy Európa-szerte használt villám-detektáló rendszer, melynek lényege, hogy különböző helyeken telepített megfigyelő antennák regisztrálják a villámlás által keltett elektromágneses-hullámok beérkezésének idejét. Egy állomás képes meghatározni a villám idejét. Ha több antenna együtt dolgozik, akkor az ismert villámidőkből képesek a helyet is megállapítani. [10]

A zivatarok keletkezése (amelyekkel együtt járnak a villámkisülések) rendkívül összetett légköri folyamatok, éppen ezért nehéz előrejelezni őket. A villámok előrejelzéséhez mind a mai napig jellemzően inkább csak közvetett paraméterek állnak rendelkezésre. A legutóbbi pár évben jelentek meg különféle diagnosztikus eljárások a villámtevékenység előrejelzéséhez, de a legtöbb meteorológiai szolgálatnál, így az OMSZ-nál sem alkalmaznak még ilyet a napi gyakorlatban. [8][14]

Ahhoz, hogy zivatar alakuljon ki egy adott területen alapvetően 3 feltétel szükséges: a légkör labilitása, elegendő megfelelő nedvesség a teljes légoszlopban, illetve valamilyen

emelő/trigger hatás. Labilitásról akkor beszélünk, amikor a függőleges légmozgást, légcserét elősegítő légköri állapot áll fent. Ezeket a függőleges feláromlásokkal járó légköri jelenségeket nevezzük konvekciónak. Megfelelő nedvesség alatt azt a vízgőztartalmat értjük, ami a kezdeti feláramlások során szükséges a levegő telítődésének, vagyis a felhőképződés kialakulásához. Ezeket a hatásokat jellemző változókkal dolgoztam. A munkát egy meteorológus segítségével végeztem, aki az Országos Meteorológiai Szolgálat munkatársa. Ő segített nekem a kérdéses adatok kivizsgálásánál, a megfelelő változók kiválasztásánál, illetve a meteorológiai fogalmak megértésében.

A villámlokalizációs mérési adatok mellett az Országos Meteorológiai Szolgálat rendelkezésemre bocsátott - fentebb már említett - légköri állapotra vonatkozó paramétert 2012-től 2018-ig. Minden év május 1-től augusztus 30-ig tartó időintervallumból kaptam adatokat, mert az év ezen periódusában jóval gyakoribbak a zivatarok, mint a téli félévben, illetve el is térhetnek a statisztikai összefüggések a zivatar valószínűsége és a vizsgált paraméterek értékei között. Ezek olyan értékek, amelyekről feltételezzük, hogy szoros kapcsolat van köztük és a villámkisülések darabszáma között. Ezen adatokat felhasználva alkalmaztam a regresszió módszerét a villámcsapások számának előrejelzésére.

Egyrészt többszörös lineáris regresszió módszerével becsültem az adott napi villámkisülések számát. Ennek célja a zivatarok intenzitásának előrejelzése. Minél több villámkisülés várható, annál nagyobb intenzitású zivatart valószínűsíthetünk az adott területen. A fentebb taglalt zivatarok létrejöttéhez szükséges légköri jelenségeket jellemző magyarázó változókat elemeztem oly módon, hogy a legmegfelelőbb modellt válasszam ki a villámkisülések számát jellemző függő változó értékének becslésére. Ehhez elemeztem a változókat, egymással való kapcsolatukat, detektáltam az esetleges hibás mérési eredményeket, illetve ezeket egyeztettem a meteorológussal.

Másrészt logisztikus regresszióval becsültem a villámkisülések létrejöttének valószínűségét. Ennek célja az volt, hogy megbecsüljem annak valószínűségét, hogy egyáltalán várható-e legalább 1 villámkisülés Magyarország területén az adott napon. Ez azért fontos, mert a veszélyjelző szolgálatok (pl. OMSZ) többletfeladatot kell ellássanak, amennyiben zivatar várható a térségben.

5.2. Adatok bemutatása

Az adatok Magyarországról, illetve közvetlen környezetéről származnak. Az ECMWF (European Centre for Medium-Range Weather Forecasts, azaz a Középtávú Időjárás-

előrejelzések Európai Központja) által készített modell operatív, nagy felbontású $00UTC+24h$ előrejelzéseit használtam fel. A modell előrejelzési mezőiből praktikussági szempontokat figyelembe véve $0,5$ fokos rácsfelbontással, 3 óránként kerültek leválogatásra a Magyarországot reprezentáló rácsponti értékek. Az adatok számának redukálása, így a könnyebb vizsgálhatóság érdekében minden nap az adott paraméter összes napon belüli időpontban és minden rácspontban felvett értékhalmazának percentiliseit kaptam meg a modellezéshez. Ez bevett módszer hasonló statisztikai módszerek alkalmazásánál.

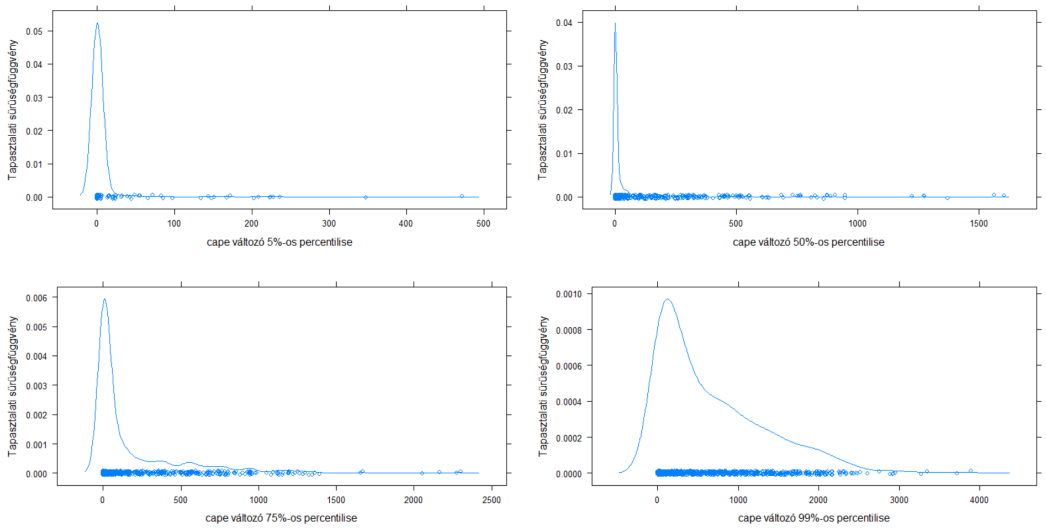
A változók, amelyek értékeit rendelkezésemre bocsátották:

$CAPE$: Konvektív hasznosítható potenciális energia, az angol Convective Available Potential Energy kezdőbetűi, mértékegysége J/kg. A légköri instabilitás, vagy más néven labilitást jellemző változó. Segítségével becsülhető, hogy a légkörben egy felfelé haladó részecskének mekkora a maximálisan elérhető mozgási energiája. Feltételezzük, hogy az ECMWF modell első 24 órára vonatkozó előrejelzett légköri paraméterei lényegesen nem térnek el a valóságban felvett értékektől. Minél magasabb az értéke, annál valószínűbbnek tartják a zivatar létrejöttét az adott területen (más, a konvekcióhoz szükséges feltételek fennállásakor). Sokféle beállítás szerint szokták számolni, én az ECMWF által használt metódus szerint számoltat használtam. Akár már $50-100J/kg$ is elég lehet a zivatarokhoz, de minél nagyobb az érték, annál nagyobb a valószínűsége a zivatarnak, azonban bizonyos szakirodalmak [7] szerint konkrét küszöbérték fölött már nem feltétlen alakul ki több zivatarfelhő ($500-1000J/kg$). A $cape1, cape5, cape25, cape50, cape75, cape95, cape99$ változók a megfelelő percentiliseket tartalmazzák az összes adott napon mért értékre vonatkozóan.

```
    cape_1          cape_5          cape_25         cape_50         cape_75         cape_95         cape_99
Min.   :  0.000   Min.   :  0.000   Min.   :  0.00   Min.   :   0.00   Min.   :   0.0   Min.   :   0.0   Min.   :   0.0
1st Qu.:  0.000   1st Qu.:  0.000   1st Qu.:  0.00   1st Qu.:   0.00   1st Qu.:   0.7   1st Qu.:  30.9   1st Qu.: 109.2
Median :  0.000   Median :  0.000   Median :  0.00   Median :   0.20   Median :  18.3   Median : 209.4   Median : 430.9
Mean   :  1.044   Mean   :  4.888   Mean   : 28.11   Mean   :  79.55   Mean   : 185.5   Mean   : 444.6   Mean   : 671.5
3rd Qu.:  0.000   3rd Qu.:  0.000   3rd Qu.:  0.40   3rd Qu.:  35.60   3rd Qu.: 225.5   3rd Qu.: 690.4   3rd Qu.:1057.5
Max.   :270.400   Max.   :471.200   Max.   :951.50   Max.   :1604.20   Max.   :2299.4   Max.   :3342.8   Max.   :3889.0
```

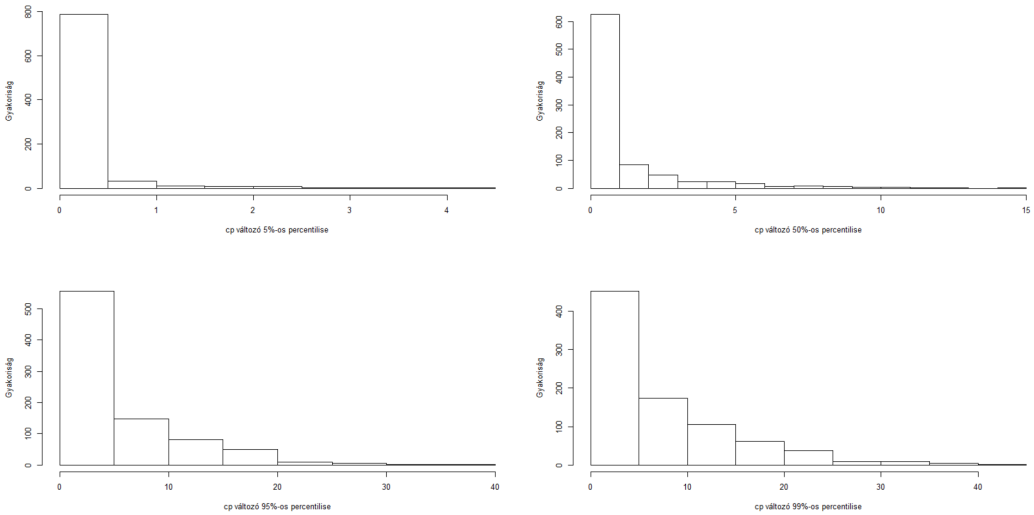
5.1. ábra. A $CAPE$ változó különböző percentiliseinek néhány leíró statisztikája

CP: Szintén az ECMWF modell által becsült konvektív (azaz a záporokhoz, zivatarokhoz társítható) csapadék mennyiség 6 óra alatt mm-ben megadva. Itt nem feltétlen csak olyan csapadékról van szó, amely villámokat is képes produkálni, de alapvetően a tornyosodó gomolyokból, zivatarfelhőkből hulló csapadékot próbálja a modell becsülni területi átlagban. A parametrizációban komplex módon vannak figyelembe véve a körülmények, köztük a labilitási viszonyok. A számok a változók végén itt is az adott napon mért összes



5.2. ábra. A $CAPE$ változó különböző percentiliseinek tapasztalati sűrűségfüggvénye

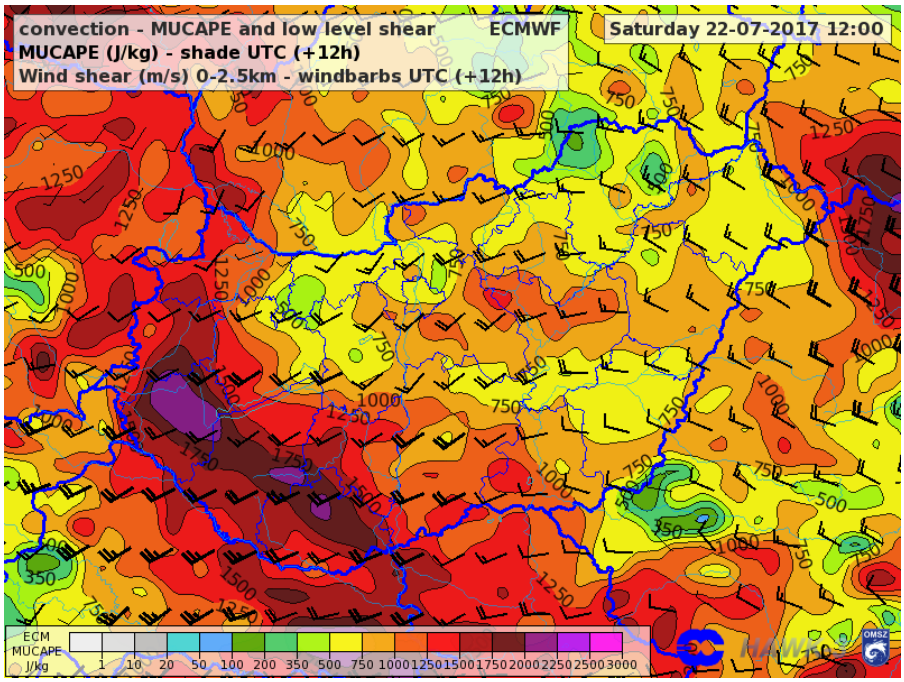
adat megfelelő percentilisét jelenti.



5.3. ábra. A CP változó különböző percentiliseinek hisztogramja

$w\_mean\_850-500$: Nagy skálájú feláramlási sebesség (Pa/s). A negatív előjeles értékek jelentik a feláramlást. A feláramlás segíti a zivatarfelhők beindulását, míg a leáramlás inkább gátolja.

$rh\_mean\_850-500$: $850, 700, 500$ hPa-os szintek átlagos relatív nedvessége (%). 50% alatt drasztikusan esik a zivatarfelhők kialakulási valószínűsége, ellenben kb. $70-80\%$ felett már nem nő a valószínűség, sőt a teljesen telített levegő akár kedvezőtlen is lehet.



5.4. ábra. $CAPE$ értékek Magyarország területén 2017.07.22-én

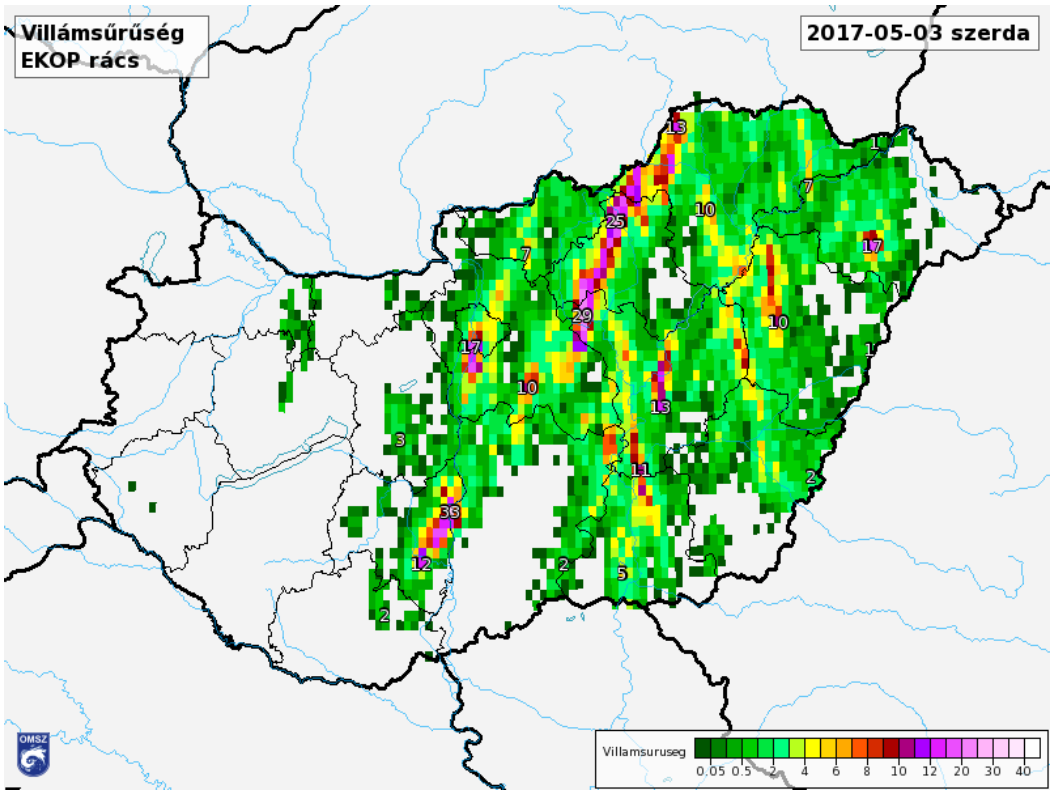
$q\_mean\_1000-850$: Átlagos specifikus nedvesség (g/kg) az $1000, 925, 850$ hPa-os szinteken. Adott hőmérsékleti profil mellett minél magasabb az abszolút vízgőztartalom az alsó szinteken, annál kedvezőbb a zivatarfelhők kialakulása.

$Lifted\_850to500$, $Lifted\_925to500$, $Lifted\_1000to500$: A Lifted index egy labilitást jellemző számérték, mely egy adott nyomási szintről (esetünkben 1000 (Li), 925 ($Li92$), 850 hPa ($Li85$) 500 hPa-ra fölemelkedő légrész hőmérsékletét hasonlítja össze az 500 hPa-os szint eredeti hőmérsékletével. Az index értéke az eredeti és a fölemelkedett légrész hőmérsékletének különbsége. Ha a légrész, mint légbuborék hőmérséklete magasabb, mint a környező levegőé, akkor az index negatív előjelű, ami labilitást jelez, míg ellenkező esetben stabil a légállapot. Az index értékét egyértelműen meghatározza a légrész kiindulási helyének légnyomása, hőmérséklete, nedvességtartalma, illetve az érkezési hely (500 hPa) hőmérséklete.

$FNUM$: A detektált villámkisülések darabszáma az egész nap során.

$FAREA$: A rácspontok közötti területek összessége, ahol az adott napon villámcsapás volt. A mértékegység km^2.

$FPERC$: Azon terület százalékos aránya a teljes területhez, amelyen detektáltak villámkisülést.



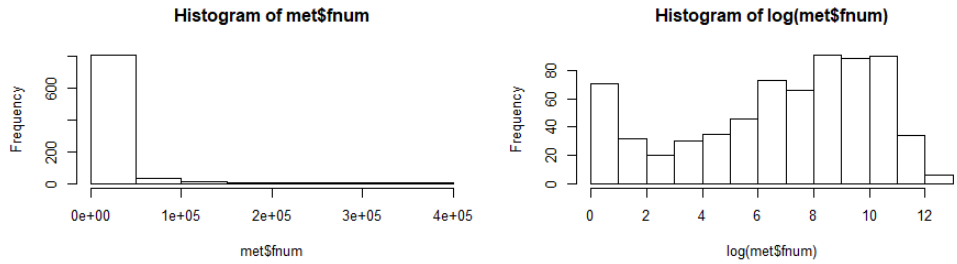
5.5. ábra. Villámsűrűség 2017.07.03-án Magyarország területén $0,05°$-os rácsra kiszámolva

FD: Villámsűrűség (villámok darabszáma osztva az érintett területtel). A számok szintén a megfelelő percentilisekre utalnak. A vizsgálatok során ezzel a változóval külön nem foglalkoztam.

Li_Rh_omega: Három paraméter szorzata. Az eredeti Li, rh_mean, w_mean paraméterek egy 0 és 1 közötti számmá lettek konvertálva, ahol azok meghatározott küszöbei között lineárisan változik a szám. Az Li esetén -7 és 3 a küszöbértékek ($Li < -7$ esetén: 1, $Li > 3$ esetén: 0), az rh_mean változó esetében 30 és 75 a küszöb ($rh\_mean < 30$ esetén 0, $rh\_mean > 75$ esetén 1), a w_mean változó esetében pedig -5 és 10 ($w\_mean < -5$ esetén 1, $w\_mean > 10$ esetén 0).

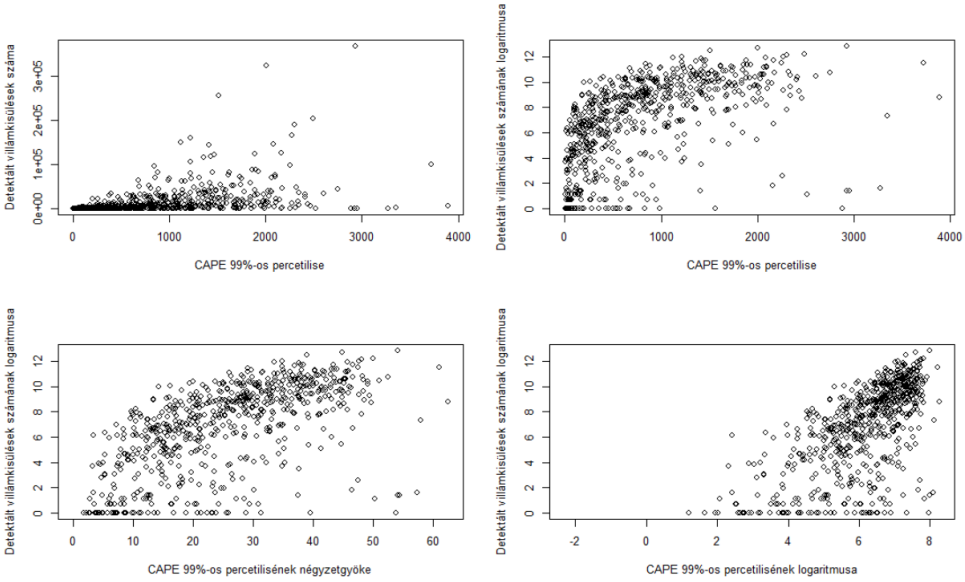
5.3. Előzetes vizsgálatok

Az $FNUM$ változó néhány napon nem tartalmazott adatot, így a hozzá tartozó sorokat töröltem az adathalmazból. A változó jobbra elnyújtott eloszlást mutatott, ezért a természetes alapú logaritmusával dolgoztam.



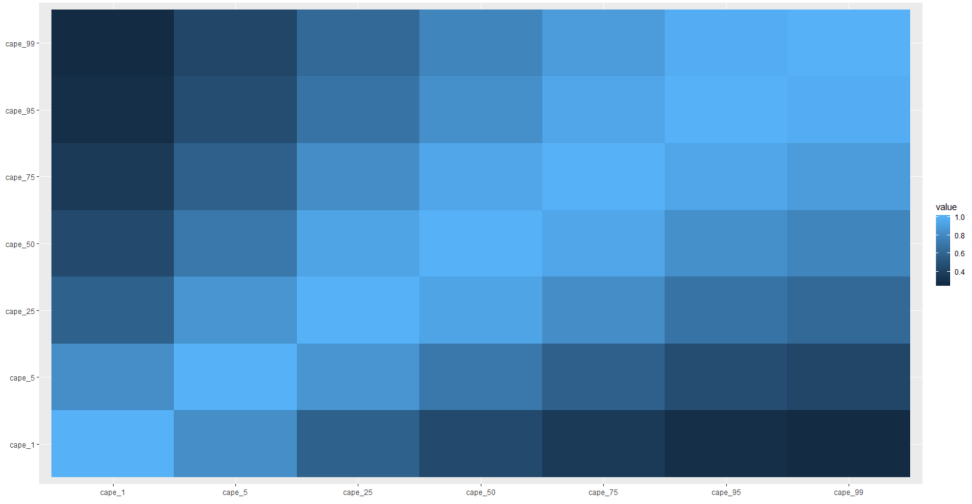
5.6. ábra. Az $FNUM$ változó és logaritmusának hisztogramja

Az $FNUM$ értékeket megvizsgálva találtam olyan napokat, ahol $1-2$ villám volt, azonban a néhány napos környezetükben nem volt villám. Ezek felülvizsgálását kértem a meteorológustól és kiderült, hogy néhány esetben fals értéket kaptunk. Ezeket javítottam 0-ra (2014.06.07.,2016.08.24.,2017.05.09.,2017.05.18. és 2017.05.27.) A magyarázó változókon is végeztem transzformációkat, majd közösen ábrázoltam őket a függő változóval a közöttük lévő kapcsolat megvizsgálásának céljából.



5.7. ábra. Különböző transzformációk alkalmazása a $CAPE$ és $FNUM$ változókon

Először az egyes változók különböző percentiliseinek egymással vett korrelációját vizsgáltam meg. Az alábbi ábrán a $CAPE$ változó különböző percentiliseinek egymással számolt korrelációs mátrixának hőtérképe látható, amelyből kiderül, hogy a legkisebb, illetve legnagyobb percentiliseket tartalmazó változók kevésbé korreláltak.



5.8. ábra. A $CAPE$ változó különböző percentiliseinek egymással számolt korrelációs mátrixának hőtérképe

Hasonló tulajdonsággal rendelkeztek még a cp,rh,w,td,$Li85$, illetve $Li92$ változók. Azonban néhány változónál már a legkisebb és legnagyobb percentilis is legalább $0,7$-es korrelációval rendelkezett. Ilyen volt a $t850$,q,$te85$, illetve az Li nevű változó. Ezt követően a különböző változók egymással számolt korrelációját vizsgáltam. Megfigyelhető volt, hogy a hasonló meteorológiai jelentéssel bíró változók erősen korreláltak egymással (pl. $Li85$, $Li92$).

5.4. Modell felépítése

5.4.1. Tanuló és teszt adatokra bontás

A modell validációjához tanuló és teszt adathalmazra bontottam az adatokat. Ez egy gyakori módszer a gépi tanuló algoritmusok tesztelésére. Lényege, hogy a teljes adathalmaz nagyobb részét használjuk a modell építésére és a megfelelő változók kiválasztására. A teszt adathalmazon a modell általánosító-képességét vizsgáljuk, mennyire képes jól előrejelezni a függő változó értékét olyan magyarázó változó értékekkel amilyeneket a tanulás alatt nem látott. Ezért fontos, hogy a tanuló és teszt adatbázis független legyen egymástól. A modellépítéshez a 2012-2017-ig terjedő adatokat használtam, míg a 2018. évi adatok alkották a teszt adathalmazt.

5.4.2. Lineáris regresszió felépítése

A lineáris regresszió egyik feltétele, hogy a magyarázó változók között ne legyen multikollinearitás. Ennek megállapításához a VIF mérőszámot használtam. Első lépésben az összes magyarázó változóra kiszámoltam a VIF értéket, majd amelyiknél a legnagyobb volt, azt elhagytam. Ezt a módszert iteráltam addig, amíg a maradék változók VIF értéke 5-nél kisebb lett. A megmaradt változók: $cape\_1$, $cape\_25$, $cape\_99$, cp_1, cp_99, $Li\_Rh\_omega.1$, rh_1, rh_99, w_1, w_75, w_99, $td85\_1$, $td85\_99$,$Li85\_1$ és Li_99.

A megmaradt változókat felhasználva kezdtem el építeni a modellt. Függő változónak az $fnum$ változó logaritmusát választottam (később: $logfnum$). Azonban ez azzal jár, hogy az összefüggés nem additív, hanem multiplikatív lesz a modellben. Az összes megmaradt változót betettem a modellbe, majd a nem szignifikánsak közül hagytam el egyet, addig, amíg nem maradt nem szignifikáns változó. Ennek megállapítására az egyes változókhoz tartozó p-értéket használtam. Az i. változó p-értéke a $H_0 : \beta_i = 0$ nullhipotézis tesztelésére használható. Minél kisebb az értéke, annál magasabb szignifikanciaszinten állítható, hogy az adott változó értékének megváltozásával a függő változó értéke is változik. Miután elhagytam egy változót megismételtem ezt a módszert. A megfelelő változók kiválasztásakor a meteorológus véleményét is kikértem, ami néhol befolyásolta a választást (pl. *cape* változónak kitüntetett szerepe volt). A megmaradt változók: $cape\_99$, cp_99, $Li\_Rh\_Omega.1$, w_1, w_75, $td85\_1$ és Li_99. A meteorológus javaslatára a *cape*, *cp* és *Li* változó 75-ös percentiliseit választottam a 99-es helyett.

A modell javításához néhány változó transzformáltjának bevonását használtam. A modellek összehasonlításához a korrigált determinációs együtthatót vizsgáltam (R-ben: Adjusted R-squared). A $cape\_75$ és cp_75 négyzetgyökét bevonva a modellbe sikerült jelentősen növelni a korrigált determinációs együttható értékét. A választott modell R outputja alább látható.

A hibák becsült szórása 2,1, amely az outputban a "Residual standard error" mellett található. A korrigált determinációs együttható mutatja, hogy a függő változó (villámkisülések számának logaritmusa) varianciájának 62%-át magyarázza a jelenlegi modell.

Látható, hogy a modell szignifikánsan különbözik attól a modelltől, amelyben minden magyarázó változó együtthatója 0 (p-érték $2,2 \cdot 10^{-16}$), vagyis létezik legalább egy magyarázó változó, amely a függő változót szignifikánsan magyarázza.

A "Coefficients" részben lévő oszlopok közül az első a becsült együtthatókat mutatja. Ez alapján írható fel pontosan a modell:

```
Call:
lm(formula = logfnum ~ cape_75 + sqrt(cape_75) + cp_75 + sqrt(cp_75) +
    Li_Rh_omega.1 + w_1 + w_75 + td85_1 + Li_75, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-6.6706 -1.1559  0.1678  1.4464  6.2431

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.8565161  0.3888003   7.347 7.21e-13 ***
cape_75       -0.0040088  0.0008342  -4.806 1.98e-06 ***
sqrt(cape_75)  0.1308864  0.0368732   3.550 0.000418 ***
cp_75         -0.6190914  0.0736699  -8.404 3.59e-16 ***
sqrt(cp_75)    2.6264784  0.2727297   9.630  < 2e-16 ***
Li_Rh_omega.1  6.7701274  1.2820527   5.281 1.84e-07 ***
w_1           -0.2378077  0.0459656  -5.174 3.20e-07 ***
w_75          -0.3543585  0.2057505  -1.722 0.085572 .
td85_1        -0.0859369  0.0266273  -3.227 0.001322 **
Li_75         -0.3066857  0.0500478  -6.128 1.68e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.146 on 560 degrees of freedom
Multiple R-squared:  0.6246,    Adjusted R-squared:  0.6186
F-statistic: 103.5 on 9 and 560 DF,  p-value: < 2.2e-16
```

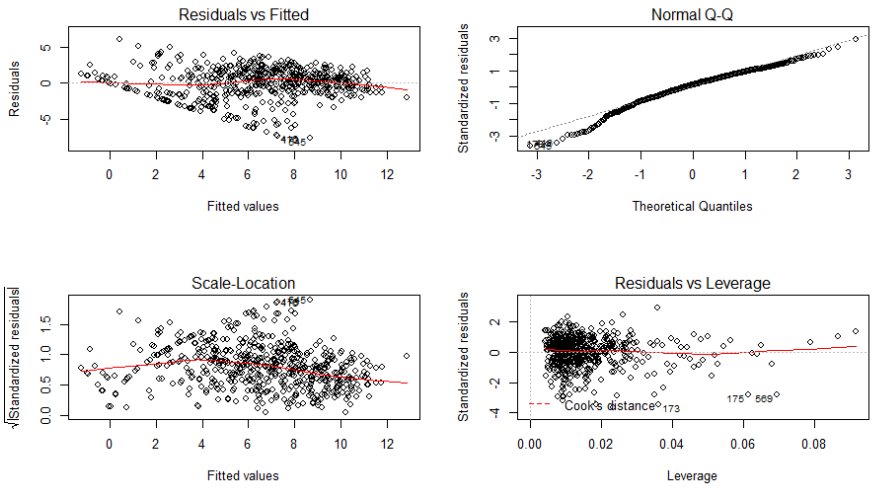
5.9. ábra. A modell R-ben kapott outputja

$$logfnum = 2,86 - 0,004cape\_75 + 0,13\sqrt{cape\_75} - 0,62cp\_75 + 2,63\sqrt{cp\_75} +$$
$$+ 6,77Li\_Rh\_omega.1 - 0,24w\_1 - 0,35w\_75 - 0,09td85\_1 - 0,3Li\_75$$

Az együtthatók segítenek megérteni mennyit változik a függő változó becsült értéke a magyarázó változók növelése/csökkenése esetén. Például ha a 850 hPa-n számított harmatpont ($td85\_1$) 1%-os percentilise $1C^{\circ}$-el nő (miközben a többi változó értéke fix marad), akkor a villámkisülések számának logaritmusának becsült értéke $0,09$-vel csökken, vagyis a villámkisülések számának becsült értéke kb. 91%-ára csökken ($e^{-0,09}$).

Az alább látható ábrákat használtam a lineáris regresszió további feltételeinek teljesülésének ellenőrzésére. Többszörös lineáris regresszió esetén szükséges feltétel, hogy a hibák normális eloszlást kövessenek és azonos szórással rendelkezzenek.

Az "residuals vs. fitted" ábrán a modell által becsült válaszértékek (x tengely) és a megfelelő reziduálisok láthatóak, amelyen nem látható jele nemlineáris kapcsolatnak. A "standardized residuals vs. fitted" ábrán a reziduálisok standardizálás után vannak ábrázolva. Ezen ellenőriztem a homoszkedaszticitást. Egy vízszinteshez közel álló tengely mentén egyenlően szóródnak a reziduálisok, ami a feltétel teljesülésére utal. Az ábrán látható kiugró értékek kivizsgálását kértem a meteorológustól, de nem tartalmaztak mérési



5.10. ábra. A reziduálisok néhány ábrázolása

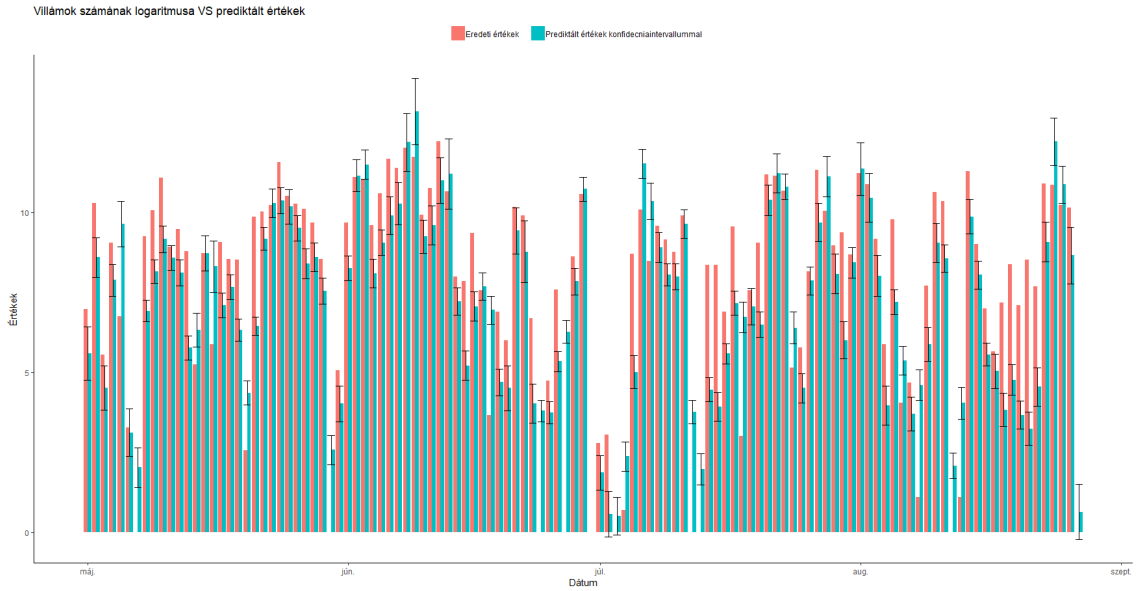
hibát. Az 5.10. ábra jobb felső részén egy Q-Q plot látható annak ellenőrzésére, hogy normális eloszlásból származnak-e a reziduálisok. Ha az egyenes mentén helyezkednének el, akkor teljesülne ez a feltétel. Látható, hogy ez nem teljesül tökéletesen, az alacsonyabb értékeknél mutat némi eltérést. A negyedik ábra a kiugró értékek detektálására alkalmas Cook-távolság alapján. A Cook-távolság azt méri, hogy a regressziós paraméterek mennyire változnak meg az adott érték elhagyásával. Az ábra alapján nincs olyan érték, aminek nagy lenne a Cook-távolsága. A feltételek nem tökéletes teljesülése miatt bootstrap módszerrel illesztettem konfidenciaintervallumot a paraméterekre, de ezek nem tértek el nagy mértékben az eredeti konfidenciaintervallumtól.

	becsült paraméterek	bootstrap 2,5%	bootstrap 97,5%	eredeti 2,5%	eredeti 97,5%
X.Intercept.	2.867	2.209	3.688	2.101	3.633
cape_75	-0.004	-0.005	-0.002	-0.006	-0.002
sqrt.cape_75.	0.131	0.060	0.195	0.059	0.204
cp_75	-0.611	-0.719	-0.444	-0.756	-0.466
sqrt.cp_75.	2.613	1.974	2.970	2.075	3.151
Li_Rh_omega.1	6.764	4.234	9.095	4.247	9.281
w_1	-0.231	-0.310	-0.139	-0.321	-0.140
w_75	-0.328	-0.700	0.080	-0.733	0.076
td85_1	-0.091	-0.136	-0.035	-0.144	-0.039
Li_75	-0.302	-0.371	-0.190	-0.401	-0.203

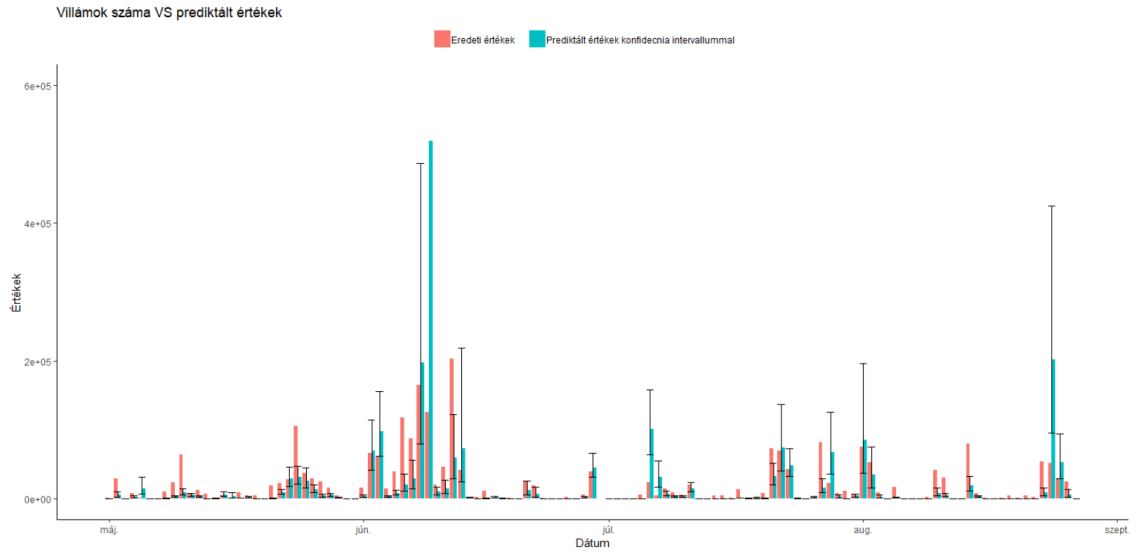
5.11. ábra. Konfidenciaintervallum a paraméterekre

5.4.3. Előrejelzés a lineáris modell alapján

A tanuló adathalmazon épített modellt a 2018. évi adatokat tartalmazó teszt adathalmazon teszteltem. A modell függő változójaként a villámkisülések logaritmusát választottam,



5.12. ábra. Prediktálás a villámkisülésekre logaritmikus skálán



5.13. ábra. Prediktálás a villámkisülésekre

így először azzal hasonlítottam össze a prediktált értékeket. Majd visszatranszformáltam az értékeket, hogy a valós villámadatokhoz tudjam hasonlítani a modell előrejelzéseit. A visszatranszformált és a logaritmikus skálán való prediktálás látható az 5.12 és 5.13 ábrán konfidenciaintervallumokkal együtt.

Általánosságban kijelenthető, hogy a nagyságrendi változást előre lehet jelezni, azonban a villámkisülések igen összetett meteorológiai tényezők következményei, amelyeket

nehéz ennél pontosabban előrejelezni. A pontosabb prediktálás további vizsgálatok útján lenne lehetséges.

5.4.4. Logisztikus modell felépítése

A tanuló és teszt adathalmazra bontást itt is elvégeztem, így a 2018-as adatok alkották a teszt, míg az összes többi a tanuló adathalmazt. Először az $fnum$ változóból kialakítottam a logisztikus regresszió függő változóját (bin_fnum). Ha volt egy adott napon villámkisülés, akkor 1-es, ha nem volt, akkor 0 értékké transzformáltam. A lineáris regresszióhoz hasonlóan a logisztikus modellnek is feltétele, hogy ne legyen a magyarázó változók között multikollinearitás, így a VIF mutató alapján kiválasztott változókból indultam ki a logisztikus modell építésénél is.

Stepwise szelekciót alkalmaztam a modell kiválasztásához. Mindkét módszert (forwards, backwards) lefuttatva ugyanazokat a változókat kaptam eredményül ($cape\_25$, $cape\_99$, cp_99, $Li\_Rh\_omega.1$, $t850\_99$, w_75, w_99, $td85\_1$, $Li85\_1$ és Li_99), így ezekkel folytattam a további vizsgálódást.

A nem szignifikáns változókat egyesével elhagyva hasonlítottam össze az így kapott modelleket. Egyrészt az AIC értéket figyeltem, amelyiknél kevesebb volt, azt választottam. Másrészt igénybe vettem a meteorológus segítségét, hogy a megfelelő változók bent maradjanak a modellben. A *cape* változó esetében alacsonyabb percentilist választottam, mert meteorológiai értelemben jobban jellemzi a függő változót. Így kaptam a lentebb látható R outputot.

Ez alapján felírható a logisztikus modell:

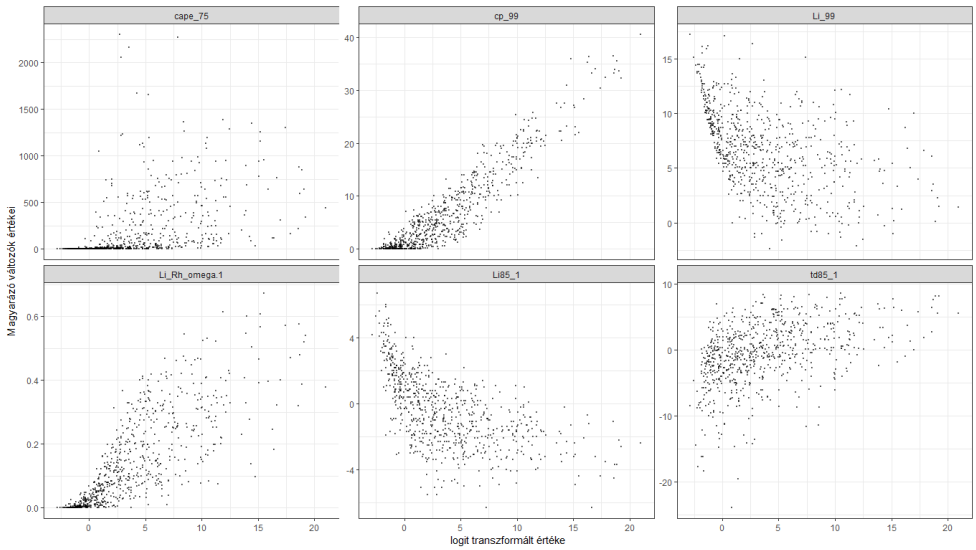
$$\begin{aligned} logit(bin\_fnum) = 0,32 + 0,37cp\_99 - 0,31Li\_99 - \\ - 0,12td85\_1 - 0,16Li85\_1 - 0,002cape\_75 + 9,36Li\_Rh\_omega.1 \end{aligned}$$

Az együtthatók úgy értelmezhetőek, hogy ha például egységnyit növeljük az $Li85\_1$ lifted index értékét az $0,16$-tal csökkenti a $logit(bin\_fnum)$ értékét. Ebből következik, hogy az *oddsz* értéke a 85%-ára ($e^{-0,16}$) csökken, ami a villámkisülés bekövetkezése valószínűségének és a villámkisülésmentes nap valószínűségének hányadosa.

Fontos még ellenőrizni a feltételeket. A mintanagyság megfelelő, mert 6 db magyarázó változó szerepel a modellben, de több, mint 60 megfigyelésen tanítottuk a modellt.

```
Call:
glm(formula = bin_fnum ~ cp_99 + Li_99 + td85_1 + Li85_1 + cape_75 +
    Li_Rh_omega.1, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0947  -0.4279   0.0477   0.4420   2.5664

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     0.3189425  0.4033915   0.791 0.429147
cp_99           0.3700362  0.0416843   8.877  < 2e-16 ***
Li_99          -0.3136924  0.0532525  -5.891 3.85e-09 ***
td85_1         -0.1199189  0.0324437  -3.696 0.000219 ***
Li85_1         -0.1572093  0.0920754  -1.707 0.087748 .
cape_75        -0.0023354  0.0005189  -4.501 6.76e-06 ***
Li_Rh_omega.1   9.3607816  2.1642126   4.325 1.52e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 973.89  on 734  degrees of freedom
Residual deviance: 469.43  on 728  degrees of freedom
AIC: 483.43

Number of Fisher Scoring iterations: 7
```

5.14. ábra. A logisztikus modell R outputja

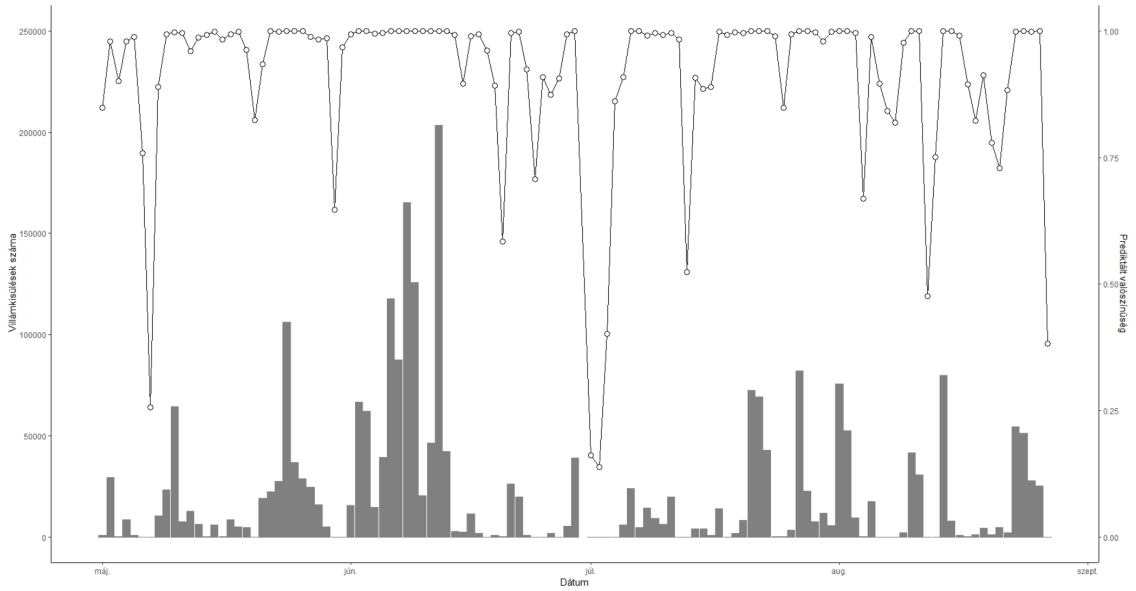


5.15. ábra. Magyarázó változók függő változóval ábrázolása

A linearitás ellenőrzésére az 5.14. ábrát használtam. Ideális esetben a pontok közel egy egyenes mentén helyezkednének el, de ez nem minden esetben teljesül tökéletesen. Box-Tidwell módszerét nem tudtam alkalmazni, mert csak pozitív változókra alkalmazható, de a modellben használt nemnegatív magyarázó változók értékei között vannak 0 értékek is.

5.4.5. Előrejelzés a logisztikus modell alapján

Teszteléshez a 2018-as adatokat használtam. Ábrázoltam a villámkisülések számát és a felépített modell által prediktált valószínűségeket. Az alább látható ábrán láthatóak a prediktált valószínűségek az adott napon bekövetkezett villámkisülésekkel közösen ábrázolva.



5.16. ábra. A logisztikus modell alapján prediktált valószínűség vs valós villámkisülés db

70%-os prediktált valószínűség fölött nem volt olyan nap, amikor ne lett volna ténylegesen is villámkisülés. Néhány alacsonyabb prediktált valószínűségű napon (2018.07.01 és 2018.07.02) voltak valós villámkisülések. Azonban a magyarázó változók értékei közelebb álltak olyan napok változó értékeihez, amelyeken nem voltak villámkisülések. Azokra a napokra, amelyeken nem volt villámkisülés elmondható, hogy a modell is alacsonyabb ($< 40\%$) valószínűséget prediktált. 1 esetben fordult elő (2018.06.27), hogy bár magasabb valószínűség lett előrejelezve (67%), mégsem volt villámkisülés. A meteorológus kivizsgálta ezt a napot és valóban nagy mennyiségű konvektív csapadék hullott, de ennek ellenére nem volt villámkisülés.

6. fejezet

Összegzés, további vizsgálatok

Dolgozatomban a regresszió módszerét alkalmaztam egy nagyon specifikus, meteorológiai adathalmazra. Az adatok amelyekkel dolgoztam összetett légköri jelenségeket leíró változók voltak, ebből következően nem voltak a legideálisabbak egyszerű statisztikai vizsgálatokhoz. Ennek ellenére látható, hogy mégis használható eredményeket lehet elérni a különböző regressziós modellekkel.

A lineáris regresszió alkalmazásával nagyságrendben előrejelezhetővé vált a következő napi villámkisülések száma, ami jó becslést ad a másnapi zivatarok intenzitásának megállapításához. A logisztikus regressziós modell nagy mértékben segített annak megállapításában, hogy Magyarország területén egyáltalán várható-e villámkisülésekkel járó zivatarfelhő a modell futtatásának első 24 órájában (az ECMWF modell 24 órás előrejelzéseivel dolgoztam, ezért lehet a következő napra előrejelezni az általam készített modellel).

További lehetséges munka lehet az eddig elkészített modell fejlesztése a fentebb részletezett módszerekkel, például a multikollinearitás kiküszöbölésére használható Ridge-regresszió alkalmazása vagy az adatokban jelenlévő autokorreláció korrigálása. Eddig nem vizsgált további releváns változók beválasztása is javíthat a modelleken. A regressziós összefüggések alkalmazását az ECMWF modell következő pár napra vonatkozó előrejelzéseire is ki lehet terjeszteni, figyelembe véve azt, hogy az idő előrehaladtával a becslés hibája növekszik. Esetleg az adatok összegzése helyett, valamilyen térstatisztikai módszert is használhatunk. Egy ilyen módszer a krigelés, amely a geostatisztikában elterjedt becslési eljárás, melynek segítségével kiszámítható egy paraméter tetszőleges helyen várhatóan felvett értéke. [17]

A feltételek nem tökéletes teljesülése miatt a logisztikus regresszió mellett lehetne még az általánosított additív modellen alapuló additív logisztikus regressziót is alkalmazni. [16]

Irodalomjegyzék

[1] Raymond H. Myers, *Classical and Modern Regression with Applications (Second edition)*, 1990

[2] David W. Hosmer, Stanley Lemeshow, *Applied Logistic Regression*, 2000

[3] Márkus László, *Regresszió jegyzet*,
http://web.cs.elte.hu/probability/markus/ElemzoTS1/RegressionSlides2018_12_15.pdf

[4] Walter W. Hauck Jr., Allen Donner, *Wald's Test as Applied to Hypotheses in Logit Analysis*, 1977

[5] Hunyadi László, Mundruczó György, Vita László, *Statisztika*, 1997

[6] Richard A. Johnson, Gouri K. Bhattacharyya, *STATISTICS Principles and Methods*, 2010

[7] Anja T. Westermayer, Pieter Groenemeijer, Georg Pistotnik, Robert Sausen, Eberhard Faust, *Identification of favorable environments for thunderstorms in reanalysis data*, 2016

[8] Thorsten Simon, Georg J. Mayr, Nikolaus Umlauf,Achim Zeileis, *Lightning prediction using model output statistics*, 2018

[9] Bolla Marianna, Krámli András, *Statisztikai következtetések elmélete*, 2015

[10] Hans D. Betz, Kersten Schmidt, Wolf P. Oettinger, *LINET - An International VLF/LF Lightning Detection Network in Europe*, 2009

[11] ECMW hivatalos dokumentáció
https://www.ecmwf.int/en/forecasts/documentation-and-support

[12] Galway, J.G., *The lifted index as a predictor of latent instability.*, 1956

[13] Barbara G. Tabachnick, Linda S. Fidell, *Using Multivariate Statistics*, 2013

[14] Lopez, P,*Promising results for lightning predictions (ECMW Newsletter No. 155 ? Spring 2018)*, 2018

[15] Francis Galton wikipedia szócikk, https://hu.wikipedia.org/wiki/Francis_Galton

[16] T. Rädler, P. Groenemeijer, G. Pistotnik, R. Sausen, *Identification of thunderstorms in reanalysis data and development of a statistical convective initiation model*, 2015

[17] Krigelés wikipedia szócikk, https://hu.wikipedia.org/wiki/Krigel%C3%A9s

A. függelék

Részlet az R programkódból

```
##############################
#     Lineáris regresszió     #
##############################

#függő változó kreálása
met_alap$logfnum <- log(met_alap$fnum)
met_alap$logfnum[which(met_alap$logfnum==-Inf)] <- NA
#tanuló és teszt adathalmazra bontás
train <- subset(met_alap,
substr(met_alap[,1],1,4) != "2018",
select=1:length(met_alap[1,]))
test <- subset(met_alap,
substr(met_alap$datum,1,4) == "2018",
select=1:length(met_alap[1,]))
##############################################
# 1. Multikollinearitás vizsgálata VIF alapján #
##############################################
#VIF-hez csak a magyarázó változók kiválasztása
#(vif_func: kiszámolja a VIF értéket minden változóra,
# és kidobja a legnagyobbat, amíg az értékek kisebbek lesznek, mint a tresh)
regs <- subset(train,
```

```
select=-c(fnum,logfnum,datum,ev,honap,index))
vegso <- vif_func(regs, thresh = 5,trace = TRUE)
# Nem szignifikáns változók vizsgálata #
#######################################
model <- lm(logfnum~cape_1+cape_25+cape_99+cp_1+cp_99+
Li_Rh_omega.1+rh_1+rh_99+w_1+w_75+w_99+td85_1+Li_1+Li_99,
 data=train)
summary(model)
plot(model)

#változók egyesével elhagyása után keletkező modell

model <- lm(logfnum~cape_75+cp_75+Li_Rh_omega.1+
w_1+w_75+td85_1+Li_75, data=train)
summary(model)
plot(model)

#transzformációk alkalmazása a modell javítása végett
model <- lm(logfnum~cape_75+sqrt(cape_75)+cp_75+sqrt(cp_75)
+Li_Rh_omega.1+w_1+w_75+td85_1+Li_75, data=train)
summary(model)
#Választott modell#
###################
fitt <- lm(logfnum~cape_75+sqrt(cape_75)+cp_75+sqrt(cp_75)
+Li_Rh_omega.1+w_1+w_75+td85_1+Li_75, data=train)
summary(fitt)
plot(fitt)
#reziduálisok vizsgálata
shapiro.test(rstandard(fitt))
hist(rstandard(fitt))
#   Bootstrand konfidencia-intervallum  #
########################################
formula <- logfnum~cape_75+sqrt(cape_75)+
cp_75+sqrt(cp_75)+Li_Rh_omega.1+w_1+w_75+td85_1+Li_75
```

```
train_na <- na.omit(train)
fitt <- lm( formula, data = train_na );  summary(fitt)

cipred <- predict(fitt, interval = "confidence")
cipars <- confint(fitt, level = 0.95); cipars
resids <- residuals(fitt)
preds  <- fitted(fitt)

set.seed(2019)
nb <- 900  # bootstrap minta száma
cnum <- length(fitt$coeff) # paraméterek száma (beleértve az intercept-et)

coefmat <- matrix(0, nb, cnum)
predmat <- matrix(0, nb, length(preds))

for(i in 1:nb) {
booty <- preds + sample(resids, replace = TRUE)   # bootstrap minta
booty[booty<0] <- 0
bmod <- update(fitt, booty ~ . )   # modell illesztése
coefmat[i,] <- coef(bmod)
predmat[i,] <- fitted(bmod)
}
colnames(coefmat) <- names(fitt$coeff)
coefmat <- data.frame(coefmat)
cis <- apply(coefmat, 2, function(x) quantile(x, c(0.025, 0.975)))
t(cis)   #bootstrap konfidencia-intervallum a paraméterekre

confs <- round(data.frame(t(cis),cipars),3)
confs$coefs <- round(fitt$coefficients,3)
colnames(confs) <- c("bootstrap 2,5%","bootstrap 97,5%",
"eredeti 2,5%","eredeti 97,5%",
"becsült paraméterek")
confs
#   Prediktálás a lineáris modell alapján   #
```

```
##########################################
test_data <- data.frame(test$cape_75,
sqrt(test$cape_75),test$cp_75,
sqrt(test$cp_75),test$Li_Rh_omega.1,
test$w_1,test$w_75,test$td85_1,test$Li_75)
names(test_data) <- c("cape_75",
"sqrt(cape_75)","cp_75","sqrt(cp_75)",
"Li_Rh_omega.1","w_1","w_75","td85_1"
,"Li_75")
predict(fitt,newdata=test_data)
test_val_conf <- predict(fitt,
newdata=test_data,interval="confidence")
predict_plot <- data.frame(test_val_conf)
predict_plot$datum <- as.Date(as.character(test$datum),"%Y%m%d")
predict_plot$log_fnum <- test$logfnum
#összehasonlító ábra -logaritmikus skálán

date <- rep(predict_plot$datum,2)
log_fnum <- test$logfnum
log_fnum_p <- test_val_conf[,1]
values <- c(log_fnum,log_fnum_p)
type <- c(rep("data",length(log_fnum)),
rep("pred",length(log_fnum_p)))
lower <- c(rep(NaN,length(log_fnum)),test_val_conf[,2])
upper <- c(rep(NaN,length(log_fnum)),test_val_conf[,3])
p <- data.frame(date,values,type,lower,upper)

ggplot(p, aes(date, values)) +
ggtitle("Villámok számának logaritmusa VS prediktált értékek") +
xlab("Dátum") + ylab("Értékek")+
theme_classic()+
theme(legend.title=element_blank())+
theme(legend.position="top")+
scale_fill_discrete(labels=c("Eredeti értékek",
```

```
 "Prediktált értékek konfidecnia intervallummal"))+
geom_bar(stat = "identity", aes(fill = type), position = "dodge") +
geom_errorbar(aes(ymin=lower, ymax=upper))
####################################
#       LOGISZTIKUS REGRESSZIŐ      #
####################################
# függő változó kreálás

met_alap$bin_fnum <- cut(
met_alap$fnum,
breaks = c(-Inf, 0, Inf),
labels = c(0,1),
right  = TRUE
)
#Tanuló és teszt adathalmazra bontás
train<-subset(met_alap,
substr(met_alap[,1],1,4) != "2018",
select=1:length(met_alap[1,]))
test<-subset(met_alap,
substr(met_alap$datum,1,4) == "2018",
select=1:length(met_alap[1,]))
#Stepwise modellépítés
logm <- glm(bin_fnum~cape_1+cape_25+cape_99+cp_1+cp_99+
Li_Rh_omega.1+t850_99+rh_1+rh_99+w_1+w_75+w_99+
td85_1+Li85_1+Li_99, family="binomial",data=train)
summary(logm)
logm_backw <- step(logm)
summary(logm_backw)
loginull <- glm(bin_fnum ~ 1, data = train, family = binomial)
logm_forw <- step(loginull,
scope = list(lower=formula(loginull),
 upper=formula(logm)),direction= "forward")
#választott modell#
summary(logm_forw)
```

```
model<-glm(bin_fnum~cp_99+Li_99+td85_1+Li85_1+
cape_99+Li_Rh_omega.1,family="binomial",data=train)
model<-glm(bin_fnum~cape_99+Li_Rh_omega.1+cp_95+
td85_90+Li_95,family="binomial",data=train)
summary(model)
logi<-glm(bin_fnum~cp_99+Li_99+td85_1+Li85_1+
cape_75+Li_Rh_omega.1,family="binomial",data=train)
summary(logi)
nc <- names(logi$coeff)
metv <- train[c( nc[which(nc!="(Intercept)")] )]

# villám valószínuségének (p) becslése
probabilities <- predict(logi, type = "response")
predicted.classes <- ifelse(probabilities > 0.5, "pos", "neg")
table(predicted.classes)
mydata <- mydata %>%
mutate(logit = log(probabilities/(1-probabilities))) %>%
gather(key = "predictors", value = "predictor.value", -logit)

# Linearitás ellenőrző ábrák:
ggplot(mydata, aes(logit, predictor.value))+
geom_point(size = 0.5, alpha = 0.5) +
theme_bw() +
xlab ("logit transzformált értéke") +
ylab("Magyarázó változók értékei") +
facet_wrap(~predictors, scales = "free_y")
logi<-glm(bin_fnum~cp_99+Li_99+td85_1+Li85_1+
cape_75+Li_Rh_omega.1,family="binomial",data=train)
#prediktálás
test_data=data.frame(test$cp_75,test$Li_75,
test$Li85_1,test$cape_75,test$Li_Rh_omega.1)
names(test_data)=c("cp_75","Li_75","Li85_1","cape_75","Li_Rh_omega.1")
pred<-predict(model, test, type="response")
test$fnum
```

```
prediction<-data.frame(test$fnum,pred)
names(prediction)=c("fnum","pred")
prediction
date <- test$datum
prediction$date=date
date<-as.Date(as.character(date),"%Y%m%d")
ggplot(prediction,aes(x=date))+
geom_bar(mapping = aes(x = date, y = fnum),
stat = "identity", colour = gray(0.5), fill = gray(0.5)) +
geom_line(mapping = aes(x = date, y = pred*250000)) +
geom_point(mapping = aes(x = date, y = pred*250000),
size = 3, shape = 21, fill = "white") +
scale_y_continuous(
name = expression("fnum"),
sec.axis = sec_axis(~ (. + 0) / 250000 , name = "prediction"),
limits = c(0, 250000)) +
theme_classic()
prediction$id=1:length(prediction$fnum)
```