NYILATKOZAT

Név: MARICA EDINA
ELTE Természettudományi Kar, szak: MATEMATIKA BSC
NEPTUN azonosító: TEOKN8
Szakdolgozat címe: JÁRVÁNYOK KISZŰRÉSE IDŐSOROS ADATOKBÓL

A **szakdolgozat** szerzőjeként fegyelmi felelősségem tudatában kijelentem, hogy a dolgozatom önálló szellemi alkotásom, abban a hivatkozások és idézések standard szabályait következetesen alkalmaztam, mások által írt részeket a megfelelő idézés nélkül nem használtam fel.

Budapest, 2022. május 31.

Marica

a hallgató aláírása

Eötvös Loránd Tudományegyetem

Természettudományi Kar

Marica Edina

Járványok kiszűrése idősoros adatokból

Szakdolgozat

Matematika BSc

alkalmazott matematikus szakirány

Témavezető:

dr. Zempléni András

egyetemi docens

Valószínűségelméleti és Statisztika Tanszék

ELTE

EÖTVÖS LORÁND

TUDOMÁNYEGYETEM

Budapest, 2022

Köszönetnyilvánítás

Szeretném megköszönni Zempléni András tanár úrnak a türelmet, odaadást, segítséget, és azt az időt amit az elmúlt egy évben hetente rám áldozott. Köszönöm szüleimnek, tanáraimnak és hallgatótársaimnak is a biztatást és támogatást.

Tartalomjegyzék

1. fejezet

Bevezetés

A koronavírus világjárvány Kínában, Vuhanban jelent meg először 2019. decemberében. Hazánkban 2020. március 4-én jelentették be az első fertőzöttet, így márciustól kezdődött a vírus első hulláma. Ezt követte a második hullám 2020. augusztus végi kezdettel, a harmadik hullám 2021. januárjában, végül a negyedik hullám ugyanezen év kora őszén. Az első hullám viszonylag enyhe volt Magyarország területén, a következő három viszont már több százezer fertőzést okozott.

1.1. A halálozási adatok használatának motivációja

Egy járvány helyzetét a legjobban a fertőzöttségi adatokkal lehet kiszűrni. Ezen adatok viszont nagyon szorosan összefüggnek a tesztelési számokkal. Olyan járványok esetén, amelyek erős és egyértelmű tünetekkel járnak, a tesztelések száma ténylegesen tükrözheti a megbetegedések számát. A Covid19 esetében viszont nagyon gyakori a tünetmentesség, vagy enyhe megfázási tünetek fellépése, amelyek nem késztetik a beteget tesztelésre. Ezért a megbetegedések számát nem lehet pontosan meghatározni. Tekintve ezt a hátrányt, most vizsgáljuk a koronavírus okozta halálozási számokat. A halandóság egyben egy járvány súlyosságának a legjobb fokmérője is.
Ezek az adatok pontosabbak, hiszen minden halál be van jelentve. Az viszont még mindig nem biztos, hogy a halál okának a koronavírus van feltüntetve. Előfordulhat, hogy egy személy a vírus miatt veszti életét, de mivel nem volt tesztelve, nem annak nyilvánítják, illetve a fordított eset is lehetséges. Viszont mégis pontosabb a megbetegedési számoknál, hiszen a koronavírus létezésének tudatában általában tesztelték a nagyon beteg állapotbn levő lakosokat. Végső soron tehát ez a mérési szám a gyakran nem egyértelmű haláloki

besorolásra támaszkodik.
Nézzünk meg egy harmadik lehetőséget, amelyben a halálozási okoktól független halálozási számokat vizsgáljuk. Ezzel kiküszöböljük a tesztelések számának is és a haláloki besorolásnak is a problémáját. Ha kiszámoljuk, hogy mennyi lett volna a várt halálozási szám, ha nem tört volna ki a világjárvány, akkor ezt kivonva a tényleges halálozási adatokból, megkapjuk a vírus hatását. Viszont ez a direkt és indirekt hatást is magába foglalja. Indirekt hatás például az influenza gyenge hatása az intézkedések bevezetése során, vagy a kevesebb autóbaleset, de negatív indirekt hatásai is vannak a világjárványnak, például a kórházi ellátás csökkenése a más betegségekben szenvedők számára, magasabb kábítószer fogyasztás, és több öngyilkosság.

Lehetséges hátrányai a módszernek

A haláloktól független halálozási számok vizsgálatakor néhány problémába ütközhetünk. Az első a halálok bejelentéséből adódó eltolódás, hiszen több hétbe is beletelhet, míg egy halál bekerül az összesítésbe. Így, ez az indikátor lassú, a régebbi járványhelyzetet tükrözi, és nem a jelenét. Ez olyankor jelenthet problémát, mikor intézkedések következményeit kell vizsgálni. A korlátozások létesítése után még egy hónappal is lehetnek bejelentve halálok, amelyek nem ezt az időszakot, hanem az előtte levőt tükrözik - torzítva az intézkedések hatásosságáról kapott képet. Ebben a dolgozatban viszont ez nem egy számottevő probléma, mert a célunk az összesített vizsgálat, nem a járvánnyal érintett időpontok behatárolása.
Egy másik, nagyobb jelentőségű kérdés a dolgozat szempontjából az, hogy hogyan lehet minél pontosabban meghatározni, hogy mennyi lett volna a halálesetek száma, a koronavírus kitörése nélkül. Ehhez több módszert is bemutatok és összehasonlítom őket. Emellett, ez a módszer egybeméri a járvány direkt és indirekt hatásait is. Erről bővebben a 3.9 fejezetben lesz szó.

1.2. Célkitűzés

A cél a koronavírus időszakában a halálozási számok növekedésének vizsgálata. A halálozási adatok elemzése segítséget nyújthat a hatóságoknak is egyes intézkedések szükségességének felismerésére, illetve a járvány súlyosságának elemzésében és a folyamatok megértésében is fontos szerepet játszik. Ezért a cél Magyarországon a Covid19 világjár-

vány hatásának vizsgálata a halálozási adatokból.

1.3. Adatok

A magyarországi 2015-2019-es halálozási adatokat hasonlítottuk a 2020-2021-es évek adataihoz, amikor már a koronavírus járvány elterjedt Magyarország területén is. A 2015-2021-es években elhunytak száma hetekre és megyékre lebontva (az utóbbi százezer lakosra nézve) a Központi Statisztikai Hivatal adatbázisából letölthető ([1], [2], [3]). A heti covid halálozási számokat pedig a Koronamonitor oldal adatai szolgáltatták ([4]).

2. fejezet

Elméleti háttér

Ebben a fejezetben a dolgozathoz szükséges matematikai fogalmakat és módszereket mutatom be. A definíciókat és tételeket az egyetem statisztikát tanító tanárainak a jegyzeteiből gyűjtöttem össze és dolgoztam fel ([11], [24], [25]). A módszerek és modellek kidolgozására pedig különböző statisztika jegyzeteket és cikkeket használtam ([10], [14], [16], [18], [19], [20], [21], [23], [26], [27]).

2.1. Idősorok

2.1.1. Definíció. *Az $X_1, X_2, ..., X_n$ valószínűségi változókat idősornak hívjuk, ha az indexeik időként is értelmezhetők.*

Az idősor elemzés tehát az időtől függő adatsorok változásainak elemzése. Ez egy sztochasztikus folyamat, azaz olyan folyamat, amelyet valószínűségi változók jellemeznek.

2.1.1. A modell

Az idősorok a következő modellel írhatók le:

$$X(t) = T(t) + S(t) + \epsilon(t), \tag{2.1}$$

ahol $T(t)$ egy lassú időbeli változást, azaz a trendet írja le, $S(t)$ az idő periodikus függvénye, amely a szezonális ciklikusságot jelenti, $\epsilon(t)$ pedig a véletlen komponens.

2.2. Regresszió

A regresszió egy statisztikai módszer, amely segítségével változók közötti összefüggéseket vizsgálunk. A megfigyelt mennyiségeket nevezzük magyarázó változóknak, ezeket a valószínűségi változókat jelöljük $X_1, X_2, ..., X_k$-val. Y-al pedig a velük feltehetően összefüggésben álló mennyiséget jelöljük, melyet válasznak nevezünk. A változók közötti kapcsolatot szeretnénk a megfigyelt adatok alapján minél pontosabban közelíteni.

$$\begin{matrix} Y \\ \\ \begin{bmatrix} y_1 \\ y_2 \\ .. \\ y_n \end{bmatrix} \end{matrix} \quad <= \quad \begin{matrix} X_1 & X_2 & X_3 & \dots & X_k \\ \\ \begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1k} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & x_{n3} & \dots & x_{nk} \end{bmatrix} \end{matrix}$$

ahol az x_{ij}-k és y_i-k ($i = 1, .., n$, $j = 1, .., k$) a megfigyelt értékek.
Mivel a dolgozatban sztochasztikus kapcsolatokat fogunk vizsgálni, ezért véletlen hiba is fellép. Ezeket összegezve, az $Y = f(X_1, X_2, ..., X_k) + \epsilon$ regressziós egyenletet kapjuk, ahol f egy k változós függvény, ϵ pedig független értékű zaj, melynek várható értéke 0, szórásnégyzete $< \infty$.

2.2.1. Definíció. *Egy többdimenziós regressziós modellt lineárisnak nevezünk, ha az egyes y_i értékek közelítése felírható mint $y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + ... + \beta_k x_{i,k}$ azaz a magyarázó változók β valós együtthatóira nézve lineáris. β_0 a konstans hatás.*

Leggyakrabban a β együtthatókat akarjuk becsülni és a segítségükkel előrejelzést adni y-ra az aktuális x-ek esetén.

Írjuk fel mátrixos alakban a többdimenziós lineáris modellt. Ahhoz, hogy β_0-t is kezeljük a többi együtthatóval együtt, vegyük a $B = (\beta_0, \beta_1, ..., \beta_k)$ $k+1$ hosszú vektort, és a magyarázó változók vektorához vegyük hozzá X_0-ként a csupa 1-esekből álló oszlopvektort. Így kapunk egy $n \times (k+1)$ méretű X mátrixot.
Ezek alapján a mátrixegyenlet a következő:

$$Y = XB + \epsilon \tag{2.2}$$

Átrendezve az egyenletet azt kapjuk, hogy $Y - XB = \epsilon$. Ez mutatja, hogy a megfelelő együtthatókkal vett magyarázó változók lineáris kombinációja csak a zajban tér el a választól.

2.2.2. Definíció. *Predikciónak nevezzük a regresszióval becsült együtthatókkal előállított* $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + ... + \hat{\beta}_k x_{i,k}$ *egyenletet.*

Predikciós hiba

2.2.3. Definíció. *A válasz és predikciójának a különbségét, azaz az* $\epsilon_i = y_i - \hat{y}_i$ *értéket reziduálisnak nevezzük.*

A reziduálisok négyzetösszege i=1,..,n-re adja meg a leggyakrabban használt predikciós hiba értékét:

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i,1} - ... - \hat{\beta}_k x_{i,k})^2 = \| Y - XB \|^2 \tag{2.3}$$

2.2.1. Regressziós együtthatók becslése

Legkisebb négyzetes (OLS) becslés

A reziduális négyzetösszeget, azaz a predikciós hiba értékét minimalizáljuk.

$$\| Y - XB \|^2 = (Y - XB)^T (Y - XB) \tag{2.4}$$

ezt deriválva B szerint, 0-t szeretnénk kapni, tehát

$$\frac{\partial}{\partial B} (Y - XB)^T (Y - XB) = 0 \tag{2.5}$$

azaz

$$\frac{\partial}{\partial B} (Y - XB)^T (Y - XB) =$$
$$\frac{\partial}{\partial B} (Y^T Y - Y^T XB - B^T X^T Y + B^T X^T XB) =$$
$$-2X^T Y + 2X^T XB$$

amelyből B-t kifejezve, a következő eredmény adódik:

$$B = (X^T X)^{-1} X^T Y \tag{2.6}$$

Ellenőrizhető, hogy ez valóban minimumhely.

Súlyozott legkisebb négyzetes (WLS) becslés

A legkisebb négyzetes becslés általánosítása a súlyozott legkisebb négyzetes becslés, ahol az adatpontokhoz hozzá vannak rendelve nemnegatív súlyfüggvény-értékek (az OLS becslés esetén minden súly 1). Ez a becslés akkor használatos, amikor egyes megfigyeléseket nagyobb preferenciával szeretnénk kezelni a regresszió illesztése során.
Illetve, olyankor is erre a módszerre van szükség, amikor az adatoknak nem ugyanaz a szórásnégyzetük (ez az ún. heteroszkedaszticitás). Hiszen, ha a lineáris regressziónál az ϵ_i hibák szórása nem egyezik meg, akkor ezek eltorzíthatják a minimumhelyét a reziduális négyzetösszegnek.
A képlet a következő:

$$\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2, \tag{2.7}$$

ahol a w_i-k a súlyok. Ezt kell minimalizálni, hasonlóan a legkisebb négyzetes becsléshez.

2.3. Általánosított lineáris modell (GLM)

2.3.1. Exponenciális eloszláscsalád

Az exponenciális eloszláscsalád olyan dominált valószínűségi eloszlások családja, amelyek sűrűségfüggvénye az alábbi alakú:

$$f_X(x|\vartheta) = h(x)e^{a(\vartheta)T(x)-b(\vartheta)}, \tag{2.8}$$

Ahol $b : \Omega \to \mathbb{R}$ és T egy statisztika. Az exponenciális eloszláscsaládba tartoznak például a normális, exponenciális, Poisson, gamma és a geometriai eloszlások is.

A lineáris regresszió általánosítása

Az általánosított lineáris modell a lineáris regresszió általánosítása, amelyben a válaszváltozó és magyarázó változók közötti függést egy ún. kapcsolati függvény (link function) határozza meg. Az Y feltételes várható értéke a következőképpen függ az X válaszváltozók értékétől: $g(E(Y|X)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$, ahol g a kapcsolati függvény.
A lineáris regresszió is egyik fajtája, ahol a változók normális eloszlásúak, a kapcsolati függvény pedig az azonosság. Viszont X és Y sok esetben nem lineárisan függ egymástól,

például amikor Y exponenciálisan növekszik X növekedésével, ezért van szükség a link függvényre.

2.3.2. Poisson eloszlás

2.3.1. Definíció. *Egy X valószínűségi változó pontosan akkor λ paraméterű Poisson eloszlású, ha*

$$P(X = k) = \frac{\lambda^k}{k!}e^{-\lambda}, \quad k = 0, 1, 2, ... \tag{2.9}$$

ahol $\lambda > 0$ konstans.

A Poisson eloszlású valószínűségi változó sajátossága, hogy a várható értéke és szórásnégyzete megegyezik, azaz $E(X) = D^2(X)$. A Poisson eloszlás például események adott idő alatti bekövetkezéseinek számát fejezheti ki.

2.3.3. Negatív binomiális eloszlás

2.3.2. Definíció. *Egy X valószínűségi változó pontosan akkor r rendű, p paraméterű negatív binomiális eloszlású, ha*

$$P(X = k) = \binom{k-1}{r-1}p^r(1-p)^{k-r}, \quad k = r, r+1, r+2... \tag{2.10}$$

ahol $0 < p \leq 1, r \in \mathbb{N}$.
$E(X) = \frac{r}{p}, \quad D(X) = \frac{\sqrt{r(1-p)}}{p}$.

Míg a Poisson eloszlás sajátossága az volt, hogy a várható érték egyenlő a szórásnégyzettel, a negatív binomiális eloszlás szórásnégyzete nagyobb is lehet a várható értékénél.

2.3.4. Gamma eloszlás

Egy X valószínűségi változó pontosan akkor α rendű, λ paraméterű gamma eloszlású, ha sűrűségfüggvénye

$$f(x) = \frac{1}{\Gamma(\lambda)}\lambda^\alpha x^{\alpha-1}e^{-\lambda x} \quad x > 0 \tag{2.11}$$

ahol α, $\lambda > 0$ és $\Gamma(\lambda) = (\lambda - 1)!$.

2.3.5. Loglineáris modell

A loglineáris, vagy másnéven Poisson modell egy fajtája az általánosított lineáris modelleknek. Tipikusan akkor használatos, amikor a válaszváltozó darabszámot jelöl, hiszen ekkor lehet a feltételezett eloszlás a Poisson eloszlás. Legyenek az $Y_1, Y_2, ..., Y_k$ egymástól független megfigyelések, melyekre $Y_i \sim \text{Poi}(\lambda_i)$. Vegyük kezdetben a $\lambda_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + ... + \beta_k x_{i,k}$ modellt minden $i = 1, .., n$-re, azaz legyen g az identitás. A problémája a modellnek az, hogy baloldalon egy nemnegatív szám kell álljon, hiszen a darabszám várható értéke nem lehet negatív, miközben a jobboldal negatív is lehet. Erre egy természetesen adódó megoldás az, hogy a kapcsolati függvény legyen a természetes logaritmus. Tehát

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + ... + \beta_k x_{i,k} \tag{2.12}$$

Ebből pedig

$$\lambda = e^{\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + ... + \beta_k x_{i,k}} \tag{2.13}$$

Így, egy egységgel növelve egy β_i-hez tartozó magyarázó változót, a várható érték e^{β_i}-szeresére nő. Az előnye tehát a log kapcsolati függvénynek az, hogy additív helyett multiplikatív hatása van a változóknak, amely jelllemző a számadatokra.

2.3.6. Regressziós együtthatók becslése

Maximum likelihood becslés

Az együtthatók becslése a loglineáris modell esetén maximum likelihood becsléssel történik. A likelihood függvény n db független Poisson megfigyelésre az 2.3.1 definícióban megadott valószínűségeknek a szorzata, azaz

$$L(\beta_0, \beta_1, ..., \beta_k, y) = \prod_{i=1}^{k} \frac{\lambda_i^{y_i}}{y_i!} e^{-\lambda_i} \tag{2.14}$$

ahol $\log(\lambda_i)$ a $\beta_0, \beta_1, ., \beta_k$ és az $x_{i,1}, x_{i,2}, ..., x_{i,k}$ lineáris függvényeként van kifejezve az 2.12 képletnek megfelelően. Ha $x_{i,0}$-t 1-nek vesszük minden i-re, logaritmálva a fenti egyenletet, a következőt kapjuk:

$$l(\beta_0, \beta_1, ..., \beta_k, y) = \sum_{i=1}^{n} y_i \log \lambda_i - \lambda_i - \log y_i! = \sum_{i=1}^{n} y_i (\sum_{j=0}^{k} \beta_j x_{i,j}) - e^{\sum_{j=0}^{k} \beta_j x_{i,j}} - \log y_i! \tag{2.15}$$

A maximum likelihood becsléssel maximalizálni akarjuk a likelihood függvényt. Tehát az együtthatók értékei a következő egyeneletrendszer megoldásai:

$$0 = \frac{\partial l}{\partial \beta_p} = \sum_{i=1}^{n} x_{i,p}(y_i - e^{\sum_{j=0}^{k} \beta_j x_{i,j}}) \tag{2.16}$$

minden $p = 1, ..., k$-ra, amelynek ugyan nincs zárt alakban megadható megoldása, de numerikus módszerekkel könnyen közelíthető.

Newton-módszer

Az $f(x) = 0$ egyenlet megoldása a feladat. Feltesszük, hogy létezik megoldás, és, hogy már ismert egy x_n közelítése. Ekkor definiáljunk egy újabb x_{n+1} közelítést a következőképpen:

$$x_{n+1} := x_n - \frac{f(x_n)}{f'(x_n)}, \quad n = 0, 1, 2, ... \tag{2.17}$$

2.3.3. Tétel. *Ha f kétszer folytonosan differenciálható, f-nek van x^* gyöke egy (a, b) intervallumban és $f'(x^*) \neq 0$, akkor a Newton-módszer konvergál minden, az x^* pontos megoldáshoz elég közeli x_0 kezdeti közelítés esetén, azaz létezik olyan $0 < C < 1$, hogy $| x_{n+1} - x^* | \leq C \cdot | x_n - x^* |$, ahol $n = 0, 1, 2, ...$*

Tehát ezzel a módszerrel közelíthető a megoldás, majd ellenőrizendő, hogy ez tényleg maximumhely.

2.3.7. Túlszóródás

A Poisson eloszlás egyedi tulajdonsága az, hogy a várható értéke egyenlő a szórásnégyzetével, azaz $E(X) = D^2(X)$.
Sok esetben azonban, olyan adatoknak, amelyek első ránézésre Poisson eloszlásúaknak feltételezhetőek, a szórásnégyzete nagyobb a várható értéknél. Ezt nevezzük túlszóródásnak (overdispersion). Erre egy példa a John Hinde, Clarice G.B. Demetrio: Overdispersion jegyzetében található ([15]), ahol az éves légi balesetekre alkalmazták a Poisson modellt.
A dolgozatban a heti lebontású halálozási számoknak a várttól vett különbségét fogjuk a koronavírus halálokkal modellezni és erre fogjuk kiszámolni, hogy túlszóródás lép fel ([9] alapján).

A szóródási paramétert a következő képlet adja:

$$\hat{\Phi} = \sum_i \frac{(y_i - \hat{\mu}_i)^2/\hat{\mu}_i}{n-p} \tag{2.18}$$

azaz szummázzuk a négyzetes eltérés és várható érték hányadosát és osztunk $n-p$-vel, ahol n a megfigyelések száma, p értéke pedig a becsült paraméterek száma. Megjegyezzük, hogy a $\hat{\Phi}$ értéke egynél kisebb szám is lehet. Ez az alulszóródás. Ez a jelenség viszont sokkal ritkább mint a túlszóródás.
A túlszóródás figyelmen kívül hagyásának következménye a modell hibás elemzése. A standard hibára és a szórásra rossz eredményeket kapunk, amelyek befolyásolhatják az egyes változók szignifikanciáját a modellben.
A Poisson modellben a túlszóródás kezelésére két lehetőség is van: a kvázi-Poisson modell, vagy a negatív binomiális modell alkalmazása.

2.3.8. Kvázi-Poisson modell

A kvázi-Poisson modell egy becslési módszer, amely nem egy valódi eloszlás. Míg a Poisoon modell egyenlőséget tételez fel a szórásnégyzet és a várható érték között, a kvázi-Poisson modellben ez az összefüggés lineáris, $D^2(Y) = \Phi E(Y)$, ahol Φ a túlszóródási paraméter. A Poisson modellhez hasonlóan maximum likelihood becsléssel történik az együtthatók becslése, a túlszóródási paraméter pedig a modellből van számolva a Túlszóródás alfejezetben (2.3.7) felírt képlettel. Tehát ugyanazokat a regressziós együtthatókat kapjuk, mint a Poisson modellben, viszont a standard hibák nagyobbak.

2.3.9. Negatív binomiális modell

A negatív binomiális eloszlás egy Poisson-gamma keverék, ahol Poisson eloszlású valószínűségi változóink vannak, a paraméterre pedig gamma eloszlást tételezünk fel. Legyen X egy Poisson eloszlású valószínűségi változó, λ paraméterrel.

$$P(X = x|\Theta = \theta) = \frac{e^{-\theta}\theta^x}{x!} \quad x = 0, 1, 2, ... \tag{2.19}$$

Tegyük fel, hogy Θ Gamma eloszlást követ, α renddel és λ paraméterrel. Ekkor θ sűrűségfüggvénye

$$f(\theta) = \frac{1}{\Gamma(\lambda)}\lambda^\alpha \theta^{\alpha-1} e^{-\lambda\theta} \quad \theta > 0. \tag{2.20}$$

Ebből

$$P(X = x|\Theta = \theta)f(\theta) = \frac{e^{-\theta}\theta^x}{x!}\frac{1}{\Gamma(\alpha)}\lambda^\alpha\theta^{\alpha-1}e^{-\lambda\theta}. \tag{2.21}$$

Innen

$$\begin{aligned}
P(X = x) &= \int_0^\infty P(X = x|\Theta = \theta)f(\theta)d\theta \\
&= \int_0^\infty \frac{e^{-\theta}\theta^x}{x!}\frac{1}{\Gamma(\alpha)}\lambda^\alpha x^{\alpha-1}e^{-\lambda\theta}d\theta \\
&= \int_0^\infty \frac{\lambda^\alpha}{x!\Gamma(\alpha)}\theta^{x+\alpha-1}e^{-(\lambda+1)\theta}d\theta \\
&= \frac{\lambda^\alpha}{x!\Gamma(\alpha)}\frac{\Gamma(x+\alpha)}{(\lambda+1)^{x+\alpha}}\int_0^\infty \frac{(\lambda+1)^{x+\alpha}}{\Gamma(x+\alpha)}\theta^{x+\alpha-1}e^{-(\lambda+1)\theta}d\theta \\
&= \frac{\lambda^\alpha}{x!\Gamma(\alpha)}\frac{\Gamma(x+\alpha)}{(\lambda+1)^{x+\alpha}} \\
&= \frac{\Gamma(x+\alpha)}{\Gamma(x+1)\Gamma(\alpha)}\left(\frac{\lambda}{\lambda+1}\right)^\alpha\left(\frac{1}{\lambda+1}\right)^x \\
&= \binom{x+\alpha-1}{x}\left(\frac{\lambda}{\lambda+1}\right)^\alpha\left(\frac{1}{\lambda+1}\right)^x \quad x = 0, 1, 2, ...
\end{aligned} \tag{2.22}$$

$r = \alpha$ és $p = \frac{\lambda}{\lambda+1}$ helyettesítéssel látható, hogy ez az r rendű, p paraméterű negatív binomiális eloszlás képlete.

A negatív binomiális modell az exponenciális eloszláscsaládhoz tartozik. A kapcsolati függvény, a Poisson modellhez hasonlóan a logaritmus, az együtthatók becslése pedig itt is a maximum likelihood becsléssel történik. A Poisson modellel szemben, megengedi a várható értéknél nagyobb szórásnégyzetet, hiszen a következő összefüggést tételezi fel közöttük:

$$D^2(Y) = E(Y) + \kappa E^2(Y), \tag{2.23}$$

ahol $\kappa > 0$.

A likelihood függvény az

$$\begin{aligned}
l(\beta_0, \beta_1, ..., \beta_n, \alpha) = &\sum_{i=1}^n \left[\left(\sum_{j=0}^{y_i-1} ln(j + \alpha^{-1})\right)\right] \\
&- \sum_{i=1}^n \left[ln(\Gamma(y_i + 1) - (y_i + \alpha^{-1})ln(1 + \alpha\lambda_i) + y_i ln(\lambda_i) + y_i ln(\alpha)\right],
\end{aligned}$$

ahol

$$\lambda_i = e^{\beta_0+\beta_1 x_{i,1}+\beta_2 x_{i,2}+...+\beta_k x_{i,k}} \tag{2.24}$$

és α a Gamma eloszlás paramétere.
A függvény minimalizálásához, a következő egyenletrendszer megoldásai lesznek a becslések:

$$0 = \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{x_{i,j}(y_i - \lambda_i)}{1 + \alpha\lambda_i}, \tag{2.25}$$

minden $j = 1, ..., k$-ra és

$$0 = \frac{\partial l}{\partial \alpha} = \sum_{i=1}^{n} \left[\alpha^{-2} \left(\ln(1 + \alpha\lambda_i) - \sum_{j=0}^{y_i - 1} \frac{1}{j + \alpha^{-1}} \right) + \frac{y_1 - \lambda_i}{\alpha(1 + \alpha\lambda_i)}) \right] \tag{2.26}$$

2.4. Simítás

Idősorok elemzésénél gyakran egy simított görbe jobban mutatja a trendeket és mintázatokat. Ezért, főleg a nagyobb felbontású adatsorokon használják, például napi adatoknál. A dolgozatban heti adatokra fogjuk használni, amikor a koronavírus előtti időszak átlagértékeit számoljuk.

Polinomiális regresszió

A polinomiális regresszió a lineáris regresszió egyik formája, ahol az X magyarázó változó az idő, az Y válaszváltozó a simítandó idősor és a köztük levő összegfüggés egy n-edfokú polinommal van modellezve.

Egyszerű mozgóátlag

A mozgóátlag egy statisztikai mutató, amelyben egy adatsor értékeihez az adott érték és meghatározott számú körülötte lévő elem átlagát rendeljük hozzá. Ez a módszer idősoros adatok simítására használt, például a heti ciklus kiküszöbölésére. Képlete:

$$\hat{y}_t = \frac{y_{t-k} + ... + y_{t-1} + y_t + y_{t+1} + ... + y_{t+k}}{2k + 1}, \tag{2.27}$$

ha a tagok száma páratlan és

$$\hat{y}_t = \frac{\frac{y_{t-k}}{2} + ... + y_{t-1} + y_t + y_{t+1} + ... + \frac{y_{t+k}}{2}}{2k}, \tag{2.28}$$

ha a tagok száma páros.

LOESS simítás

A LOESS (Locally estimated scatterplot smoothing) egy lokális regresszió, amely általánosítása a mozgóátlagnak és a polinomiális regressziónak.
Az algoritmus a következő:
Megadunk egy $0 < s \leq 1$ paramétert, amely azt jelöli, hogy hány százalékát használja a megfigyeléseknek a lokális regresszió. Ha n a megfigyelések száma, akkor a $k = \lfloor ns \rfloor$, x_0-hoz legközelebbi megfigyelés alkotja az x_0 lokális környezetét.
Ezután, egy harmadfokú súlyfüggvénnyel súlyok vannak hozzárendelve az x_0 pont kiszámításához használt értékeknek. Az i. pont súlya $w_i = (32/5)(1 - (d_i/D)^3)^3$, ahol D a legnagyobb távolság x_0 környezetében és d_i a távolsága az i. pontnak.
Végül lokális súlyozott regresszióval megkapjuk a simított értékeket.

2.5. Konfidencia-intervallum

2.5.1. Definíció. *Legyen $X = (X_1, ..., X_n)$ minta F_ϑ eloszlásból, ahol ϑ valós paraméter. Azt mondjuk, hogy a $(T_1(X), T_2(X))$ statisztikapár $1-\alpha$ megbízhatósági szintű konfidencia-intervallum ϑ-ra, ha*

$$P_\omega(T_1(X) < \omega < T_2(X)) = 1 - \alpha, \tag{2.29}$$

A konfidencia-intervallumot gyakran úgy konstruáljuk, hogy a pontbecslés köré szimmetrikus intervallumot veszünk fel. Például, ha $X = (X_1, ..., X_n)$ minta független, ahol minden X_i normális eloszlású m várható értékkel és σ szórással, és σ ismert, akkor

$$(T_1(X), T_2(X)) = \overline{X} \pm \frac{z_{1-\frac{\alpha}{2}}\sigma}{\sqrt{n}}, \tag{2.30}$$

ha pedig σ ismeretlen, akkor

$$(T_1(X), T_2(X)) = \overline{X} \pm \frac{t_{n-1,1-\frac{\alpha}{2}}s_n^*}{\sqrt{n}}, \tag{2.31}$$

ahol $\overline{X}$ a mintaátlag, s_n^* a korrigált tapasztalati szórás, azaz $s_n^* = \sqrt{\frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n-1}}$, z pedig a standard normális, t_{n-1} az $n-1$ szabadsági fokú Student eloszlás megfelelő kvantiliseit jelölik.

2.5.1. Student eloszlás

2.5.2. Definíció. *Legyenek $z_0, z_1, ..., z_n$ függetlenek, standard normális eloszlásúak. Akkor $x := \frac{z_0}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} z_i^2}}$ eloszlása n szabadsági fokú Student eloszlás.*

Egymintás t-próba

2.5.3. Definíció. *Legyenek $X_0, X_1, ..., X_n$ függetlenek, normális eloszlásúak m várható értékkel és σ szórással és legyen m_0 adott. Ekkor a $H_0 : m = m_0$, $H_1 : m \neq m_0$ hipotézis mellett akkor utasítjuk el a nullhipotézist, ha $\mid T(x) \mid > t_{n-1,\frac{\alpha}{2}}$, ahol $T(x) = \sqrt{n}\frac{\overline{X}-m_0}{s_{n*}}$, és $t_{n-1,\frac{\alpha}{2}}$ az $n-1$ szabadsági fokú Student eloszlás megfelelő kvantilisét jelöli.*

2.6. Harmadfokú spline interpoláció

A gyakorlatban sok olyan eset van, amikor adottak pontok a síkon (vagy térben), ezek az interpolációs alappontok és mindegyikhez tartozik egy érték. A feladat az, hogy keressünk olyan, előírt tulajdonságokkal rendelkező függvényt, amely az adott interpolációs pontokban az adott értékeket veszi fel. A harmadfokú spline interpoláció egy módszer arra, hogy ilyen görbét keressünk.

2.6.1. Definíció. *Az $f : [a, b] \to \mathbb{R}$ függvényt harmadfokú és másodrendű spline függvénynek nevezzük, ha szakaszonként legfeljebb harmadfokú polinom és az illesztési pontokban kétszer folytonosan deriválható.*

Ha adva vannak az $a = x_0 < x_1 < ... < x_n = b$ interpolációs alappontok és a hozzájuk rendelt $f_0, f_1, ..., f_n$ értékek, akkor definiáljuk a $f_0', f_1', ..., f_n', f_0'', f_1'', ..., f_n''$ első- és másodrendű deriváltértékeket. Minden $[x_k, x_{k+1}]$ részintervallumon $f(x)$ alakja $f(x) = A_k + B_k x + C_k x^2 +_k Dx^3$. Illetve, az alappontokban az elsőrendű deriváltak értékei meg kell egyezzenek. Előírva, hogy a másodrendű deriváltak a végpontokban 0-k legyenek, az összes egyenletet felírva kapunk egy egyértelműen megoldható egyenletrendszert. Ezt megoldva megkapjuk f-et.

3. fejezet

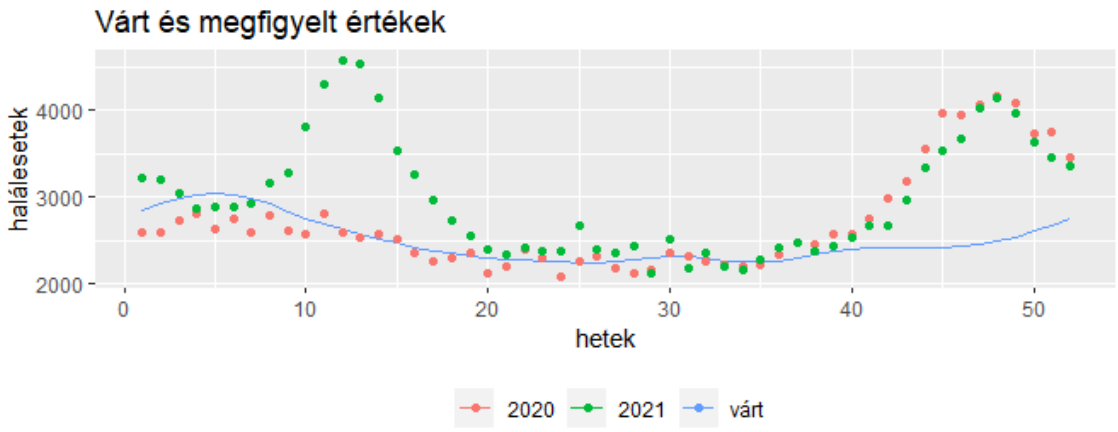
Adatelemzés

Ebben a fejezetben az előző részben bevezetett módszereket fogom alkalmazni egy valós probléma vizsgálatára: a koronavírus magyarországi hatásának elemzésére. Ezt, mint ahogy a bevezetésben is írtam, a halandósági adatokból fogom megtenni. A számolásokat Rstudioban, az R programozási nyelvben végeztem. Az ötleteim egy részét Ferenci Tamás weboldaláról vettem ([7]), aki egy teljes honlapot dolgozott ki a koronavírus elemzésének céljából. Többek között ő is foglalkozik a többlethalálozással, de sok más szempontból is vizsgálja a járványt ([8]).

3.1. Többlethalálozás

A többlethalálozás az összes haláleset és a "várt" halálesetek számának különbsége. A "várt" esetszámnak, más néven alapadatoknak a $2015 - 2019$ évek haláleseteinek átlagát veszem de a későbbiekben egy másik módszert is alkalmazok ennek kiszámítására.
2020-ban 141002 volt az összhalálozás, a $2015 - 2019$-es évek átlaga pedig 130214, tehát a nyers többlethalálozás a 2020-as évben 10788. A továbbiakban ezt a kezdeti becslést finomítjuk és részletezzük.
Ábrázoljuk a várt halálozási számokat és a megfigyelt értékeket a 2020 és 2021-es évben. Azonban, mivel az adatok heti felbontásban vannak megadva, és így sok véletlen hatást is tartalmaznak, a várt halálesetek számát simítsuk először. Ezt az Rstudio loess függvényével tesszük. Mivel éves adataink vannak, először megháromszoroztuk a várt értékeket (ezek a 2015-19 közötti évek heti átlagai), és a középső szakaszt simítottuk, hogy december és január között ne legyen szakadás. A simítási paramétert $s = 0,1$-nek választjuk, azaz a lokális regresszió minden megfigyeléshez a legközelebbi $\lfloor 52 \times 0,1 \rfloor = 5$ megfigyelést

használja.



3.1. ábra. A 2015-2019-es évek alapján kiszámolt várt értékeket ábrázoljuk a 2020 és 2021-es megfigyelt értékek mellett. Nagyon jól látható a járvány második, harmaik és negyedik hulláma, amikor a halálesetek száma jóval nagyobb, mint a várt értékek.

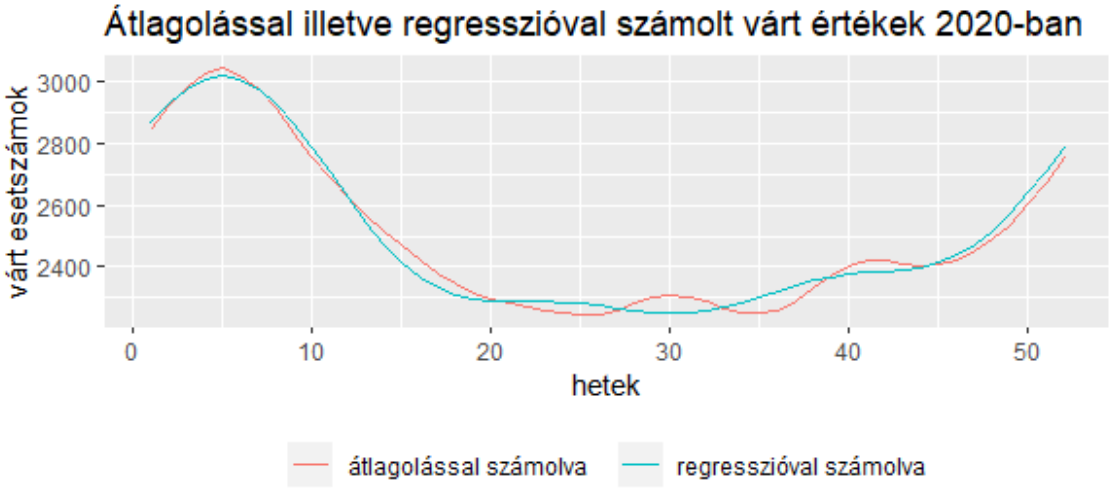
Egy második módszer az alapadatok kiszámítására

A fentiekben bemutatott nemparméteres simítás helyett pontosabb közelítések kaphatóak, ha egy olyan modellt alkalmazunk, amely paraméteresen figyelembe veszi a trendet és a periodikus komponenst (22). Ehhez a következő lineáris regressziót írjuk fel a heti adatokra, megengedve a trendet és a periodikus tagokat is (téli/nyári csúcsok jellemzőek):

$$M_h = \beta_0 + \beta_1 h + \sum_{k=1}^{3} [\beta_{2,k} \sin(\frac{2\pi kh}{52,18}) + \beta_{3,k} \cos(\frac{2\pi kh}{52,18})] \tag{3.1}$$

A beták a regressziós együtthatók, a h a heti indexet jelöli, amely 1-től 261-ig fut. A képlet második részében, felhasználva, hogy minden függvény felírható Fourier sorként, a szezonális komponenst modellező függvényt is így közelítettük. Az $52,18$-as osztó az egy évben levő átlagos hét-számot jelöli. A k három értéke pedig az évenkénti, 6 hónapos és 4 hónapos periodicitást viszi be a modellbe. A tagok számát, azaz k értékeinek számát az alapján állítottuk be 3-ra, hogy megnéztük, hogy hány tagig kapunk szignifikáns együtthatókat. A harmadik szinusos tag még $6 \cdot 10^{-5}$ körüli szignifikanciájú, ezért benne hagyjuk a modellben, a negyedik viszont már kevésbé szignifikáns, a koszinuszos tagra pedig $0,1$-nél is nagyobb a p-érték, ezért a sorfejtést csak az első három tagig írjuk fel.
Ez a modell viszont a hét sorszámának együtthatójára $0,88$ szignifikanciaszintet ad, ami

azt jelenti, hogy nincs szignifikáns hosszú távú trend, mint például az egészségügy fejlődése. Tehát a modellt e nélkül a tag nélkül írjuk fel.
A regressziós modell és az R predict függvénye segítségével meghatároztuk a 2020-as heti "várt" halálesetek számát. Ezeket összehasonlítottuk az átlagolással kapott értékekkel a 3.2 ábrában.



3.2. ábra. Simított átlagolással, illetve 3.1 képlet szerinti regresszióval kiszámított alapadatok vizualizációja a 2020-as évben. A két módszerrel számolt értékek kis mértékben térnek el egymástól, legfőbbképpen a nyári időszakban.

Mindkét esetben jól látható, hogy télen és kora tavasszal a legnagyobbak a halálozási számok. A legkisebb értékek pedig nyáron fordulnak elő.
A két módszerrel kiszámolt "várt" halálesetek számai 2000-3000-es nagyságrendeknél tizes nagyságrendben különböznek. Ez 2% körüli különbségeket jelent.
Az év elején és végén, a három hullám időszakában mindkét módszer nagyon hasonló várt értékeket eredményez, tehát a továbbiakban nem nagy eltérést okoz, ha csak az egyikkel számolunk.

3.2. Lineáris modell

Magyarország sajnos az élen szerepel európai viszonylatban a koronavírus okozta halálozások terén. Gyakran olvasható az, hogy azért ennyire nagy a koronavírus okozta halálok száma, mert mindent annak jelentenek be, más országokkal szemben, ahol ez nem így történik. Ezért a következő a fejezetekben különböző regressziós modellekkel azt vizsgálom,

mekkora szerepet játszott a koronavírus a többlethalálozás növekedésében.
Az első és egyben legegyszerűbb modell a lineáris modell. A 12 cikkben hasonlóan lineáris modellel vizsgálták az Egyesült Államokban, államok szerint a lehetésges aluldetektálást. Vegyük a megfigyelt és várt halálozási számok különbségét mint válaszváltozó, és a koronavírusos halálokat mint magyarázó változó. Mindezt heti felbontásban, szeptembertől kezdődően, mert ekkor kezdett a vírus hatása erősebben érződni hazánkban. A vírus első (2020. márciusi) hulláma még nem okozott többlethalálozást a korai intézkedések következtében. Ezért a szeptembertől bekövetkező második hullámtól kezdődően vizsgáljuk az adatokat. A modell a következő lesz:

$$T(i) = \beta_1 + \beta_2 C(i), \tag{3.2}$$

ahol $T(i)$ a különbség (többlet) az i. héten, $C(i)$ pedig az i. heti koronavírusos halálozási számot jelöli. Az együtthatókat legkisebb négyzets becsléssel kiszámolva (2.2.1), $\beta_1 = 80,338$, $\beta_2 = 0,91005$. (Ezen együtthatókról egy összesített táblázat a 3.5 alfejezetben található.) Ez azt jelenti, hogy koronavírus nélkül 80 körüli lett volna a többlethalálozás, a $\beta_2 = 0,91005$ pedig azt jelentené, hogy kissé felül volt detektálva a koronavírus. Viszont ha t-próbával (2.5.3) kiszámoljuk, hogy ez az érték szignifikánsan eltér-e 1-től, akkor azt kapjuk, hogy nem ($\mid T \mid < t_{68,\frac{\alpha}{2}}$, ahol α-t $0,02$-nek választjuk). Tehát a koronavírusnak, mint halálozási oknak bejelentett halálok száma nem szignifikánsan tér el a ténylegesen a vírus miatt elhunytak számától. Vizsgáljuk meg egy második modellel is az adatokat.

3.3. Poisson modell

Mivel a halálozási adatok esetszámot jelölnek, egy természetes ötlet az, hogy a Poisson modellel becsüljünk. Ehhez az Rstudio glm nevű függvényét használjuk, ami kiszámolja maximum likelihood becsléssel az együtthatókat.
Azt szeretnénk látni, hogy mennyire voltak szignifikánsak a covid okozta halálok a többlethalálozás mérésénél, tehát hasonlóan a lineáris modellhez, a többlethalálozás lesz a válaszváltozónk, a covid halálok vektora pedig a magyarázó változó.
Fontos megjegyezni, hogy a lineáris modellben 2020 szeptemberétől kezdődő adatokkal számoltunk, hiszen ettől a hónaptól növekedett meg a halálok száma a várthoz képest. Tehát a számok 2020. szeptember és 2021 december közötti megfigyelések. Ebben az időszakban azonban 8 héten negatív volt a többlethalálozás értéke. Az egyik ilyen periódus 2021 februárja, amikor az influenza miatt a várt halálozás magasabb volt, viszont 2021-ben

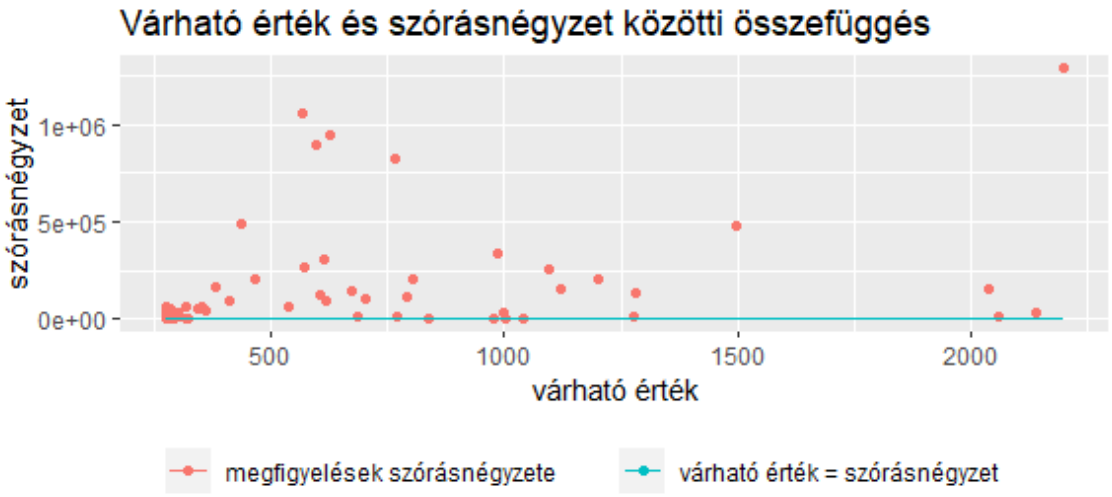
nagyon kicsi volt az influenza hatása, illetve korlátozások is hatályban voltak. A másik ilyen időszak ugyanebben az évben a késő nyár, kora ősz volt, amikor újra nagyon sok korlátozás lett behozva Magyarország területén. Mivel a Poisson modell csak nemnegatív értékeket enged, ezért ezt a nyolc hetet kihagytuk a számolásból.

Eredményül azt kaptuk, hogy egy covid halál $e^{0,00182}$-ször, azaz $1,00182$-ször növelte a többlethalálozási számot. Ez újra alátámasztja, hogy a bejelentett covid halálok száma reális, hiszen t-próbával számolva, $1,00182$ nem szignifikánsan nagyobb 1-nél, 1000 bejelentett covid halálozásnál 1001,82 tényleges halált jelent.

3.4. Túlszóródás kezelése

A Poisson eloszlás egy egyedi tulajdonsága az, hogy a várható értéke egyenlő a szórásnégyzetével, azaz $E(X) = D^2(X) = \lambda$. Az előbbiekben feltettük, hogy a megfigyeléseink Poisson eloszlásúak, és a loglineáris modellel becsültünk. Most viszont ellenőrizzük, hogy teljesül-e, hogy a várható érték megegyezik a szórásnégyzettel.

Ehhez rajzoljuk ki a megfigyelések várható értékeit és szórásnégyzeteit (3.3 ábra), ahol a várható értékek a Poisson modellből kapott együtthatókkal kiszámolt értékek (a 2.13 képlet alapján), a szórásnégyzetek pedig a várható értékeknek a megfigyelt értékektől vett különbségeinek a négyzetei.



3.3. ábra. Ábrázoltuk a várható értéket a szórásnégyzet függvényében. A szórásnégyzet a legtöbb esetben jóval nagyobb a várható értéknél. Az egyenes az $x = y$ egyenletű egyenes, tehát ha nem lenne túlszóródás, a pontok ezen az egyenesen kellene elhelyezkedjenek.

A 3.3 ábrán látható tehát, hogy a szórásnégyzetek nagyobbbak a várható értékeknél. Ellenőrizzük ezt matematikailag is. A szóródási paraméter 2.18 képletébe

$$\hat{\Phi} = \sum_i \frac{(y_i - \hat{\mu}_i)^2/\hat{\mu}_i}{n-p} \tag{3.3}$$

behelyettesítve a megfigyelt értékeket, Poisson regresszió által megkapott együtthatókból kiszámolt várt értékeket és p-t 1-nek véve, mert nincs hosszú távú trend, 250-et kapunk. Tehát egyértelműen nagy a túlszóródás.

Ekkor viszont a Poisson modell helyett más modellt fogunk alkalmazni.
Két lehetőség is van: a kvázi-Poisson és a negatív binomiális modell. A két modell között az az egyik jelentős különbség, hogy a kvázi-Poisson modellben feltesszük, hogy a variancia a várható érték egy lineáris függvénye, azaz $E(Y) = \mu$ és $D^2(Y) = \Phi\mu$, ahol $\mu > 0$ és $\Phi > 1$, míg a negatív binomiális modellben $E(Y) = \mu$ és $D^2(Y) = \mu + \kappa\mu^2$, ahol $\mu > 0$ és $\kappa > 0$.
Tehát, míg a kvázi Poisson modellben lineárisan függ a szórásnégyzet a várható értéktől, a negatív binomiális modell esetén ez az összefüggés négyzetes.

3.4.1. Kvázi-Poisson modell

Az egyik megoldás a túlszóródás figyelembe vételéhez tehát az, hogy a Poisson helyett a kvázi-Poisson modellt alkalmazzuk. Ilyenkor, a túlszóródási paramétert a modellből becsüljük a szóródási paraméter képletével (3.3). A válaszváltozókra azt feltételezzük, hogy $D^2(Y) = \Phi E(Y)$, ahol Φ a modellből becsült túlszóródási paraméter. A glm függvényt használjuk most is.
A kvázi-Poisson modellben az együtthatók ugyanazok lesznek mint a Poisson modellben, csak a standard hibák és a p értékek változnak. Ezen kívül, a modell 250 körülinek becsülte a túlszóródási paramétert, hasonlóan az előbbi számolásainkhoz. Egy összefoglaló táblázat a vizsgált modellek együtthatóiról a 3.5 alfejezetben található.

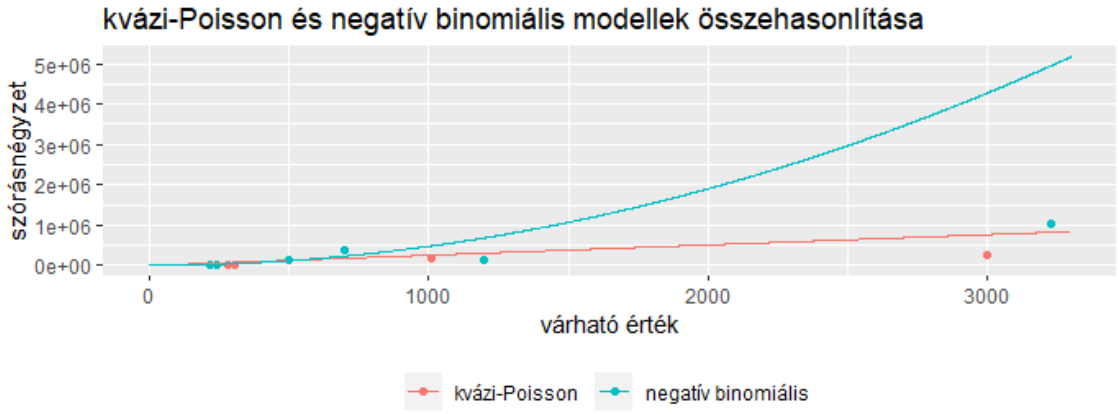
3.4.2. Negatív binomiális modell

Egy másik módszer a túlszóródás kezelésére a negatív binomiális modell alkalmazása.
A Poisson modellhez hasonlóan, most is a glm függvényt használjuk.
Ebben az esetben nem sokkal, de eltérő eredményeket kaptunk az együtthatók értékeire

(a 3.5 alfejezetben vannak az együtthatók összesítve), θ-nak pedig $2,1$-t ad a program (ahol $\theta = \frac{1}{\kappa}$ és κ a 2.23 egyenletben szereplő másodfokú tag együtthatója) .

3.4.3. A kvázi-Poisson és negatív binomiális modellek összehasonlítása

Láttuk, hogy a mindkét modell a túlszóródás kezelésére megfelelő, most pedig azt vizsgáljuk, hogy melyik modellt érdemesebb használni ebben az esetben.
Ehhez, felhasználva, hogy $\Phi = 250$ a kvázi-Poisson modellben és $\kappa = \frac{1}{\theta} = \frac{1}{2,1} = 0.476$ a negatív binomiális modellben, ábrázoljuk a szórásnégyzetet a várható érték függvényében. Azaz, ábrázoljuk a $D^2(Y) = \Phi\mu$ egyenest és $D^2(Y) = mu + \theta\mu^2$ másodfokú függvényt. Illetve, ugyanezen az ábrán tüntessük fel a reziduális négyzetösszegeket is a következőképpen: μ szerint csoportosítsuk az adatokat 6 kategóriába úgy, hogy mindegyikben ugyanannyi megfigyelés legyen. Majd vegyük egy kategórián belül a reziduális négyzetösszegek értékeinek átlagát. Ezt ábrázoljuk a 3.4 ábrán.



3.4. ábra. Az egyenes a kvázi-Poisson, a parabola a negatív binomiális modell alapján ábrázolt szórásnégyzet a várható érték alapján. A piros pontok a kvázi-Poisson, a kékek a negatív binomiális modell alapján kiszámolt reziduális négyzetösszegek 6 egyforma gyakoriságú kategóriába osztva a várható értékük szerint

A kvázi-Poissonnál a 6 kategória a következő: $0 < \hat{\mu}_i \leq 282$, $282 < \hat{\mu}_i \leq 305$, $305 < \hat{\mu}_i \leq 500$, $500 < \hat{\mu}_i \leq 700$, $700 < \hat{\mu}_i \leq 1010$, $1010 < \hat{\mu}_i$.
A negatív binomiális modell esetén a 6 kategória a $0 < \hat{\mu}_i \leq 214$, $214 < \hat{\mu}_i \leq 240$, $240 < \hat{\mu}_i \leq 500$, $500 < \hat{\mu}_i \leq 700$, $700 < \hat{\mu}_i \leq 1200$, $1200 < \hat{\mu}_i$.

Mindegyik kategóriához 10-11 megfigyelés tartozik.
Ez a módszer segíthet megfigyelni egy lineáris vagy négyzetes összefüggést a várható érték és a szórásnégyzet között. Látható, hogy a kisebb várható értékekre a negatív binomiális modell illeszkedik jobban, de összeségében az egyenesre illeszkednek jobban az adatok, azaz a kvázi-Poisson modellre.
Ez azt jelenti, hogy érdemes a kvázi-Poisson modellel számolni, és esetleg előrejelzés is ennek a modellnek az alapján adható.

3.5. A három modell összehasonlítása

Négy különböző modellt illesztettünk az adatokra. A Poisson modellt leszámítva (hiszen ott túlszóródás miatt a másik két modellt használtuk), ábrázoljuk egy táblázatban a kapott együtthatókat, annak érdekében, hogy jobban megfigyelhessük a különbségeket.

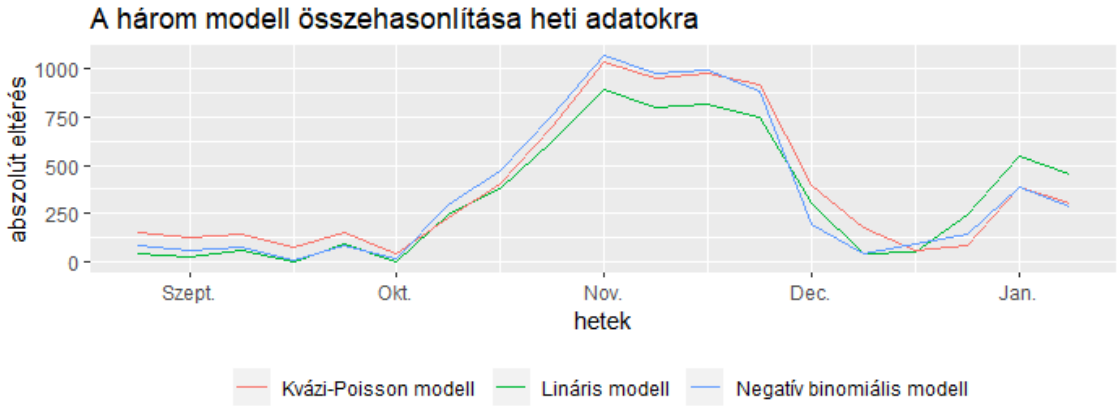
A modellek összehasonlítása			
Modell	Együtthatók (szabadtag, covid)	Szignifikancia szintek (szabadtag, covid)	p-értékek (szabadtag, covid)
Lineáris modell	80,33854 0,91005	69,99273 0,09262	0,255 1,3e-14
Kvázi-Poisson modell	272,46342 1,0011825	0,1541574 0,0001387	<2e-16 7,07e-12
Negatív binomiális modell	204,54737 1,0015593	0,1304126 0,0001648	<2e-16 <2e-16

3.1. táblázat. Összehasonlítjuk a lineáris, kvázi-Poisson és negatív binomiális modellel számolt együtthatókat, az együtthatók szignifikancia-szintjeit és p-értékeit.

A táblázatból látszik, hogy hasonló eredményeket kaptunk mindhárom esetben. A covid együtthatója egyik modellben sem tér el szignifikánsan 1-től, tehát a bejelentett covid halálok száma reális, nem volt sem felül-, sem alulbecsülve.

Láttuk az előbbi fejezetben (3.4.3), hogy a kvázi-Poisson és a negatív binomiális modellek közül a kvázi-Poisson kicsivel jobban illeszkedik a megfigyelt adatokra. Azonban,

mivel lineáris modellel is számoltunk, hasonlítsuk össze ezt is velük. Ehhez, ábrázoljuk az abszolút eltéréseket mindhárom esetben a második (2020 végi) hullámon.

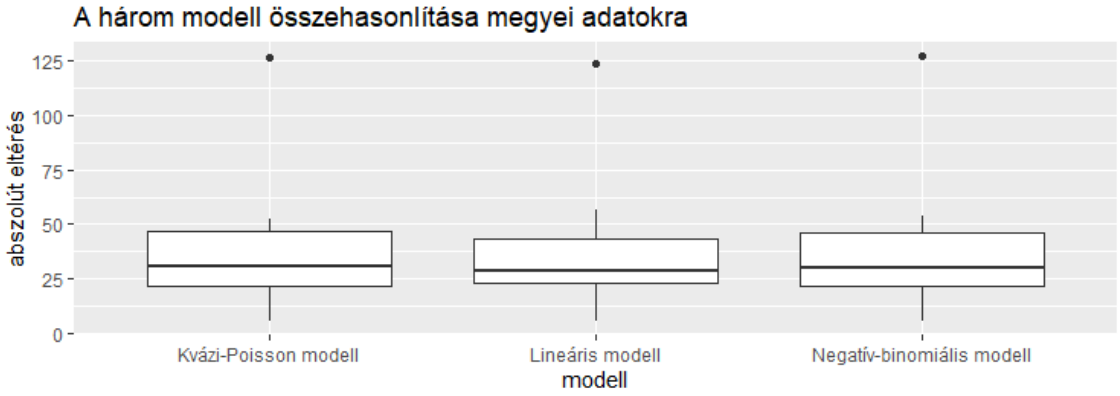


3.5. ábra. A koronavírus második hullámán vizualizáljuk a három modellel kiszámolt abszolút eltéréseket.

A három modellel hasonló eltéréseket kaptunk, a negatív binomiális modell összeségében kicsivel kisebb eltéréseket eredményez a kvázi-Poisson modellnél (a 3.4.3 alfejezettel összhangban), a lineáris modell azonban mindkettőnél enyhén kisebb különbségeket ad.

3.6. Megyei adatok elemzése

Hasonlóan a fenti alfejezetekhez, most a 2020-as megyei adatokat is vizsgáljuk a különböző regressziós modellekkel. A módszerek ugyanazok mint az előzőekben, azzal az eltéréssel, hogy minden modellben a megyék népességével súlyozunk. A halálozási arányok és a koronavírus okozta halálozási arányok százezer lakosra nézve vannak megadva (3) és a várt halálozási arányok a $2015 - 2019$-es évek átlagolásával vannak számolva. Tehát a modell a következő: A többlethalálozás (százezer lakosra vetítve) a válaszváltozó, a megyei koronavírusos halálok aránya (százezer lakosra nézve) pedig a magyarázó változó. A súlyok a megyénkénti lakosság számai.
Ábrázoljuk az abszolút eltéréseket (3.6). Azt várjuk el, hogy a modellek jósága az előbbiekhez hasonló legyen, hiszen más csoportosításban, de ugyanazokkal az adatokkal dolgozunk.



3.6. ábra. A három modellel kiszámolt abszolút eltérések a megyei adatokon. A vastagított szakaszok a mediánokat mutatják.

A három modellel ebben az esetben is -hasonlóan az előbbi 3.5 fejezethez- csak kissé eltérő eredményeket kaptunk. Az abszolút eltérések a lineáris modellben voltak a legkisebbek, a másik két modell pedig majdnem egyforma különbségeket eredményezett. Mindhárom esetben az egy-egy magas érték a Szabolcs-Szatmár megyei különbséget mutatja. Ez Magyarország legnépesebb megyéje, ehhez képest mégis kevés bejelentett koronavírus okozta halál következett itt be.

3.7. EXCESSMORT programcsomag

Az EXCESSMORT programcsomag az R programozási nyelv egy csomagja, amely pontosan a többlethalálozás vizsgálatára volt megírva Rolando J. Acosta1 és Rafael A. Irizarry biostatisztikusok által (6).
Az eredmények és ábrák egyszerűbb elemzése céljából előbb összesítem egy táblázatban a világjárvány 2020-21-ben bekövetkezett négy hullámát (28).

A koronavírus négy hulláma		
Hullám száma	Hullám kezdete	Regisztrált covid halálozás (fő)
Első hullám	2020. március	609
Második hullám	2020. szeptember	11589
Harmadik hullám	2021. február	17848
Negyedik hullám	2021. szeptember	9058

3.2. táblázat. A koronavírus négy magyarországi hulláma 2020-21-ben. Mindegyik hullám a következő kezdetéig tartott, a negyedik december 31-ig.

Az alapadatok kiszámítása

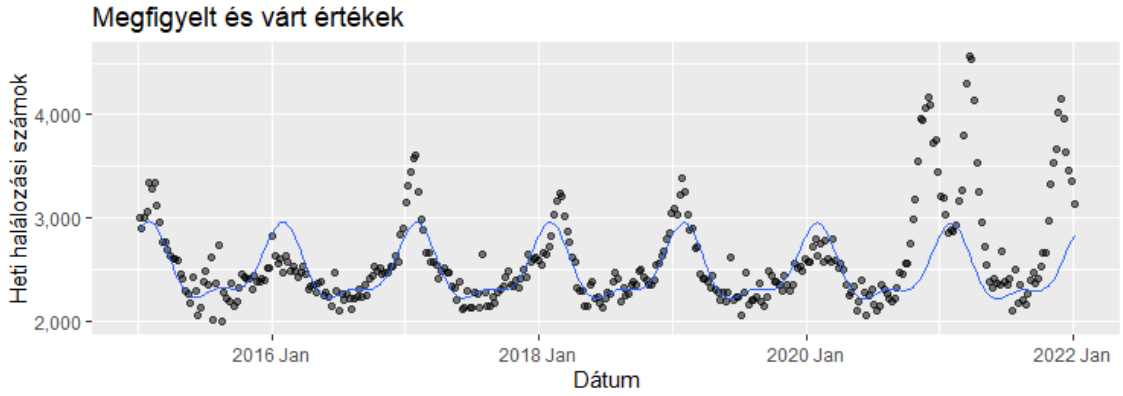
A modell az alapadatok kiszámításához (várt halálozás) a következő:

$$\mu_h = N_h e^{\alpha(h)+s(h)}, \tag{3.4}$$

ahol $\alpha(h)$ hosszú távú trendeket modellez (mint például az egészségügyi fejlődések az évek során), N_h a populáció a h. héten, $s(h)$ pedig a szezonális komponenst modellező harmonikus függvény. Mivel viszont csak 2015-2021-es adatokkal dolgozunk, $\alpha(h)$-t az idő lineáris függvényének tekintjük. $s(h)$-t harmonikus modellel számoljuk, hasonlóan a 3.1 alfejezetben írtakkal:

$$s(h) = \sum_{k=1}^{3} [\beta_{1,k} sin(\frac{2\pi kh}{52,18}) + \beta_{2,k} cos(\frac{2\pi kh}{52,18})]. \tag{3.5}$$

Az adatokat továbbra is heti felbontásban modellezzük.
A modell tehát hasonló a 3.1 alfejezet alapadatainak kiszámolásához használt modellhez. Abban tér el az utóbbitól, hogy offset-ként szerepel a népesség száma is. Ez hasznos, hiszen 2015 és 2021 között csaknem 150000-rel csökkent a népesség száma, azonban a heti lebontású népességi adatokat csak közelíteni lehetett, mert kizárólag évi adatok vannak közölve a népesség számáról ([1]).
Ábrázoljuk a megfigyelt és a modellel kapott várt halálozási számokat.



3.7. ábra. Megfigyelt és várt értékek a teljes 2015-2021-es periódusban, a 2015-2019-es adatokból számolva.

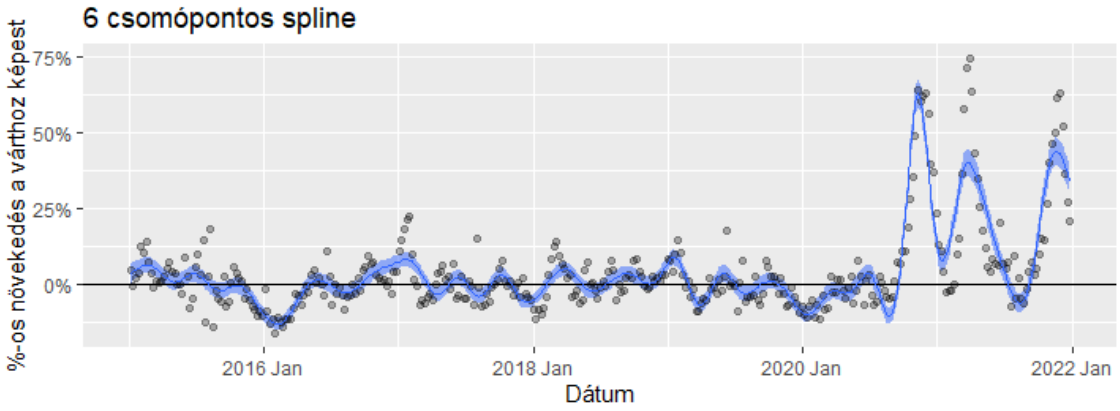
Egyértelműen látható a koronavírus hatása a halálozási számokra, és tisztán leolvasható a három erős hulláma is a járványnak. Ilyenkor a várthoz képest a halálok csaknem másfélszeresére nőttek.

A többlethalálozás kiszámítása

A módszer a többlethalálozás kiszámításához, amelyet a program használ a következő:

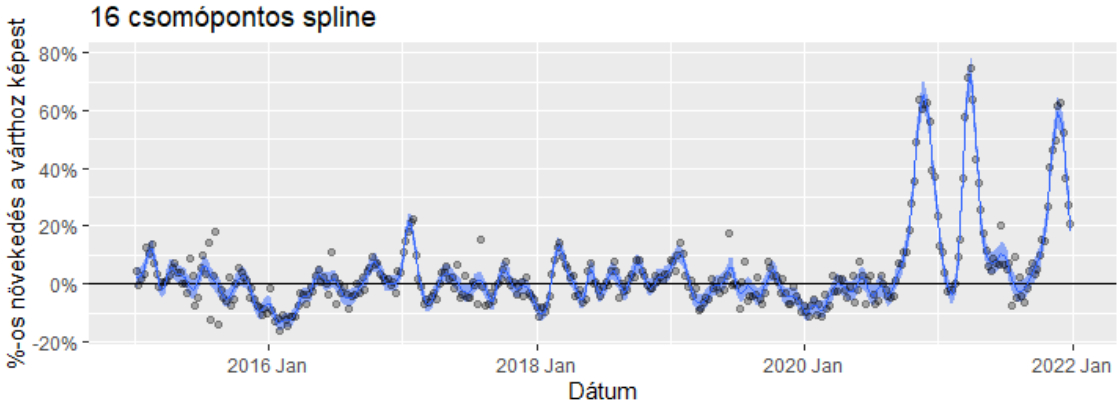
$$Y_h \mid \epsilon_h \ Poi(\mu_h[1 + f(h)]\epsilon_h) \tag{3.6}$$

ahol az Y_h-k a heti megfigyelt halálozási számokat jelölik, $100 \times f(h)$ a százalékban megadott többlet, μ_h pedig a feljebb számolt várt halálozás.
$f(h)$-t akarjuk kiszámolni, ez jelöli a keresett, covid által bekövetkezett százalékos többletet. Feltesszük, hogy $f(h) = 0$ olyan időszakokban, amikor nincs semmilyen járvány vagy más természeti katasztrófa.
A μ_h-t is abból az időszakból számoljuk, amikor még nem volt koronavírus, azaz a 2015-2019-es adatokból.
A járványos időszakokban egy évi 12 csomópontos harmadfokú splinenal modellezzük f-et. A 12 csomópontot az alapján választottuk, hogy a görbe ebben az esetben jól megfogta a 2020 előtti influenza járványok és a 2020-21-es koronavírus hatásait is. 6 csomópont nem volt elég ahhoz, hogy ezeket a változásokat is jól modellezze. A több csomópont, például 16 viszont már nagyon sok, túl sok hullámzása van egy éven belül. Ennek vizualizációjához ábrázoljuk 6 és 16 csomópontos spline használatának esetén a százalékos többlethalálozást.



3.8. ábra. Százalékos többlethalálozás 6 *csomópontos modellel.*

Látható, a 3.8 ábrán, hogy a harmadik hulláma a koronavírusnak nagyon el van simítva, illetve az év eleji influenzás időszakok is túl simítottak.

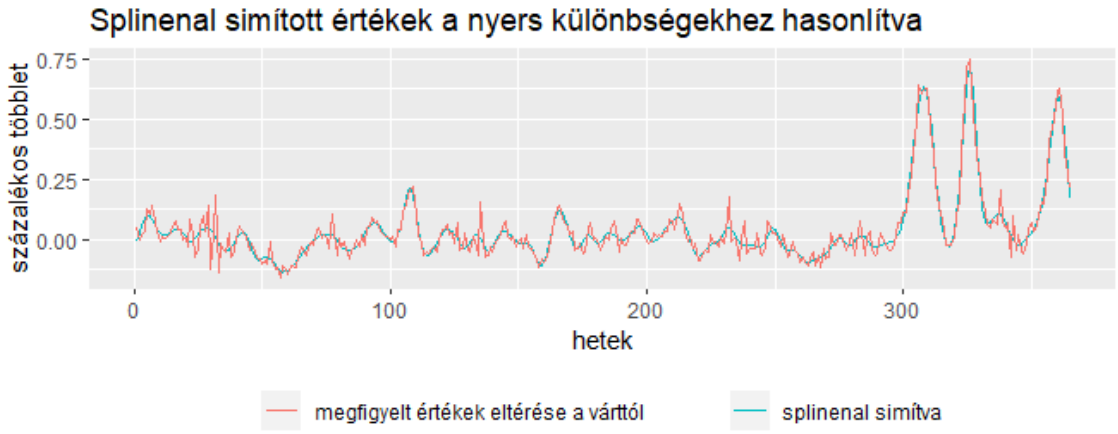


3.9. ábra. Százalékos többlethalálozás 16 *csomópontos splinenal.*

A 16 csomópontos esetben (3.9 ábra) már kevésbé simított a görbe, évi $5-6$ hullám már túl sok ingadozás.

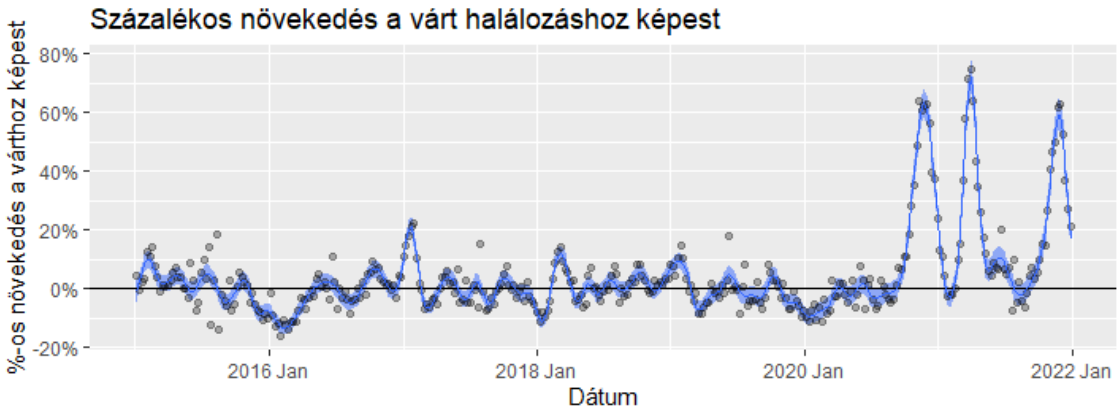
Ezek alapján tehát az évi 12 csomópontos splinenal számolunk. Megjegyezzük, hogy a programcsomag fejlesztői is ezt a csomópont-számot javasolták (5). A 3.6 képlettel becsültük $f(h)$ értékét, legyen ez $\hat{f}(h)$. A becslés kvázi-Poisson modellel történik a túlszóródás fennállása miatt. A centrális határeloszlás tételéből adódóan, $\hat{f}(h)$ tekinthető normális eloszlásúnak, így a konfidenciaintervallumokat ennek megfelelően számoljuk (5, 24).

A $\hat{\mu}_h\hat{f}(h)$ simítása a $Y_h - \hat{\mu}_h$ különbségnek. Ha annyi csomópontos splinet használnánk, mint ahány megfigyelésünk van, akkor $\hat{\mu}\hat{f}$ konvergálna ehhez a különbséghez. Ábrázoljuk is ezt:



3.10. ábra. $\hat{\mu}_h \hat{f}(h)$ *simítása a* $Y_h - \hat{\mu}_h$ *különbségnek. A teljes 2015-2021-es periódust ábrázoltuk.*

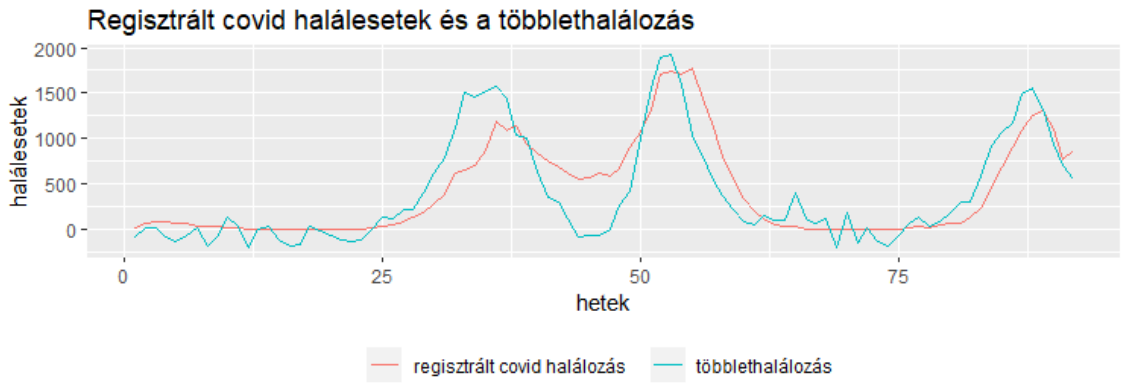
Ábároljuk a százalékos többlethalálozást.



3.11. ábra. A százalékos többlethalálozás és a megfigyelt heti halálozás ábrárája.

A 12 csomópontos splinenal becsült százalékos többlethalálozás ábrájából (3.11) leolvasható a koronavírus három erős hulláma, amikor $70-80\%$ körüli volt a többlethalálozás. A legnagyobb többlethalálozást a második covid hullám okozta, míg a legerősebb influenza hatás a halálozásra 2017-ben volt megfigyelhető.

Ábrázoljuk a bejelentett covid halálozások számát és a mért többlethalálozást.

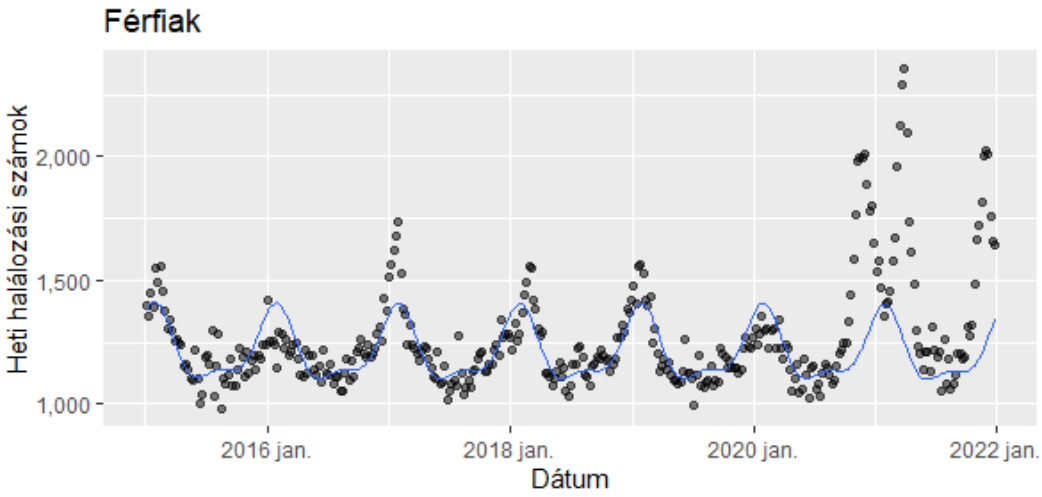


3.12. ábra. Az Excessmort programcsomaggal számolt többlethalálozást összehasonlítjuk a bejelentett covid halálozási számokkal a 2020. március - 2021. december időszakban, azaz a koronavírusos periódusban.

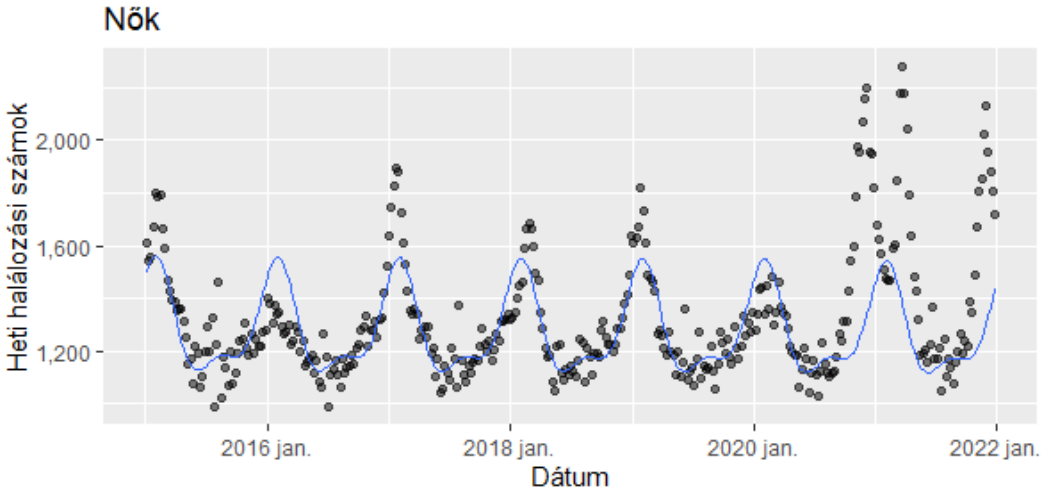
Látható a 3.12 ábrán, hogy a 45-50 közötti heteken -amelyek január vége- február elejének felelnek meg- nagy a különbség az adatok között. Ezen eltérések az influenza járvány hiánya miatt alakultak ki. Ennel a megoldásáról a 3.9 fejezetben lesz szó.
Az ábrán a covid halálok regisztrációjának elcsúszása is megfigyelhető.

3.8. Várt és többlethalálozás nemek és korcsoportok szerint

Eddig végig a népesség rétegzése nélkül számoltunk, most viszont ábrázoljunk néhány adatot nemek és korcsoportok szerint is, hogy jobban megfigyelhető legyen a vírus hatása a különböző csoportokon belül. Az eddig használt módszerekkel először a megfigyelt és várt halálozási számokat ábrázoljuk, továbbra is heti felbontásban.



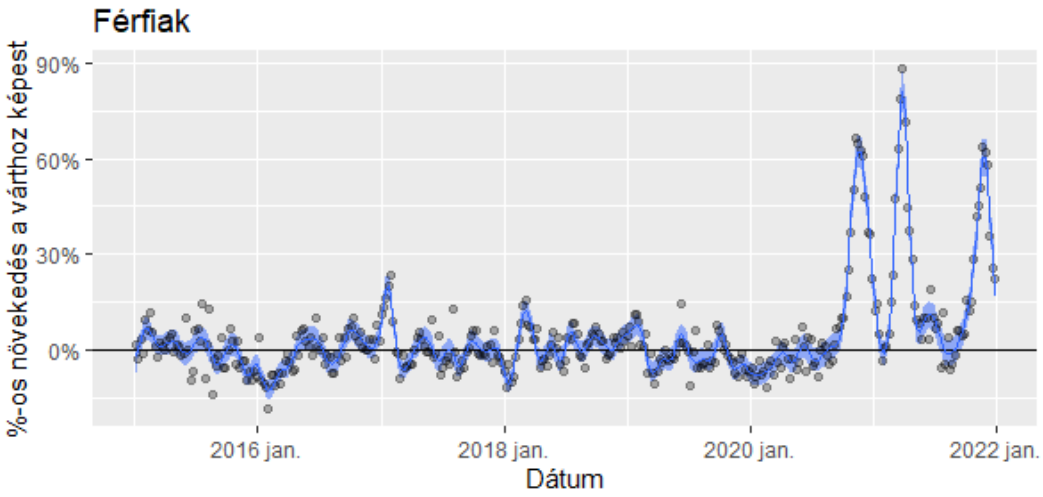
3.13. ábra. Megfigyelt és várt adatok férfiaknál.



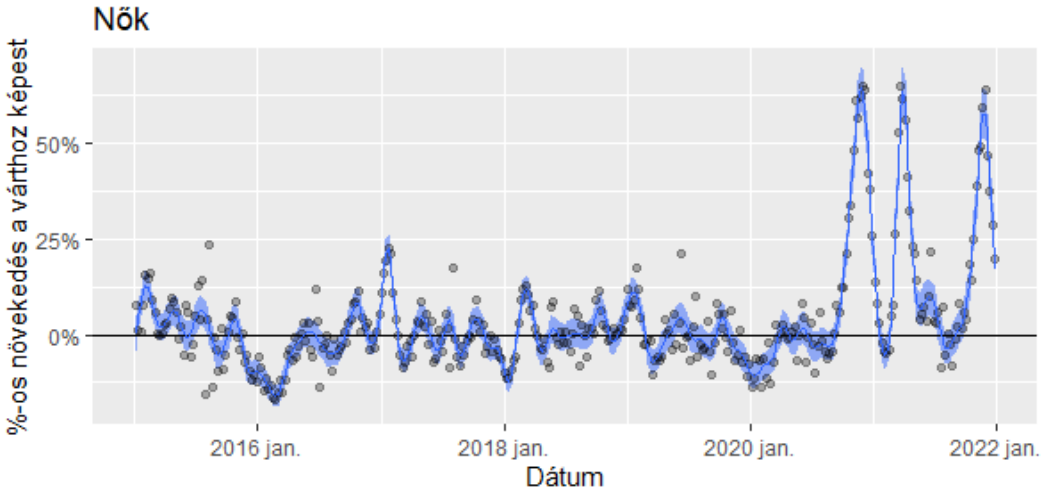
3.14. ábra. Megfigyelt és várt adatok nőknél.

Mindkét nemnél 2020. szeptemberétől kezdve leolvasható a járvány három erős hulláma, amikor 2000 feletti halálozási számok is voltak mérve a várt 1500-1600 körüliekhez képest a hullámok csúcsidején.

Ezek után pedig ábrázoljuk a százalékos többlethalálozást, ahol jobban látjuk majd a nemek közötti különbségeket:



3.15. ábra. Százalékos többlethalálozás férfiaknál.

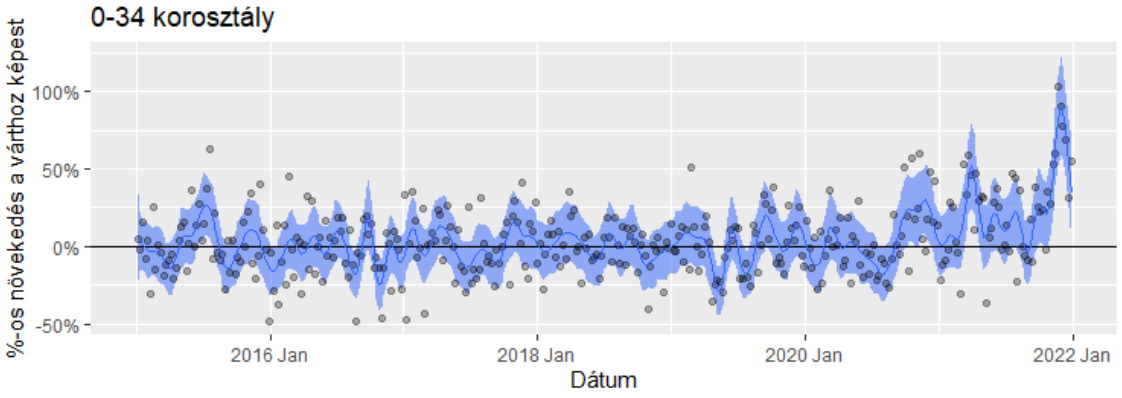


3.16. ábra. Százalékos többlethalálozás nőknél.

Látható, hogy a férfiak körében a harmadik hullám nagyobb többlethalálozást okozott, mint a nőknél, a többi hullám viszont hasonlóan hatott a két nemre. Összesítve, a 2020-21-es időszakban a nőknél kisebb volt a százalékos többlethalálozás, mint a férfiaknál. Érdemes azt is megfigyelni, hogy 2016-ban mennyire gyenge volt az influenza járvány, hiszen a várthoz képest, azaz más évek járványos időszakaihoz hasonlítva, sokkal kevesebb áldozat volt, a többlethalálozás negatív. A legerősebb influenzás év pedig 2017 volt, illetve, a nőknél jól láthatóan a 2015-ös év is erős volt.

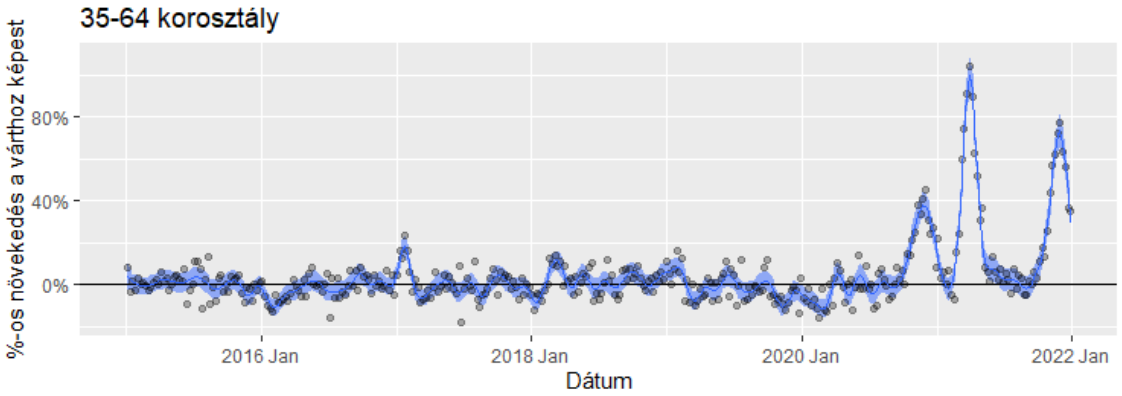
Ábrázoljuk a százalékos többlethalálozást korcsoportonként is annak érdekében, hogy

megfigyeljük a koronavírus hatásait a különböző korosztályoknál. Három korcsoportra osztjuk a populációt: 0-34, 35-64, 65+. A három korcsoport választása az alapján történt, hogy a koronavírusos adatok a 0-34 éves korcsoportról egyben vannak közölve (4). A felnőttek csoportosítását pedig a 65 éves életkor osztja ketté, hiszen az oltást kezdetben a 65 éven felülieknek ajánlották, mert ennek a korcsoportnak a számára a legnagyobb a vírus veszélye.



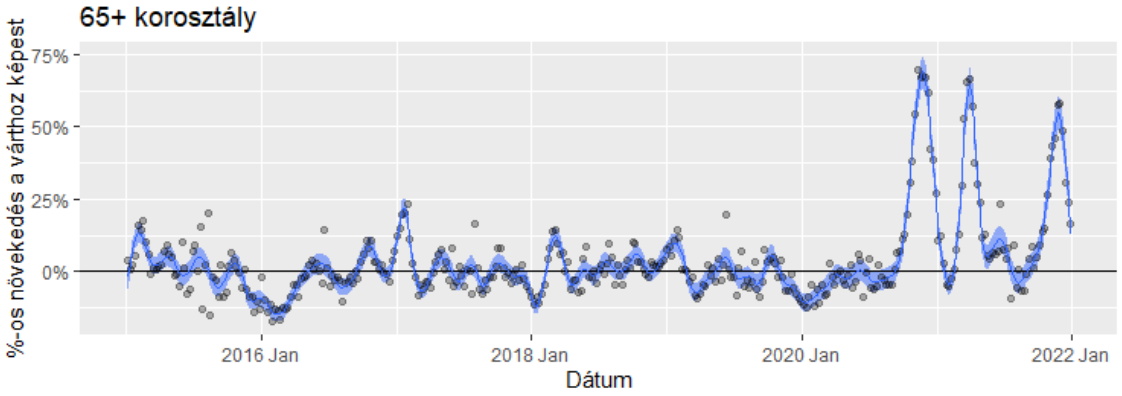
3.17. ábra. Százalékos többlethalálozás a 0-34 éves korosztálynál.

A koronavírus első három hulláma kevésbé tűnik ki, viszont a negyedik hullámnak ebben a korcsoportban is számottevő hatása van, csaknem 90%-kal nagyobb a többlethalálozás..



3.18. ábra. Százalékos többlethalálozás a 35-64 éves korosztálynál.

A legnagyobb hatása a harmadik hullámnak volt, 100% körüli emelkedéssel, ezt követte a negyedik. A konfidenciaintervallumok láthatóan szűkebbek, mint a fiataloknál.



3.19. ábra. Százalékos többlethalálozás a 65 évnél idősebb korosztálynál.

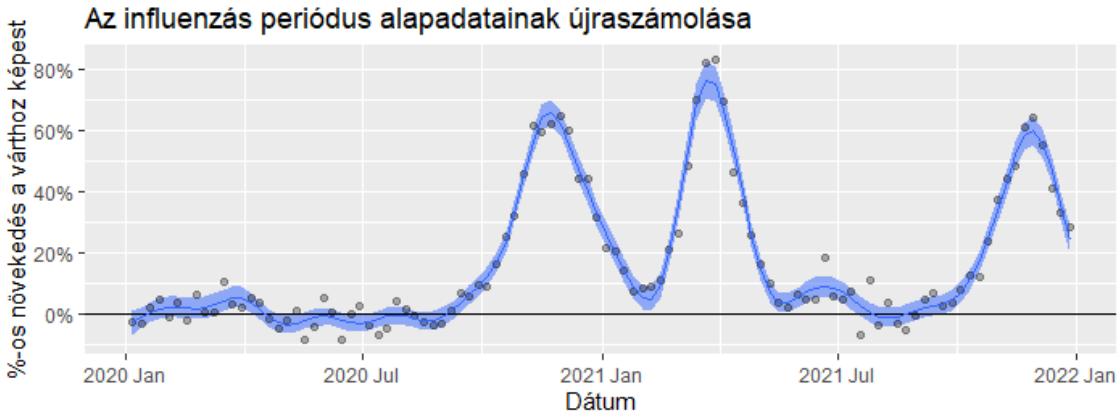
A második hullám okozta a legnagyobb többlethalálozást ebben a korcsoportban, 70%-kal volt nagyobb, mint kellett volna. Ezt követte a harmadik, majd a negyedik hullám 55% alatt. Valószínűleg az oltás hatása miatt érzékelhető ez a csökkenő trend.
A 3.17, 3.18, 3.19 ábrákból tehát jól látható, hogy az első hullám egyik korosztályra sem volt nagy hatással, a második az idősebb korostályt, a harmadik a 35-64, a negyedik a fiatal korosztályt érintette a legjobban halálozás szempontjából.

3.9. Az influenza hatásának elkülönítése

A koronavírus okozta halálokat nehéz pontosan számolni több okból is. A teszteléssel nem bizonyított koronavírus okozta halálokat nem tulajdonítják a covid áldozatainak, miközben lehetséges, hogy fertőzöttek voltak, és az okozta a halálukat.
Emellett, sok indirekt halálozáshoz is vezet a járvány, ilyenek a kórházak elérhetőségének hiánya miatt következett halálok, illetve a vírushelyzet miatt számos öngyilkosság, kábítószer túladagolás is történik.
De vannak pozitív indirekt hatások is, mint a más légúti fertőzések csökkenése, vagy kevesebb autóbaleset bekövetkezése.

A dolgozatban a többlethalálálozás méréséből próbáljuk meghatározni a koronavírus járvány hatását. Azonban tehát, a többlethalálozás nem csak a vírus direkt halálozásainak számát adja meg, hanem az indirekt hatásokat és az előrejelzésben való tévedést is beleszámolja. Az alapadatok kiszámítását már a 3.1 fejezetben vizsgáltuk, viszont az influenza indirekt hatását nem vettük eddig figyelembe. Az összes indirekt hatást nem lehet

a modellből kiszűrni, de mint a 3.11 ábra és a rétegzett ábrák is mutatják, az influenza volt az a szezonális hatás, amely a koronavírus mellett nagymértékben befolyásolta a halálozást. A járványügyi intézkedések következtében 2020-ban és 2021-ben számottevően kevesebb influenzás megbetegedés és halálozás volt más évekkel szemben.
Ha megnézzük a 2021. január végét és februárt, a többlethalálozás 0 körüli, vagy negatív, miközben a koronavírus okozta bejelentett halálok 500-700 fő között vannak. Ez a hónap tehát jól mutatja, hogy az influenza hiányából adódó pozitív hatás "kinullázta" a koronavírus okozta halálokat.
Ezért, megpróbáljuk elkülöníteni az influenza hatását a következőképpen: tegyük fel, hogy az év első három hónapja az influenzás időszak. Két lehetőségünk van: vagy teljesen kivesszük ezeket a hónapokat a 2015-2019-es évekből számolt várt adatok kiszámolásakor, vagy csak azon éveket hagyjuk benne, amikor kevés halálozás volt az influenzának tudható.
Az első esetben, mivel az f függvényt egy splinenal számoljuk, fogunk kapni a hiányzó hónapokra is valamilyen eredményt, viszont pontosabb lesz a becslés, ha a kis esetszámú években erre az időszakra is adunk adatokat. A 2015-2019-es években 2016-ban volt enyhébb influenza időszak, ezért ennek az egy évnek az adatait fogjuk bent hagyni az alapadatok kiszámolásakor a január-március időszakban.



3.20. ábra. Az influenza hatásának elkülönítése, a január-március időszakban csak a 2016-os évet bent hagyva a modellben

A 3.20 ábrán látható, hogy negatív értékekről 10% körülire nőtt a többlethalálozási arány 2021-nek februárjában, amely kiadja azt a néhány száz koronavírus halált, ami be volt jelentve.

4. fejezet

Összegzés

A dolgozat célja az volt, hogy a halálozási adatokból vizsgáljuk a koronavírus világjárvány hatásait Magyarországon.
Fontos megjegyezni, hogy az eredmények csak irányadóak, hiszen a többlethalálozási számok tartalmazzák a vírus direkt és indirekt hatásait is, amelyeket nem lehetett szétválasztani és külön kezelni a rendelkezésre álló adatok alapján. Illetve láttuk, hogy ezen indirekt hatások a koronavírus nélküli várt halandósági számokat is befolyásolják.
A dolgozatban elvégzett adatelemzések és számítások azt mutatják, hogy a járvány második, harmadik és negyedik hulláma nagymértékű többlethalandóságot okozott, amelyet viszonylagos pontossággal meg is lehetett becsülni. A hullámok csúcsain $70-80\%$-kal több haláleset volt, mint amennyi a koronavírus hiányában lett volna.
Illetve az is kiderült, hogy a Magyarországon bejelentett covid halálok reálisak, tényleg ennyire sok áldozata volt hazánkban a világjárványnak. 2020-ban összesen 141002 haláleset volt Magyarországon, 2021-ben 155000. Tehát a két évben 296002 haláleset volt. A várt halálozási szám 258471, tehát a különbség 37531. Ezzel szemben, a bejelentett koronavírusos halálesetek száma 39185. Ezt a különbséget viszont az influenza gyenge hatása adja, hiszen ha az influenzás időszak korrigálásával számoljuk a várt értékeket, akkor 251889-et kapunk, amihez hozzáadva a bejelentett 39185 koronavírus okozta halálesetet, 291074-et kapunk. Az 5000 körüli különbség pedig a negatív indirekt hatásokból adódó halálokkal magyarázható.
Zárásképpen pedig fontos még egyszer hangsúlyozni, hogy a megosztó vélemények ellenére, a direkt és indirekt hatásokat egybe számolva is, a világjárvány ténylegesen nagymértékben növelte a többlethalálandóságot és az intézkedések, oltások és korlátozások valóban okkal történtek.

Ábrák jegyzéke

Irodalomjegyzék

[1] Központi Statisztikai Hivatal: *Magyarország népességének száma nemek és életkor szerint, január 1.* (https://www.ksh.hu/interaktiv/korfak/orszag.html)

[2] Központi Statisztikai Hivatal: *Halálozások száma nemek és korcsoportok szerint, hetente* (https://www.ksh.hu/stadat_files/nep/hu/nep0065.html)

[3] Központi Statisztikai Hivatal: *Halálozások megye és régió szerint, negyedévente kumulált* (https://www.ksh.hu/stadat_files/nep/hu/nep0069.html)

[4] Koronamonitor: https://atlo.team/koronamonitor/

[5] Rolando J. Acosta, Rafael A. Irizzary, *Monitoring Health Systems by Estimating Excess Mortality*, 2020

[6] Rolando J. Acosta, Rafael A. Irizzary, *Introduction to excessmort*, 2021 (https://cran.r-project.org/web/packages/excessmort/vignettes/excessmort.html)

[7] Ferenci Tamás, *A magyarországi koronavírus járvány valós idejű epidemiológiája*, 2022 (https://research.physcon.uni-obuda.hu/COVID19MagyarEpi/)

[8] Ferenci Tamás, *Többlethalálozási adatok európai összevetésben*, 2022 (https://github.com/tamas-ferenci/ExcessMortEURa-direkt-hat%C3%A1s-elk%C3%BCl%C3%B6n%C3%ADt%C3%A9se-egy-k%C3%ADs%C3%A9rlet-az-influenza-j%C3%A1rv%C3%A1ny-kezel%C3%A9s%C3%A9re)

[9] C. P. Farrington, N. J. Andrews, *A Statistical Algorithm for the Early Detection of Outbreaks of Infectious Disease*, 1996

[10] P. McCullagh, J. A. Nelder, *Generalized Linear Models*, London New York Chapman and Hall, 1989 (http://www.utstat.toronto.edu/ brunner/oldclass/2201s11/readings/glmbook.pdf)

[11] Gáspár Csaba, *Numerikus Analízis 1*, 2020

[12] Andrew C. Stokes, Dielle J. Lundberg, Irma T. Elo, Katherine Hempstead, Jacob Bor, Samuel H. Preston, *COVID-19 and excess mortality in the United States: A county-level analysis*, 2021 (https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1003571)

[13] Liselotte Van Asten, *Excess Deaths during Influenza and Coronavirus Disease and Infection-Fatality Rate for Severe Acute Respiratory Syndrome Coronavirus 2, the Netherlands*, 2021 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7853586/)

[14] John Hinde, Clarice G.B. Demétrio, *Overdispersion: Models and Estimation*, 2007

[15] Jay M. Ver Hoef , Peter L. Boveng, *Quasi-Poisson vs. Negative Binomial REegression: How should we model overdispersed count data?*, 2007 (https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1141&context=usdeptcommercepub)

[16] Michaletzky György, Mogyoródi József, *Matematikai statisztika*, Tankönyvkiadó, 1986

[17] Tóth G. Csaba, *Másfél év pandémia Magyarországon: Mérséklődő különbségek a regionális és korspecifikus többlethalandóságban*, 2022 (https://kti.krtk.hu/wp-content/uploads/2022/01/CERSIEWP202204.pdf)

[18] Aris Perperoglou, *A review of spline function procedures in R*, 2019 (https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-019-0666-3)

[19] German Rodriguez, *The Negative Binomial Distribution*, 2011 (https://probabilityandstats.wordpress.com/tag/poisson-gamma-mixture/: :text=The%20Poisson%2DGamma%20Mixturetext=The%20negative%20binomial%20distribution%20can%20be%20viewed%20as%20a%20Poisson,as%20a%20Poisson%2DGamma%20mixture.)

[20] *Generalized Linear Models: Poisson Models for Count Data*, Princeton University (https://data.princeton.edu/wws509/notes/c4.pdf)

[21] German Rodriguez, *Generalized Linear Models: Models for Count Data With Overdispersion*, Princeton University, 2013 (https://data.princeton.edu/wws509/notes/c4a.pdf)

[22] Daniel M. Weinberger, *Estimation of Excess Deaths Associated With the COVID-19 Pandemic in the United States, March to May 2020*, 2020 (https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/2767980)

[23] Rick Wicklin, *What is loess regression?*, 2016 (https://blogs.sas.com/content/iml/2016/10/17/what-is-loess-regression.html)

[24] Csiszár Villő, *Statisztikai fogalmak összefoglalása*, (http://csvillo.web.elte.hu/prstat/elmelet.pdf)

[25] Márkus László, *Idősorok és többdim. stat. módszerek segédanyag: Poisson folyamat, Regresszió*, (https://web.cs.elte.hu/probability/markus/ElemzoIdosor.html)

[26] NCSS Statistical Software, *Negative Binomial Regression*, (https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Negative_Binomial_Regression.pdf)

[27] Achim Zeileis, Christian Kleiber, Simon Jackman, *Regression Models for Count Data in R*, (https://cran.r-project.org/web/packages/pscl/vignettes/countreg.pdf)

[28] Uzzoli Annamária, Kovács Sándor Zsolt, Páger Balázs, Szabó Tamás, *A hazai COVID–19-járványhullámok területi különbségei*, 2021 (https://www.ksh.hu/statszemle_archive/terstat/2021/2021_03/ts610302.pdf)