EÖTVÖS LORÁND TUDOMÁNYEGYETEM

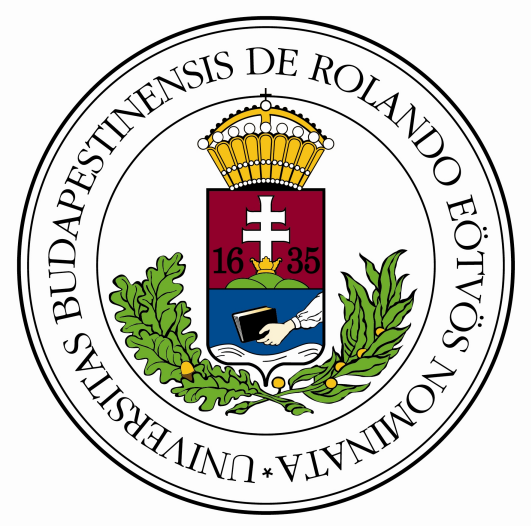
TERMÉSZETTUDOMÁNYI KAR

Mi a baj a p-értékekkel?

Szakdolgozat

Takács Noémi
Matematika Bsc
Matematikai elemző specializáció

Témavezető:
Dr. Zempléni András
egyetemi docens
Valószínűségelméleti és Statisztika Tanszék



Budapest, 2023

NYILATKOZAT

Név: Takács Noémi

ELTE Természettudományi Kar, szak: Matematika BSc

NEPTUN azonosító: HRZBHP

Szakdolgozat címe:
Mi a baj a p-értékekkel?

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, 2023. 06. 01.

Takács Noémi

a hallgató aláírása

Köszönetnyilvánítás

Hálával tartozom témavezetőmnek, Dr. Zempléni András professzor úrnak, aki a téma ajánlása mellett időt és energiát nem sajnálva segítette hasznos tanácsaival és észrevételeivel szakdolgozatom létrejöttét.

Szeretnék köszönetet mondani Barczi Péter tanár úrnak, aki igazán megszerettette velem a matematikát, hogy a gimnáziumi évek alatt hozzájárult szakmai fejlődésemhez.

Végül, de nem utolsósorban köszönöm családomnak és páromnak a tőlük kapott szeretetet, támogatást és biztatást, melyre mindig számíthatok.

Takács Noémi

Tartalomjegyzék

1. Bevezetés

Szakdolgozatomban a statisztikai tesztelésekből származó döntések alapjául szolgáló p-értékeket mutatom be, koncentrálva a tudományos közösségekben körülöttük zajló vitákra. A különböző álláspontok felkutatása mellett az R program használatával végzett szimulációkon és elemzéseken keresztül vizsgálom bizonyos tényezők gyakorlati hatását a szóban forgó p-értékekre.

Dolgozatom elején definiálom az alapvető statisztikai fogalmakat a hipotézisvizsgálat és a lineáris regresszió témaköréből, melyek szükségesek a probléma megértéséhez. Továbbá körüljárom a lineáris regresszió feltételeit, melyek ahhoz szükségesek, hogy az alkalmazott módszerek az elvárt tulajdonságokkal rendelkezzenek, ismertetem a kritériumok feltérképezésének egy-egy lehetséges technikáját, ill. bemutatok néhány eljárást, melyet akkor lehet alkalmazni, ha az egyes feltételek nem teljesülnek.

A 3. fejezetben foglalkozom a p-értékek használatából adódó problémákkal. A statisztikai próbák használata széles körben elterjedt, számos tudományos és társadalmi megfigyelésből levont következtetés alátámasztása épít a p-értékekre. Azonban nem minden eset fekete vagy fehér, még ha gyakran válaszként egy igent vagy nemet várunk. Sajnos sokan nem megfelelően alkalmazzák, nem tartanak be bizonyos feltételeket vagy rosszul értelmezik a p-értékeket, így nem valós konklúziókra jutnak, melyeknek olykor komoly következménye is lehet. A téma feldolgozásához a Nature tudományos folyóirat 2019. március 21-én megjelent számának „Scientists rise up against statistical significance" c. cikkét vettem kiindulásul.

Dolgozatom 4. fejezetében lineáris regressziós modelleknél tanulmányozom azt a kérdést, hogy hipotézisvizsgálattal be tudjuk-e statisztikailag bizonyítani, hogy az adott tényezőknek van-e hatása az eredményváltozóra. Szemügyre veszem, hogy az egyes feltételek nem teljesülése ellenére mégis alkalmazásra kerülő módszereknél ezek a feltételek mekkora hatást gyakorolnak a választ igazoló p-értékek nagyságára. A vizsgálatokhoz a 2. fejezetben ismertetett "feltérképező" és "megoldó" eljárásokat alkalmazom fiktív és generált adatokra. Generált adatok esetében ismert az elméleti eredmény, így könnyen összehasonlítható, hogy mennyiben tér el tőle a gyakorlatban kapott.

A szimulációk során a p-értékeket a standard, megszokott módon kezelem, ezzel rávilágítva, hogy a hipotézisvizsgálati eljárások nem megfelelő használata olykor téves eredményekhez vezethet. A különböző esetek jól tükrözik, hogy a statisztikai szignifikancia és a hozzá hasonló fogalmak megalapozatlan használata hibás útra vezethet.

Végül pedig az 5. fejezetben a különböző országokban várható élettartamot modellezem, valós megfigyeléseket használva. Leellenőrzöm a feltételek teljesülését, és megvizsgálom, hogy olyan tényezők, mint a megfigyelések száma vagy a modellbe bekerült változók hogyan befolyásolják az eredményeket, milyen hatással vannak a p-értékekre.

A dolgozathoz írt kódjaim a következő linken érhetők el:
`https://github.com/TNoemi5/Mi-a-baj-a-p-ertekekkel-.git`

2. Statisztikai alapok

2.1. Hipotézisvizsgálat

Az alfejezetet az egyetemi jegyzeteim alapján dolgoztam ki.

Hipotézisvizsgálat során következtetéseket próbálunk meg levonni véletlen mintákból egy sokaság egészére vonatkozóan. Általában nem tudjuk megfigyelni az egész populációt, ezért van szükség mintavételezésre, amelyeken alkalmazhatunk statisztikai teszteket állítások alátámasztására vagy éppen elutasítására.

1. Definíció. *Hipotézis: Egy állítás, melyet a vizsgálat folyamán szeretnénk bebizonyítani vagy megcáfolni.*

A feladat specifikálásához a Θ paraméterteret két diszjunkt részre bontjuk: Θ_0 és Θ_1-re. Ekkor a statisztikai hipotézisvizsgálat alapfeladata a következő:

$$H_0 : \vartheta \in \Theta_0$$
$$H_1 : \vartheta \in \Theta_1,$$

ahol H_0 a nullhipotézis, H_1 pedig az ellenhipotézis. A nullhipotézis szokásosan valamilyen elvárt dolgot, sok éves tapasztalatnak megfelelőt állít. Az ellenhipotézis vagy alternatív hipotézis ellentmond a nullhipotézisnek, jellemzően ez a bennünket érdeklő kérdés.

2. Definíció. *Statisztikai próba: olyan vizsgálati módszer, mellyel a minta alapján döntünk, hogy elvetjük-e a nullhipotézist vagy sem.*

3. Definíció. *Paraméteres próba: Olyan próba, melyet paraméterek értékének tesztelésére alkalmazunk, ahol a paramétertér a valós számok részhalmaza.*

4. Definíció. *Elfogadási tartomány ($\mathcal{X}_e$): azon megfigyelések halmaza, amelyre nem utasítjuk el a nullhipotézist.*
Kritikus tartomány ($\mathcal{X}_k$): azon megfigyelések halmaza, amelyre elutasítjuk a nullhipotézist.

Megjegyzés: $\mathcal{X}_e$ és $\mathcal{X}_k$ diszjunkt halmazok, uniójuk a $\mathcal{X}$ mintatér.

5. Definíció. *Elsőfajú hiba: Elvetettük a nullhipotézist, pedig valójában az igaz volt. Valószínűsége: $\alpha(\vartheta) = P_{\vartheta\in\Theta_0}(\mathcal{X}_k)$*

6. Definíció. *Másodfajú hiba: Nem utasítottuk el a nullhipotézist, pedig az valójában nem volt igaz. Valószínűsége: $\beta(\vartheta) = P_{\vartheta\in\Theta_1}(\mathcal{X}_e)$*

7. Definíció. *Erőfüggvény: $\psi(\vartheta) = P_{\vartheta\in\Theta_1}(\mathcal{X}_k)$, tehát annak a valószínűségét mutatja meg, hogy helyesen utasítjuk el a nullhipotézist.*

8. Definíció. *Terjedelem vagy szignifikancia szint: $\alpha := \sup_{\vartheta\in\Theta_0} \alpha(\vartheta)$. A próba elvégzése előtt szokás rögzíteni, tipikusan 5 %-on, ekkor 5 % elsőfajú hiba valószínűsége mellett, 95 %-os megbízhatósággal döntünk.*

9. Definíció. *Kétoldali paraméteres próba: $H_0 : \vartheta = \vartheta_0, \quad H_1 : \vartheta \neq \vartheta_0$*
Egyoldali paraméteres próba: $H_0 : \vartheta = \vartheta_0, \quad H_1 : \vartheta > \vartheta_0$ vagy $H_1 : \vartheta < \vartheta_0$

10. Definíció. *Próbastatisztika: Egy alkalmas $T : \mathcal{X} \to \mathbb{R}$ statisztika, melynek segítségével meg tudjuk határozni a kritikus tartományt. Egyoldali próbánál*

$$\mathcal{X}_k = \{x \in \mathcal{X} : T(x) < c\} \text{ vagy } \mathcal{X}_k = \{x \in \mathcal{X} : T(x) > c\},$$

kétoldali próbánál

$$\mathcal{X}_k = \{x \in \mathcal{X} : |T(x)| > c\},$$

ahol c (vagy c_α) a kritikus érték. Az előre rögzített α szinthez azt a c_α értéket keressük, melyre éppen α lesz a próba terjedelme: $\sup_{\vartheta \in \Theta_0} P(\mathcal{X}_k) = \alpha$.

11. Definíció. *p-érték: az a terjedelem, amire a kritikus érték megegyezik a próbastatisztikával.*

Másképpen fogalmazva: ha elutasítjuk H_0-t, akkor akkora lesz az elsőfajú hiba bekövetkezésének valószínűsége, mint amennyi a p-érték (mindig 0 és 1 közötti értéket vesz fel). Egy hipotézis tesztelésekor dönthetünk a p-érték alapján: amennyiben az kisebb az α szignifikancia szintnél, elvetjük a nullhipotézist, ha pedig nagyobb nála, akkor nem tudjuk elvetni. Ha elutasítjuk a nullhipotézist, tehát ha a minta $\in \mathcal{X}_k$, akkor erős (szignifikáns) döntést hozunk, ellenkező esetben pedig gyenge döntést.

2.2. Lineáris regresszió

A 2.2.1-es és 2.2.2-es pontokat az egyetemi jegyzeteim alapján dolgoztam ki.

2.2.1. A modell felépítése

A statisztikában regressziószámítás során egy függő változó (jelölés: Y) és egy vagy több magyarázó változó (jelölés: $X_1, \ldots, X_k$) közötti viszonyt próbáljuk meg megbecsülni. Egyrészt használhatjuk a függő változó előrejelzésére a magyarázó változó(k) ismeretében, másrészt pedig a függő és a magyarázó változók közötti kapcsolat vizsgálatára. Későbbi szimulációimban az utóbbi játszik fontos szerepet.

Jelöljük Y_j-vel a függő változó és $X_{i,j}$-vel az i-edik magyarázó változó j-edik megfigyelését, ahol $i = 1, \ldots, k$ és $j = 1, \ldots, n$. Magyarázó változó nem csak önmagukban a megfigyelések lehetnek, de azok függvénye is. Általában jóval több megfigyelésünk van, mint magyarázó változónk ($n >> k$). Továbbá a lineáris modellhez hozzáadódik egy 0 várható értékű és véges szórású valószínűségi változó is, ami a zajnak felel meg (jelölés: ε_j). Ezt azért szükséges feltüntetni, mert csak a megfigyelések fixek, a függő változó már nem, azt befolyásolják a magyarázó változók, melyek közül általában nem tudunk mindent figyelembe venni, és a véletlen is.

Lineáris modell esetében az Y_j-t a magyarázó változók lineáris kombinációjával közelítjük, ahol az együtthatókat b_i ($i = 0, \ldots, k$) jelöli. Vagyis a megfigyelt adatok pontjaira szeretnénk egy ún. regressziós egyenest illeszteni. Tehát a (zajjal kiegészült) modell a következő összefüggést írja le:

$$Y_j = b_0 + b_1 X_{1,j} + \ldots + b_k X_{k,j} + \varepsilon_j,$$

röviden:

$$Y = Xb + \varepsilon,$$

vagy a mátrixos alak:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & \dots & X_{1k} \\ 1 & X_{21} & \dots & X_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \dots & X_{nk} \end{pmatrix} \cdot \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

12. Definíció. *A függő változó közelítése: $\hat{Y} = X\hat{b}$, ahol $\hat{b}$ a paramétervektor egy becslése.*

13. Definíció. *A függő változó és annak becsült értéke közötti eltérést reziduálisoknak nevezzük: $\hat{\varepsilon}_j = Y_j - \hat{Y}_j$.*

Regressziószámítás alatt a b_i paraméterek becslése történik. Ennek leggyakoribb módszere a legkisebb négyzetek módszere. Ahogy a neve is utal rá, a módszer lényege, hogy az eltérések négyzetösszegét, vagyis a reziduálisok négyzetösszegét minimalizálja.

1. Állítás. *A paramétervektor a következőképp becsülhető: $\hat{b} = (X^T X)^{-1} X^T Y$.*

Bizonyítás. A reziduálisok négyzetösszege:

$$\hat{\varepsilon}^2 = \|Y - \hat{Y}\|^2 = \|Y - X\hat{b}\|^2 = (Y - X\hat{b})^T (Y - X\hat{b}) = Y^T Y - Y^T X\hat{b} - \hat{b}^T X^T Y + \hat{b}^T X^T X\hat{b}.$$

A kifejezés $\hat{b}$ szerinti deriváltja:

$$-2X^T Y + 2X^T X\hat{b}.$$

Miután egyenlővé tesszük nullával, kifejezhető belőle $\hat{b}$:

$$\begin{aligned} -2X^T Y + 2X^T X\hat{b} &= 0 \\ 2X^T X\hat{b} &= 2X^T Y \\ \hat{b} &= (X^T X)^{-1} X^T Y. \end{aligned}$$

□

14. Definíció. *Egy lineáris modell magyarázóerejét a determinációs együtthatóval jellemezhetjük. Ez a szám megadja, hogy a modell hány %-ban magyarázza jól az Y változó változékonyságát. Kiszámítása:*

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

ahol

$$SSE = \sum_{j=1}^{n} (Y_j - \hat{Y}_j)^2, SST = \sum_{j=1}^{n} (Y_j - \bar{Y})^2$$

a reziduális négyzetösszeg (Sum of Squares Errors), ill. a teljes négyzetösszeg (Sum of Squares Total).

15. Definíció. *A korrigált determinációs együttható figyelembe veszi a magyarázó változók számát is. Segítségével el tudjuk dönteni, hogy megéri-e egy új változót bevenni a modellbe, az szignifikánsan növeli-e a modell illeszkedését vagy sem.*

$$\textit{korrigált } R^2 = 1 - \left(\frac{(1 - R^2)(n-1)}{n-k-1} \right),$$

ahol R^2 a nem korrigált determinációs együttható, n a megfigyelések száma, k pedig a felhasznált magyarázó változók száma.

2.2.2. Hipotézisvizsgálat lineáris regresszióban

Lineáris modellekben a hipotézisvizsgálat során a függő és az egyes magyarázó változók közti összefüggőséget elemezzük a b paraméterek tesztelésével. Erre t-próbát alkalmazunk, a következő formában:

$$H_0 : b_i = 0$$
$$\underline{H_1 : b_i \neq 0}$$

A nullhipotézis azt feltételezi, hogy az X_i változóknak nincs hatásuk az Y változóra, míg az ellenhipotézis szerint az X_i megváltozása befolyásolja az Y változó értékét is.

Amennyiben igaz a H_0, a teszt próbastatisztikája n-1 szabadságfokú és t-eloszlású:

$$t = \frac{\hat{b_i}}{D(\hat{b_i})},$$

ahol

$$D^2(\hat{b_i}) = \frac{\sigma^2}{\sum_{j=1}^{n}(X_{i,j} - \bar{X}_i)^2}$$

a becsült paraméter varianciája. A reziduálisok szórása pedig a következőképp becsülhető:

$$\hat{\sigma}^2 = \frac{\hat{\varepsilon}^2}{n-k}.$$

2.2.3. Előfeltételek

A ponthoz a [8] forrást használtam fel.

Lineáris regressziós modellek esetén több előfeltételnek is teljesülnie kell, különben torzulhatnak az összefüggések és a szignifikancia tesztek, kevésbé megbízható p-értékeket kaphatunk. A bemutatott standard becslési technikákat alkalmazó modelleknek az alábbi főbb feltételeknek kell megfelelniük, ahhoz hogy az eredményeket biztonsággal lehessen értelmezni:

- **Függetlenség.** Az összes megfigyelésnek és így az összes regressziós hibának függetlennek kell lennie egymástól. Ekkor teljesülni fog az Y függő változó minden egyes megfigyelésére, hogy az X_i magyarázó változó egyetlen megfigyelésétől függ. Ezen kritérium megszegése okozhatja a legnagyobb problémát a regresszióanalízisben. Többnyire a feltétel nem teljesülésének oka a megfigyelések csoportba oszthatósága, ill. a soros hatások (pl. idősorok).

- **Multikollinearitás.** Nemcsak az egy változón belüli megfigyeléseknek kell függetlennek lenniük egymástól, hanem a magyarázó változók sem korrelálhatnak egymással. Ha két magyarázó változó erősen korrelál, akkor ugyanazt az információt hordozzák magukban, így nem lesz egyértelmű a paraméterbecslés (melyik milyen mértékben befolyásolja a függő változó változását). Szélsőséges esetben, amikor az egyik magyarázó változó konstansszorosa egy másiknak, akkor nem lesz invertálható az X mátrix, ami szükséges a legkisebb négyzetek módszerének paraméterbecsléséhez.

- **Homoszkedaszticitás.** Szükséges, hogy a reziduálisok szórása azonos legyen, függetlenül az X_i változó értékeitől, mivel a hipotézisvizsgálat során csak egy szórás értékkel számolunk. Azonban a kis eltérések kevésbé befolyásolják a vizsgálatot és sokszor kezelhetőek pl. adattranszformációval.

- **Normalitás.** A modell hibáinak normális eloszlást kell követniük, ekkor a legkisebb négyzetek módszerének becslése egyben maximum likelihood becslés is lesz. Ennek vizsgálatára léteznek különböző tesztek (pl. Shapiro-Wilk), ugyanakkor ezek kis mintaelemszám esetén alacsony hatásfokkal rendelkeznek, nagy mintáknál pedig szinte mindig jelentős eltérését mutatnak a normális eloszlástól. A hipotézisvizsgálathoz használt t-próba megfelelő működéséhez a becslés normalitása szükséges, de ez nagy elemszámú mintákra többnyire feltehető. A tudományos életben nemcsak a normalitási kikötést sértik meg kisebb-nagyobb mértékben, de változékony az ezen feltétel betartásának megítélése is, hiszen a centrális határeloszlás-tétel szerint a mintaelemszám növelésével a minta átlaga tart a normális eloszláshoz, függetlenül a minta valódi eloszlásától.

2.2.4. Az előfeltétel-teljesülés ellenőrzése

Későbbi szimulációim és elemzéseim során a következő módszerekkel térképeztem fel, hogy teljesülnek-e a lineáris regresszió megfelelő működéséhez szükséges feltételek.

Függetlenség A részhez a [20]. forrást használtam fel.

A reziduálisok autokorrelációjának vizsgálatára alkalmazható a Durbin-Watson próba. A teszt nullhipotézise, hogy a reziduálisok nem korrelálnak egymással, az ellenhipotézis pedig ennek az ellentétét állítja. A próbastatisztika értéke a következő:

$$d = \frac{\sum_{i=2}^{n}(\varepsilon_i - \varepsilon_{i-1})^2}{\sum_{i=2}^{n}\varepsilon_i^2},$$

ami mindig 0 és 4 közé esik, jelentése:

- $d < 2$: pozitív autokorreláció,
- $d = 2$: nincs autokorreláció,
- $d > 2$: negatív autókorreláció.

Az adott szignifikancia szinthez tartozó kritikus értékeket egy arra szolgáló táblázatból tudjuk leolvasni, függ a magyarázó változók számától és a megfigyelések számától is. A Durbin-Watson próbával a soros hatások (pl. idősoroknál) jelenlétét lehet kimutatni, azonban dolgozatomban ilyen jellegű adatokkal nem foglalkoztam.

Homoszkedaszticitás A részhez a [20]. forrást használtam fel.

A regressziós modellbéli heteroszkedaszticitás megállapítására szolgál a Breusch-Pagan próba, melynek nullhipotézise, hogy a reziduálisok homoszkedasztikusak, ellenhipotézise pedig, hogy a reziduálisok heteroszkedasztikusak. A teszt elvégzéséhez a következő lépéseket kell végrehajtani:

1. regressziós modell illesztése a megfigyelésekre,

2. reziudálisok négyzetének kiszámítása,

3. új regressziós modell illesztése, ahol a függő változó értékei a reziduálisok négyzetei,

4. a χ^2 statisztika kiszámítása: $\chi^2 = n \cdot R^2_{\text{új}}$, ahol n az összes megfigyelés száma, $R^2_{\text{új}}$ pedig a második regressziós modell determinációs együtthatója.

Ezt követően meg lehet határozni az ehhez a χ^2 statisztikához tartozó p-értéket, ahol a szabadsági fok a magyarázó változók száma.

Normalitás A részhez a [20]. forrást használtam fel.

A reziduálisok normalitását lehet Shapiro-Wilk próbával tesztelni. A próbastatisztika kiszámításához a reziduálisokat növekvő sorba kell rendezni. Ekkor a próbastatisztika értéke:

$$W = \frac{(\sum_{i=1}^{n} a_i \cdot \varepsilon_{(i)})^2}{\sum_{i=1}^{n} (\varepsilon_i - \bar{\varepsilon})^2},$$

ahol $\varepsilon_{(i)}$-k a sorbarendezett reziduálisok, az a_i-k pedig a megfelelő táblázatból származó együtthatók. A p-értékeket szintén egy arra szolgáló táblázat alapján lehet megállapítani.

Multikollinearitás A részhez a [15]. forrást használtam fel.

A multikollinearitás egy mérőszáma a varianciainflációs tényező (VIF - Variance Inflation Factor). Megmutatja, hogy az egyes magyarázó változók viselkedését mennyire befolyásolja a többi magyarázó változó. Kiszámításának módja:

$$VIF_i = \frac{1}{1 - R_i^2},$$

ahol R_i^2 az a nem korrigált determinációs együttható, melyet abból a regresszióból kapunk, ahol az i-edik magyarázó változót becsüljük a többivel. Ha a VIF értéke 1, akkor az adott változó nem korrelál semelyik másikkal (ekkor $R_i^2 = 0$). Ha ez a szám 1 és 5 közé esik, akkor az adott változó mérsékelten korrelál, de még nincsen szükség korrekciós intézkedésekre. Ha pedig a VIF nagyobb 5-nél, akkor erős összefüggésről van szó.

2.2.5. Megoldások az előfeltételek megsértése ellen

2.2.5.1. Box-Cox transzformáció A következő részt a [11] forrás alapján dolgoztam ki.

Az adataink gyakran nem teljesítenek minden kezdeti feltételt, de léteznek korrigáló módszerek. Ilyen a George Box és Sir David Roxbee Cox (1964) által megalkotott technika is, amely elsősorban csökkenti a lineáris modell reziduálisainak nem-normalitását, de mérsékli a heteroszkedaszticitást is. A módszer alapötlete az, hogy a jobb oldal változatlanságával alakítsuk át a függő változót úgy, hogy a reziduálisok normális eloszlást mutassanak. Nem szabad azonban elfelejteni, hogy ekkor a regressziós együtthatók is megváltoznak, azokat a transzformált változóhoz kell viszonyítani, ill. egy-egy előrejelzésnél, az eredeti mérték szerinti értelmezéshez, az értékek visszatranszformálására van szükség.

A standard Box-Cox transzformáció képlete:

$$Y_i^{(\lambda)} = \begin{cases} \frac{Y_i^{\lambda} - 1}{\lambda} & (\lambda \neq 0) \\ \log(Y_i) & (\lambda = 0). \end{cases}$$

Az inverz transzformáció képlete:

$$Y_i = \begin{cases} \left(1 + \lambda Y_i^{(\lambda)}\right)^{\frac{1}{\lambda}} & (\lambda \neq 0) \\ \exp\left(Y_i^{(\lambda)}\right) & (\lambda = 0). \end{cases}$$

A pontos transzformációt a λ paraméter határozza meg. Cél: kiválasztani azt az optimális λ értéket, amely a normális eloszlás lehető legjobb közelítését adja. Ez megoldható maximum likelihood módszerrel. (Megjegyzés: ha a λ becslése során 1-et kapunk, az azt jelenti, hogy nem szükséges az adatok transzformálása.)

A Box-Cox transzformáció hasznos lehet, ha a mintánk átalakítására van szükség, azonban ez a módszer a tapasztalati értékeket használja. Amennyiben ismert, egy elméleti átalakítás (pl. fizikai összefüggések, kémiai reakciók) pontosabb eredményt kaphatunk annak használatával.

2.2.5.2. Súlyozott regresszió A következő részt a [12] forrás alapján dolgoztam ki.

A súlyozott regresszió alapötlete az, hogy az egyes megfigyelésekhez társítsunk bizonyos súlyokat, amelyek megmutatják, hogy a modellillesztés során mennyire fontos figyelembe venni az adott megfigyeléseket. Ez a módszer lehetővé teszi, hogy a kiugró értékek kevésbé húzzák el a regressziós egyenest, és segít a homoszkedaszticitás elérésében.

Ha a reziudálisok homoszkedasztikusak (és teljesül a többi feltétel is), akkor fennáll a következő egyenlőség:

$$C = E[\varepsilon\varepsilon^T] = \sigma^2 I = \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix},$$

ahol C a hibák kovariancia mátrixa. Mivel (feltételezhetően) bármely két különböző hiba független egymástól, ezért kovarianciájuk 0, így a főátlón kívüli elemek nullák. A főátlóban pedig az egyes hibák önmagukkal vett kovarianciája, vagyis a szórásnégyzete található meg. Amennyiben heteroszkedaszticitás áll fenn, a főátló elemei nem azonosak:

$$C = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n^2 \end{pmatrix}.$$

Mivel a függő változó többdimenziós normális eloszlású, ezért a paraméterbecslés a következőképp történik (maximum likelihood módszerrel):

$$\hat{b} = \arg\max_b \frac{1}{\sqrt{(2\pi)^n |C|}} \, exp\left(-\frac{1}{2}(Y - Xb)^T C^{-1}(Y - Xb)\right),$$

a logaritmusát véve és elhagyva a b-től nem függő tagokat:

$$\begin{aligned} \hat{b} &= \arg\max_b \; -\frac{1}{2}(Y - Xb)^T C^{-1}(Y - Xb) \\ &= \arg\min_b \; (Y - Xb)^T C^{-1}(Y - Xb), \end{aligned}$$

kifejtve és a konstans tagot elhagyva:

$$\hat{b} = \arg\min_b b^T X^T C^{-1} X b - 2 b^T X^T C^{-1} Y.$$

A kifejezés deriváltjának nullával való egyenlővé tételéből kifejezhetjük a paraméter becsült értékét:

$$2X^T C^{-1} X b - 2X^T C^{-1} Y = 0$$
$$\hat{b} = (X^T C^{-1} X)^{-1} X^T C^{-1} Y.$$

Látható, hogy a súlyozott regresszió paraméterbecslése csak a kovariancia mátrix inverzének használatában különbözik egy sima lineáris esettől (a súlyozatlan eset ekvivalens a $C = I$ súlyozott esettel). Az inverz könnyen kiszámolható, hiszen egy diagonális mátrixról van szó:

$$C^{-1} = \begin{pmatrix} 1/\sigma_1^2 & 0 & \dots & 0 \\ 0 & 1/\sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1/\sigma_n^2 \end{pmatrix}.$$

Tehát az optimális súlyok a szórásnégyzetek reciproka. A levezetés azt is alátámasztja, hogy ha a hiba normális eloszlású, akkor a legkisebb négyzetek módszerének becslése maximum likelihood becslés is. Ekkor a reziduálisok szórásnégyzetének várható értéke konstans lesz:

$$C^{-1} E(\varepsilon \varepsilon^T) = \frac{1}{\sigma^2} \sigma^2 = 1.$$

Heteroszkedasztikusság esetén általában nem ismerjük az egyes hibák varianciájának pontos értékét, így nem mindig egyszerű megfelelő súlyokat találni. Ha azonosítani tudunk egy változót, amely szoros kapcsolatban áll a reziduálisok szórásával (pl. a megfigyelések értékeinek növekedésével nő a reziduálisok szórása is), akkor jó ötlet lehet a változó inverzét venni súlyvektorként, de megpróbálhatjuk becsülni a szórásnégyzeteket is például az alábbi módszerrel:

- súlyok nélküli lineáris regresszió megoldása,
- reziduálisok kiszámolása,
- a szórásnégyzetek becslése a reziduálisokból (ε^2),
- súlyozott regresszió megoldása a becsült varianciákkal.

2.2.5.3. Véletlen tengelymetszet modell A következő részt a [13] forrás alapján dolgoztam ki.

A véletlen tengelymetszet (random intercept) modell olyan esetekben alkalmazható, amikor egy magyarázó változón belül a megfigyeléseket valamilyen tulajdonság alapján csoportokba lehet sorolni, tehát a hibák nem azonos eloszlásúak. A modellbe beépíthető ez a fajta klaszterezés ún. véletlen hatásként, ekkor mindegyik csoporthoz külön metszéspont (intercept) fog tartozni.

Tegyük fel, hogy csak egyetlen magyarázó változónk van most. A hagyományos lineáris modell a következőképp nézett ki: $Y_j = b_0 + b_1 X_j + \varepsilon_j$. Az új modellben jelölje $Y_{j,k}$ a függő

változó j-edik megfigyelését, amely a k-adik csoportba tartozik, $X_{j,k}$ és $\varepsilon_{j,k}$ hasonlóan. Az egyes csoportokra jellemző véletlen hatást u_k jelölje. Ekkor $u_k \sim N(0, \sigma_u^2)$. Tehát az új modell:

$$Y_{j,k} = b_0 + b_1 X_{j,k} + u_k + \varepsilon_{j,k}$$

A b_0 és az u_k összege alkotja a random interceptet, vagyis a metszéspont a korábbiakkal ellentétben már függ a véletlentől.

2.2.5.4. Változók centrálása A következő részt a [16] forrás alapján dolgoztam ki.

A multikollinearitásnak két alapvető fajtája van [14]:

- **Szerkezeti multikollinearitás.** Ez a típus akkor fordul elő, ha létrehozunk egy új magyarázó változót az eredetieket felhasználva. Például, ha két változó szorzatát is szeretnénk vizsgálni, a külön-külön vett hatásuk mellett. Ekkor nyilván korrelál a szorzat a két változóval, hiszen az tartalmazza valamilyen szinten őket.
- **Az adatok multikollinearitása.** Ekkor a korreláció a megfigyelt változók között van jelen. Ez a multikollinearitás természetes fajtája, ellentétben az előzővel, ahol mesterségesen került be a modellbe.

A változók centrálása a szerkezeti multikollinearitás csökkentésének egy módja, ahol változók szorzata szerepel a regresszióban. Egy változó centrálása során kivonjuk a megfigyelésekből a mintaátlagot, így az új átlag 0 lesz. Vegyünk egy példát, ahol a függő változót két magyarázó változóval, ill. azok szorzatával próbáljuk előrejelezni. Ekkor a következőképp néz ki a regressziós modell:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_1 X_2 + \varepsilon.$$

X_1X_2 és X_1 korrelációja pedig:

$$r_{(X_1X_2),X_1} = \frac{cov((X_1X_2), X_1)}{D(X_1X_2) \cdot D(X_1)}, \text{ centrálás után: } \frac{cov(((X_1 - \overline{X}_1)(X_2 - \overline{X}_2)), (X_1 - \overline{X}_1))}{D((X_1 - \overline{X}_1)(X_2 - \overline{X}_2)) \cdot D(X_1 - \overline{X}_1)}.$$

Mivel többváltozós normalitás mellett [17]

$$cov(AB, C) = E(A)cov(B, C) + E(B)cov(A, C),$$

ezért

$$\begin{aligned} cov(X_1X_2, X_1) &= E(X_1)cov(X_2, X_1) + E(X_2)cov(X_1, X_1) \\ &= E(X_1)cov(X_2, X_1) + E(X_2)D^2(X_1), \end{aligned}$$

centrálás után:

$$E(X_1 - \overline{X}_1)cov((X_2 - \overline{X}_2), (X_1 - \overline{X}_1)) + E(X_2 - \overline{X}_2)D^2(X_1 - \overline{X}_1).$$

Elméletileg $E(X_1 - \overline{X}_1) = 0$ és $E(X_2 - \overline{X}_2) = 0$, így $cov(((X_1 - \overline{X}_1)(X_2 - \overline{X}_2)), (X_1 - \overline{X}_1)) = 0$, ami miatt a centrálás után X_1 és X_1X_2 korrelációjának várható értéke 0 (X_2 és X_1X_2 esetében ugyanígy járunk el). Habár a gyakorlatban a centrált változók és a szorzatuk korrelációja nem lesz mindig pontosan 0, valóban csökkeni fog az eredeti változók korrelációjához képest. Megjegyzendő, hogy az együtthatók értelmezését nem változtatja meg a centrálás.

2.2.5.5. Torzítás-variancia trade-off A következő részt a [18] forrás alapján dolgoztam ki.

A statisztikai becslő módszereknek két alapvető jellemzőjük van: a torzításuk és a szórásnégyzetük. A torzítás a becslés várható értékének pontosságát méri:

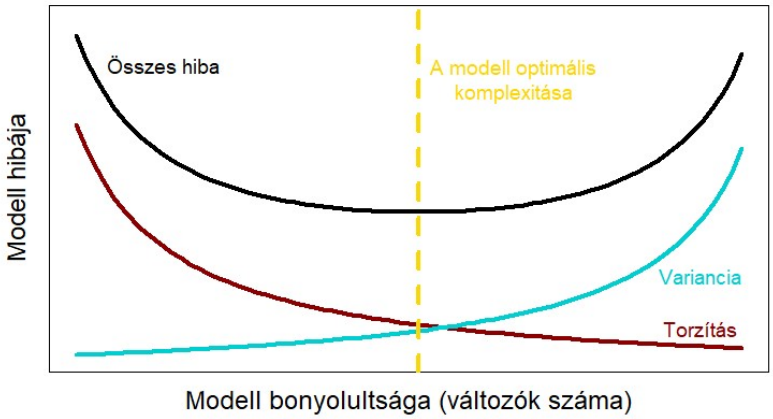
$$Bias(\hat{b}) = E(\hat{b}) - b,$$

míg a szórásnégyzet a becslés bizonytalanságát:

$$D^2(\hat{b}) = \sigma^2(X'X)^{-1}, \text{ ahol } \hat{\sigma}^2 = \frac{\hat{\varepsilon}^2}{n-k}.$$

Szeretnénk, hogy mindkét érték a lehető legalacsonyabb legyen, hiszen a modell hibája három részből tevődik össze: a nagy szórásból, a jelentős torzításból és a maradék, véletlen részből.

$$E(\varepsilon) = (E(X\hat{b}) - Xb)^2 + E(X\hat{b} - E(X\hat{b}))^2 + \sigma^2 = Bias^2 + D^2 + \sigma^2$$



1. ábra. A modell hibájának általános alakulása a változószám függvényében.[1]

A legkisebb négyzetek módszere egy torzítatlan becslést ad a célváltozóra, azonban nagy varianciával rendelkezhet. Ez különösképpen akkor fordulhat elő, ha a magyarázó változók száma megközelíti a megfigyelések számát ($k \to n, (n-k) \to 0, \hat{\sigma}^2 \to \infty$) vagy pedig ha a multikollinearitás másik, "természetes" fajtája fordul elő, amikor a megfigyelt változók korrelálnak egymással. Ekkor a legkisebb négyzetek módszere a 1. ábra jobb oldalán helyezkedik el (ahol alacsony a torzítás és magas a variancia), azonban az optimum ettől balra helyezkedik el. A modell hibáját leredukálhatjuk például regularizációval, vagyis a varianciát némi torzítás ellenében csökkenthetjük. Erre több módszer is létezik.

Ridge regresszió A Ridge regresszió alapötlete, hogy egyes változók együtthatóit a 0 közelében tartjuk, büntetjük őket, ha azok túl nagyok. Így csökken a modell komplexitása, de mégis megtartjuk az összes változót. Ekkor a veszteségfüggvény a reziduálisok négyzetösszegéből és a paraméterbecslések négyzetösszegének büntetéséből áll:

$$L_{Ridge}(\hat{b}) = \sum_{i=1}^{n}(Y_i - X_i\hat{b})^2 + \lambda\sum_{j=1}^{m}\hat{b}_j^2 = \|Y - X\hat{b}\|^2 + \lambda\|\hat{b}\|^2,$$

[1] Az ábrát a [18] forrásban szereplő alapján készítettem.

ahol λ a büntetés paramétere. A kifejezés minimalizálása:

$$\begin{aligned}\frac{d}{d\hat{b}}\left(\|Y - X\hat{b}\|^2 + \lambda\|\hat{b}\|^2\right) &= -2X'Y + 2X'X\hat{b} + 2\lambda\hat{b}\\ -2X'Y + 2X'X\hat{b} + 2\lambda\hat{b} &= 0\\ (2X'X + 2\lambda I)\hat{b} &= 2X'Y\\ \hat{b}_{Ridge} &= (X'X + \lambda I)^{-1}(X'Y).\end{aligned}$$

A regresszió torzítása:

$$\begin{aligned}Bias(\hat{b}_{Ridge}) &= E(\hat{b}) - b = E((X'X + \lambda I)^{-1}(X'Y)) - b = (X'X + \lambda I)^{-1}(X'E(Y)) - b\\ &= (X'X + \lambda I)^{-1}(X'Xb) - b = (X'X + \lambda I)^{-1}(X'X + \lambda I - \lambda I)b - b\\ &= (X'X + \lambda I)^{-1}\left((X'X + \lambda I) - (\lambda I)\right)b - b = b - \lambda(X'X + \lambda I)^{-1}b - b\\ &= -\lambda(X'X + \lambda I)^{-1}b.\end{aligned}$$

A variancia kiszámításához [19] definiáljuk a $W_\lambda = (X'X + \lambda I)^{-1}X'X$ lineáris operátort. Ekkor

$$\begin{aligned}W_\lambda b^* &= W_\lambda(X'X)^{-1}X'Y\\ &= (X'X + \lambda I)^{-1}X'X(X'X)^{-1}X'Y\\ &= (X'X + \lambda I)^{-1}X'Y\\ &= \hat{b}_{Ridge},\end{aligned}$$

ahol b^* az együtthatók maximum likelihood, tehát legkisebb négyzetes becslése. A regresszió varianciája:

$$\begin{aligned}D^2(\hat{b}_{Ridge}) &= D^2(W_\lambda b^*) = W_\lambda D^2(b^*)W_\lambda'\\ &= W_\lambda\sigma^2(X'X)^{-1}W_\lambda' = \sigma^2 W_\lambda(X'X)^{-1}W_\lambda'\\ &= \sigma^2(X'X + \lambda I)^{-1}X'X[(X'X + \lambda I)^{-1}]'.\end{aligned}$$

A következő észrevételeket tehetjük λ-ra vonatkozóan:

- ha $\lambda \to 0$, akkor a Ridge regresszió paraméterbecslése tart a legkisebb négyzetek módszerének paraméterbcsléséhez.
- ha $\lambda \to \infty$, akkor $\hat{b}_{Ridge} \to 0$
- λ minél nagyobb, annál kisebb a variancia és annál nagyobb a torzítás mértéke.

Lasso regresszió A Lasso (Least absolute shrinkage and selection operator) regresszió koncepciója hasonló a Ridge regresszióéhoz. Ebben az esetben is büntetjük a paramétereket, azonban nem a négyzetösszegüket, hanem az abszolútértékük összegét. Tehát a Lasso regresszió veszteségfüggvénye:

$$L_{Lasso}(\hat{b}) = \sum_{i=1}^{n}(Y_i - X_i\hat{b})^2 + \lambda\sum_{j=1}^{m}|\hat{b}_j|.$$

A multikollinearitást különbözőképp kezeli a két számítás: míg a Ridge regresszióban a korreláló magyarázó változók együtthatója hasonló lesz, addig a Lasso-nál az egyik magyarázó változó nagy, a másik (vagy többi) közel 0 együtthatót kap.

Rugalmas háló A rugalmas háló (elastic net) a Ridge és a Lasso regresszió kombinálásából jött létre. A λ mellett egy másik paraméterrel is rendelkezik, melyet α-val jelölünk és a kétfajta regresszió vegyítéséért felel. Ekkor a veszteségfüggvény:

$$L_{\text{Rháló}}(\hat{b}) = \frac{\sum_{i=1}^{n}(Y_i - X_i\hat{b})^2}{2n} + \lambda\left(\frac{1-\alpha}{2}\sum_{j=1}^{m}\hat{b}_j^2 + \alpha\sum_{j=1}^{m}|\hat{b}_j|\right).$$

Látható, hogy $\alpha = 0$ estén a Ridge, míg $\alpha = 1$ esetén a Lasso regressziót kapjuk vissza.

3. A p-érték problémája

Valentin Amrhein (a Bázeli Egyetem zoológia professzora, tudományos újságíró), Sander Greenland (statisztikus és epidemiológus) és Blake McShane (statisztikai metodológus, a Northwestern Egyetem Kellogg School of Management Marketing tanszékének professzora) több mint 800 aláíróval együtt int óvatosságra a statisztikai szignifikancia használatakor [1]. Sohasem szabadna olyan következtetéseket levonni, hogy a vizsgálat tárgyai között nincs összefüggés vagy nincs különbség csupán azért, mert a p-érték nagyobb egy bizonyos küszöbértéknél. Hasonlóan hamis kijelentés lehet az, hogy két tanulmány nem ért egyet, mivel az egyik egy statisztikailag szignifikáns eredményt kapott, a másik pedig nem. Az ilyen és ehhez hasonló állítások nem csak túlzóak, de teljesen megalapozatlanok lehetnek. Vegyünk például egy gyógyszert és egy lehetséges mellékhatást. Két tanulmány is vizsgálja, hogy vajon a gyógyszer tényleg okozza-e a bizonyos tüneteket. Az egyik tanulmány kutatói azt állítják, hogy nem kaptak statisztikailag szignifikáns eredményt, így nincsen összefüggés a vizsgált szer és a hatás között, míg a másik csoportnál nagyon alacsony p-érték jött ki. Egyszerűen a második esetben az elemzések sokkal precízebbek voltak (több megfigyeléssel) ettől lett egyértelműen kimutatható összefüggés, míg az első esetben hamis következtetést vontak le. Ha a p-érték nagyobb egy bizonyos küszöbértéknél, akkor az azt jelenti, hogy nincs elegendő bizonyíték a nullhipotézis elutasítására, de a statisztikailag nem szignifikáns eredményből nem következik, hogy a gyógyszer nem okozhatja a bizonyos mellékhatást. Így túlzás lenne azt állítani, hogy a két tanulmány eredményei ellent mondanak egymásnak.

Az említett írók a p-értékek és a statisztikai szignifikancia megszokott, kétosztatú használatának felhagyására szólítanak fel. Emellett szorgalmazzák a tanulmányok előzetes bejelentését (elemzési tervvel), minden eredmény nyilvánosságra hozatalát és a konfidencia intervallumok összes értékének figyelembevételét, elemzését.

3.1. A hibák forrásai

A p-értékek körül kialakult vitáknak nem csak egy gyökere létezik. Egy hipotézis vizsgálata során több helyen is véthetünk hibákat, az első lépésektől egészen az utolsókig.

A statisztikai döntéselméletben és a statisztikai következtetés elvei szerint egy tesztelési eljárás nem érvényes, ha nincsen egy előzetesen meghatározott, az adatoktól független döntési szabály. Hagyományosan úgy választjuk meg ezt a döntési szabályt, hogy az egy felső korlátot szabjon meg az elsőfajú hiba nagyságára. Fontos, hogy az elsőfajú hiba valószínűségére adott korlát a szabályt határozza meg, a p-érték pedig magát a döntést, a kettőt nem szabad összetéveszteni. Azonban előfordul, hogy a p-értékeket döntési szabályként alkalmazzák, de mivel az függ a vizsgált adatoktól, ezért innentől nem lehet érvényes következtetéseket levonni, a hipotézisről szóló döntés jogtalan lesz. [2]

Például szeretnénk vizsgálni a következő nullhipotézist: egy adott, feltehetően normális eloszlású minta 0 várható értékkel rendelkezik. Először rögzítsük le, hogy az elsőfajú hiba bekövetkezésének valószínűsége legfeljebb 5 % legyen. A döntési szabály: ha az elsőfajú hiba bekövetkezésének valószínűsége nagyobb, mint 0,05, akkor nem tudjuk elvetni a nullhipotézist míg, ha a választott küszöbnél kisebb ez az érték, akkor dönthetünk amellett, hogy a minta eloszlásának várható értéke nem 0. Magát a döntést a p-érték nagysága határozza meg, miután kiszámoltuk az adatokból, összehasonlítjuk a korábban lerögzített küszöbértékkel. Nem a p-érték kiszámítása után szabjuk meg a döntési szabályt, vagyis választunk egy korlátot az elsőfajú hiba nagyságára! Az alkalmazott statisztikai közösségben egy teszt elvégzésekor általában három különböző típusú ember létezik, a következő módszereket alkalmazva: [2]

- Az első, mielőtt látná az adatokat, meghatároz egy α küszöbértéket, majd kiszámítja a p-értéket. Végül pedig dönt az α és a p-érték összehasonlítása után.
- A második nem választ küszöbértéket, hanem kiszámolja a p-értékét, és végül azt állítja, hogy ha választott is volna α-t, az nagyobb lenne az általa kapott p-nél.
- A harmadik pedig szintén a p-érték kiszámításával kezd, és ha úgy gondolja, hogy az kicsi, akkor elutasítja a nullhipotézist. Majd a második tesztelő megközelítését folytatva választ egy p-nél nagyobb küszöbértéket, hiszen úgy véli bármilyen ilyen α megfelelő lesz, ahhoz hogy szignifikáns eredményt kapjon.

Látható, hogy csak az első tesztelő végzi el helyesen a vizsgálatot. A 2. és 3. esetben az alanyok pusztán a p-érték alapján vonnak le következtetéseket (esetleg utólag megválasztva a küszöbértéket), ami már alapvetően egy hibás elképzelés.

Tegyük fel, hogy minden szabályt betartva elvégeztük a vizsgálatot és kiszámítottuk a p-értéket. Ezt követően jön az értelmezés. Mit is jelent pontosan a kapott eredmény az adatainkra nézve? A megszokott módon, ha $p < \alpha$, akkor elutasítjuk a nullhipotézist, ha pedig az egyenlőtlenség fordítva áll fenn, akkor nem tudjuk elutasítani a nullhipotézist (ez nem azt jelenti, hogy az ellenhipotézis hamis, vagy hogy a nullhipotézis igaz). Érezhetjük, hogy az a két eset, amikor $p = 0,049$ és $p = 0,051$ nem áll messze egymástól. Azonban egy $\alpha = 0{,}05$ mellett, mégis ellentétes döntés születhet.

Az eredmények statisztikailag szignifikáns és nem szignifikáns kategóriákba sorolása arra késztetheti az embereket, hogy ez a két eset erősen, egymásnak ellentmondva különbözik. Tudósok és folyóirat szerkesztők is beleesnek abba a hibába, hogy a statisztikailag szignifikáns eredményeket előnyben részesítik, azon tévhit okán, hogy mindössze egy küszöbérték átlépése elengedő ahhoz, hogy érvényes döntést hozzunk. Ebből kifolyólag sokszor túl nagy jelentőséget tulajdonítunk a statisztikailag szignifikáns eredményeknek és leértékeljük azokat, amelyek nem szignifikánsak. Sajnos ez a fajta felfogás egy újabb problémának is kiváltója: arra bátoríthatja a kutatókat, hogy úgy válasszák meg a vizsgálati módszereiket és az adatokat, hogy azok "statisztikailag" alátámasszák az előzetesen kívánt eredményt. [1]

3.2. Javaslatok a vitás helyzetek elkerülésére

3.2.1. A p-értékek megfelelő használata

A p-érték egy hasznos statisztikai mutató lehet, ha azt helyesen, bizonyos szabályokat betartva használják. Az Amerikai Statisztikai Szövetség (American Statistical Association

- ASA) 2016-ban kiadott egy nyilatkozatot [3], melyben tisztáz néhány széles körben elfogadott alapelvet a p-érték és a statisztikai szignifikancia használatáról. Ezen iránymutatók megfogadása hozzájárul ahhoz, hogy megalapozottabb következtetéseket tudjunk levonni és kezelni a fennálló bizonytalanságokat. A nyilatkozat a következő pontokat tartalmazza:

1. **A p-értékek megmutathatják, hogy az adatok mennyire összeegyeztethetetlenek egy meghatározott statisztikai modellel.** Általában olyan modellről van szó, amely egy nullhipotézis és néhány feltételezés alapján épül fel. Minél kisebb a p-érték, annál nagyobb az adatok statisztikai összeférhetetlensége a nullhipotézissel, ha a p-érték kiszámításához használt alapfeltevések igazak. Ez az összeegyeztethetetlenség úgy értelmezhető, hogy kétségbe vonja vagy bizonyítékot szolgáltat a nullhipotézis vagy az alapfeltételezések ellen.

2. **A p-értékek nem mérik annak a valószínűségét, hogy a vizsgált hipotézis igaz, vagy annak a valószínűségét, hogy az adatokat kizárólag a véletlen alkotta.** A kutatók gyakran kívánják azt, hogy a p-értéket át lehessen változtatni egy, a hipotézis igaz voltára vonatkozó állítássá vagy valószínűséggé, hogy az adatokat a véletlen hozta létre. Azonban a p-érték ezek közül egyik sem. A p-érték az adatokról tesz kijelentést egy meghatározott hipotetikus magyarázathoz viszonyítva, és nem magáról a magyarázatról.

3. **Nem szabadna a tudományos következtetéseket és az üzleti vagy politikai döntéseket csak arra alapozni, hogy a p-érték átlép-e egy bizonyos küszöbértéket.** Azok a gyakorlatok, amelyek tudományos állítások igazolására az adatelemzést olyan egyszerű szabályokra redukálják, mint például $p < 0,05 =$ szignifikáns eredmény, $p > 0,05$ pedig nem szignifikáns, téves meggyőződésekhez és rossz döntéshozatalokhoz vezethetnek. Az alkalmazásban gyakran egy "igen-nem" válaszra van szükség, azonban a p-érték önmagában nem elegendő egy kérdés megválaszolására, hiszen egy következtetés nem válik rögtön igazzá vagy hamissá egy bizonyos határ egyik, ill. másik oldalán. A kutatóknak számos más tényezőt is figyelembe kell venniük, ahhoz hogy megalapozott tudományos következtetéseket tudjanak levonni, beleértve a vizsgálat megtervezését, a mérések minőségét, a vizsgált jelenség külső bizonyítékait és az adatelemzés alapjául szolgáló feltételezések érvényességét is. A statisztikai szignifikancia önmagában, mint bizonyíték egy állítás mögött, a tudományos folyamatok torzulásához vezet.

4. **A helyes következtetés teljes körű tájékoztatást és átláthatóságot igényel.** A p-értékek a hozzájuk kapcsolódó elemzések nélkül értelmezhetetlenek. Ha egy kutató a statisztikai eredmények alapján választja ki, hogy mit prezentál, akkor jelentősen fennáll az eredmények helytelen értelmezésének veszélye, ha az olvasót nem tájékoztatják a választásról és annak alapjairól. A kutatóknak nyilvánosságra kellene hozniuk az egy tanulmány során vizsgált hipotézisek számát, minden adatgyűjtési döntést, minden elvégzett statisztikai elemzést és minden kiszámított p-értéket. A p-értékeken és a hozzájuk kapcsolódó statisztikákon alapuló jogos tudományos következtetések levonásához szükség van legalább azt tudni, hogy hány és milyen elemzéseket végeztek, és hogy ezeket az elemzéseket hogyan választották ki.

5. **A p-érték vagy a statisztikai szignifikancia nem méri egy hatás nagyságát vagy az eredmény jelentőségét.** A statisztikai szignifikancia nem egyenlő

a tudományos, emberi vagy gazdasági szignifikanciával. Ahogy a kisebb p-értékek sem feltétlenül jelentik a nagyobb vagy fontosabb hatások jelenlétét, úgy a nagyobb p-értékek sem vonják maguk után egy hatás vagy a jelentőség hiányát. Bármilyen apró jelenség tud kis p-értéket generálni, ha a vizsgált minta mérete és/vagy a mérés pontossága elegendően nagy. Ugyanígy lehet egy tényezőnek bármilyen erős hatása, ha a mérések pontatlanok és/vagy a minta mérete kicsi, akkor nagy p-értéket kaphatunk. Hasonlóképpen egy azonos hatás jelenlétének vizsgálata különböző p-értékéket produkálhat, ha a különböző becslések pontossága eltérő.

6. **A p-érték önmagában nem nyújt megfelelő bizonyítékot egy modellre vagy hipotézisre vonatkozóan.** A kutatóknak fel kell ismerniük, hogy a p-érték kontextus vagy más bizonyíték nélkül csak korlátozott információt nyújt. Például egy 0,05 körüli p-érték önmagában csak gyenge bizonyítékot nyújt a nullhipotézis ellen. Azonban egy viszonylag nagy p-érték nem jelent bizonyítékot a nullhipotézis mellett, több más feltevés is ugyanúgy vagy még jobban is megfelelhet a megfigyelt adatoknak. Ezen okok miatt az adatelemzés nem érhet véget a p-érték kiszámításával, ha más megközelítések is alkalmasak lehetnek.

Az ASA nyilatkozata megteszi az első lépést a javuláshoz vezető úton, ugyanis az alapok megértése elengedhetetlen a fejlődéshez. Mindemellett többnyire arról volt szó, hogy mit nem szabad tennünk és hinnünk. Az elkerülendő dolgok ismerete szükséges, de nem feltétlenül elégséges a szóban forgó probléma orvoslására.

3.2.2. Túl a hagyományokon

Egyre több tudós és kutató ért egyet azzal, hogy a hipotézisek megszokott módon való tesztelése, az eredmények értelmezése és közzététele nem megfelelő, ezért változásra van szükség. Nem létezik olyan módszer, amely tökéletes lenne, vagy amelyik egyik pillanatról a másikra megszüntetne minden megkérdőjelezhető esetet, de javítani lehet a jelenleg kialakult helyzeten.

Számos statisztikus, statisztikai szakkönyvek szerzője, statisztikai folyóratok szerkesztője, statisztikai gyakorlatok bírálója és más statisztikával foglalkozó tudós ért egyet azzal, hogy kerülni kellene a "statisztikai szignifikancia" kifejezés használatát (egyébként sem ad hozzá plusz értéket ahhoz, amit a p-érték már eleve hordoz), a p-értékeket azok felcimkézése nélkül kellene megjelentetni [4]. Magát a kifejezést is gyakran rosszul értelmezik, sokszor összemosódik a szignifikáns, mint fontos/jelentős hétköznapi értelmezésével. Azonban ahogy az ASA is megfogalmazta a 2016-os nyilatkozatának 5. pontjában, a szignifikancia szó nem ugyanazt jelenti a statisztikában és más területeken.

Ugyanakkor nem csak a szóhasználattal van probléma. A "Scientists rise up against statistical significance" c. cikk írói a statisztikai szignifikancia teljes koncepciójának elhagyására szólítanak fel. A kategorizálásból való kilépés segíthet csökkenteni a túlzó és leegyszerűsített állításokat, mint például hogy két csoport között nincs különbség. Nem a p-értékek és egyéb statisztikai mutatók kiszámításával van baj, hanem azok éles kettéválasztásával. A *The American Statistician* tudományos folyóirat egyik 2019-ben megjelent cikkében [4] azt javasolják a folyóiratok szerkesztőinek, hogy a következőhöz hasonló módon, a kategorizálás elkerülésére is hívják fel a szerzők figyelmét:

> „Ma már széleskörű egyetértés van számos, a témát tanulmányozó, statisztikus között abban, hogy a p-értékeket szolgáltató statisztikai tesztek közlésekor lo-

gikátlan és helytelen a p-skála dichotomizálása, valamint az eredmények "szignifikáns" és "nem szignifikáns" megnevezése. Határozottan ellenezzük, hogy a szerzők folytassák ezt, a sohasem igazolt gyakorlatot, amely a modern statisztika korai történetében kialakult zűrzavarokból származik."

Hogyan juthatunk el egy $p < 0,05$-ön túlmutató világba? Ezt a kérdést tanulmányozza Wasserstein, Schirm és Lazar egy 2019-ben megjelent cikkben [5] megannyi más, a témában íródott publikációt összegezve. Javaslataikat a következőképp foglalják össze: „Fogadjuk el a bizonytalanságot. Legyünk meggondoltak, nyitottak és szerények."

El kell ismernünk, hogy egy vizsgálat végén nem tudunk 100%-ig biztos eredményt mondani, viszont megpróbálhatjuk ezt a bizonytalanságot megragadni, kezelni, de nem eltüntetni! Ennek egy módja lehet az, hogy a konfidencia intervallumokra "kompatibilitási intervallumok"-ként hivatkozunk [6]. Meg kell vizsgálni, hogy az intervallumon belüli értékek (a határértékeket is beleértve) milyen jelentéssel bírnak az adatainkra nézve, hiszen azok mind összeegyeztethetőek az alapfeltevéseinkkel [1]. Amrhein és társai négy dologra hívja fel a figyelmet, melyet észben kell tartani, amikor kompatibilitási intervallumokról beszélünk. Először is az intervallumon kívül eső értékek nem teljesen összeegyeztethetetlenek az alapfeltevésekkel, csak kevésbé, mint az azon belüli értékek. Így nem mondhatjuk azt, hogy egy intervallum az összes lehetséges értéket magába foglalja. Másodszor tudni kell azt is, hogy a belső értékek sem egyformán kompatibilisek az adatokkal a feltételezéseink alapján. Harmadszor a 95 %-os szint is önkényesen meghatározott, nem feltétlenül ez a megfelelő minden esetben. Végül, de nem utolsó sorban pedig figyelembe kell venni, hogy az összeegyeztethetőségi becslések függenek az intervallum meghatározásához használt statisztikai feltételezések helyességétől. A feltevéseinket minél pontosabban kell megfogalmaznunk, más alternatívákat is tesztelve, valamint az összes eredményt közzé kell tenni, nem csak az általunk legkedveltebbeket.

A bizonytalanság elfogadásából következik, hogy legyünk megfontoltak minden helyzetben. Egy tanulmányhoz az elejétől a végéig, a tervezéstől az elvégzésen át az eredmények elemzéséig, szükség van precizitásra és minden eshetőség figyelembe vételére. Anderson a következő kérdéseket fogalmazta meg, melyek megválaszolása fontos egy kutatási eredmény elemzésekor: Milyen gyakorlati következményei vannak a becslésnek? Mennyire pontos a becslés? Helyesen specifikálták a modellt? Utóbbiba beleértve a feltevések érthetőségét, érvényességét, és hogy a főbb eredmények más modellezési döntések mellett is megállják-e a helyüket. Ahhoz, hogy a teszteléseket a független személyek is meg tudják ismételni, ill. objektív kritikát megfogalmazni, minden egyes döntést és választást kellően dokumentálni kell [7].

Egy átgondolt kutatás többek között figyelembe veszi a korábbi kutatások eredményeit, a tudományos kontextust, nagy hangsúlyt fektet az adatok megbízható előállítására, ill. ellenőrzi a vizsgálat során használt módszerek hitelességét. Azonban még az adatgyűjtés és azok elemzése előtt fontos lenne meghatározni, hogy milyen eredmények jöhetnek számításba, és hogy az adott esetben mikor lesz egy hatás jelentős. Utólag már könnyebb magyarázatokat találni, és fennáll az a veszély is, hogy egy triviális hatásméretet jelentősnek ítélünk meg.

A következő lépés a nyitottság. Egy tanulmány lefolytatása során nem feltétlenül zajlik minden úgy, ahogy azt előre megterveztük, ennek ellenére befogadónak kell lenni az újra. Előfordulhat például, hogy egy vizsgálat folyamán előre nem sejtett összefüggésekre leszünk figyelmesek, amik újabb, fontos kérdéseket vethetnek fel. Nyitottnak kell lenni a kritikára is. A döntések meghozatalában jelen van a szubjektivitás, akármennyire is

törekszünk az objektivitásra, ezért lényeges lehet meghallgatni, és persze elfogadni egy kívülálló szakértő véleményét is.

Nemcsak a nyitottság, de a nyíltság is fontos szerepet játszik a statisztikai teszteléseknél. Egy eredmény nem tűnik elegendően megbízhatónak, ha csupán csak a nullhipotézis elutasítására vonatkozóan kapunk információt. A felhasznált adatok forrása, mennyisége, az alkalmazott módszerek, a p-értékékek és más mutatók stb. mind olyan tényezők, melyeknek nyílvánosnak kellene lenniük. Ha ez teljesül, akkor nem egy ember vagy csoport következtetésére kell támaszkodni, nem fog torzulni az eredmény, és a vizsgálat is könnyen reprodukálható lesz, más kutatók is tudnak végezni alternatív elemzéseket.

Végül eljutunk a szerénységhez. Tudatában kell lenni és persze nyilvánosan is el kell ismerni egy-egy kutatás korlátait. A valóság általában jóval összetettebb, mint bármilyen modell, amellyel megpróbáljuk azt elemezni. A statisztika eszközeinek is megvannak a maguk korlátai, gyakran egy p-érték önmagában nem nyújt elegendő információt egy hipotézis elfogadására vagy elutasítására. A túlzott magabiztossággal tett kijelentések félrevezethetnek. Egy vizsgálat olvasói különböző okokból tanulmányozhatják az elemzéseket, így jobb lehet az eredmények közlésénél minél objektívebbnek, semlegesnek maradni.

4. Az előfeltételek megsértése

Egy statisztikai teszt elvégzésekor figyelmet kell fordítani arra, hogy az adatok teljesítike a használandó módszer alapjául szolgáló előfeltételeket. Ezek megsértése megbízhatatlan p-értékeket eredményez, és még nagyon kicsi kapott érték mellett sem lehetünk biztosak abban, hogy tényleg egy valódi hatást mutattunk ki. Fontos megjegyezni, hogy a p-értékek befolyásoltságának mértéke függ attól, hogy mely feltétel (vagy feltételek) és milyen mértékben sérül (vagy sérülnek) meg.

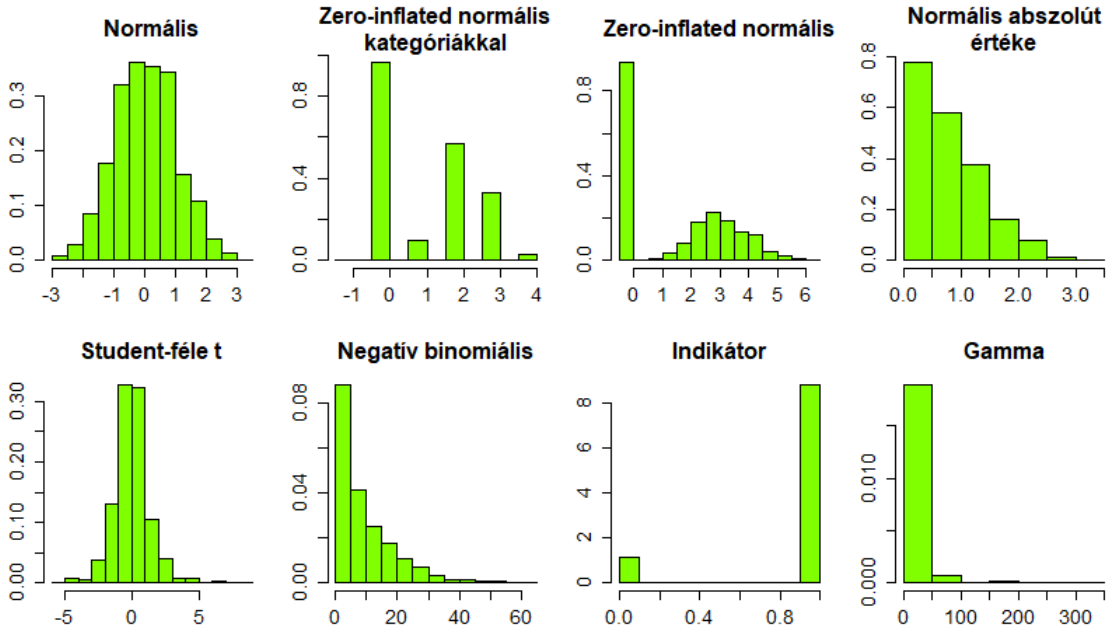
4.1. Normalitás

Mi történik, ha a minta nem normális eloszlású, de a használt teszt mégis Gauss-féle hibát feltételez? Hogyan befolyásolja a vizsgálat eredményét, mennyit változtat a p-értékeken? Knief és Forstmeier [8] azt állítja, hogy gyakran "kevésbé rossz" döntés lehet az, ha a kutatók Gauss modelleket illesztenek nem normális eloszlású adatokra, mint hogy más hibaeloszlású (pl. Poisson, binomiális) modelleket vagy randomizációs technikákat (pl. bootstrapping) alkalmaznak. Utóbbi esetek számos, sokszor nem eléggé ismert kockázattal járnak.

Reprodukáltam a [8] forrásban szereplő kísérletet, mely során a normalitási feltétel megsértésének hatását vizsgálom a p-értékekre. A cél választ kapni arra a kérdésre, hogy mekkora problémát jelent az imént említett feltétel megszegése. A kísérlet során változók függetlenségét tanulmányoztam, melyekről tudni, hogy valóban függetlenek egymástól. Ehhez lineáris regressziós modellt illesztettem az adatokra, mindig kiválasztva egy függő, ill. egy magyarázó változót és elmentettem a kapott p-értéket. Minden esetben a modell felépítésekor azt feltételeztem, hogy a változók normális eloszlásúak, annak ellenére, hogy nem minden esetben azok. Saját futtatásaimhoz a forrásban szereplő 10 különböző eloszlású változóból 8-at használtam fel, melyek közül minden lehetséges párosításra (összesen 64) végrehajtottam az elemzést. Valamint a futási idő csökkenésének érdekében minden esetben 50 000 helyett 5 000 szimulációt végeztem el. A felhasznált eloszlásokat, melyeket a 1. táblázat tartalmazza, a 2. ábrán láthatóak.

	Eloszlások	Paraméterek
1	(Standard) normális	$m = 0, \sigma = 1$
2	Zero-inflated normális, kategóriákkal	$m = 3, \sigma = 1, p = 0,5, k = 5$
3	Zero-inflated normális	$m = 3, \sigma = 1, p = 0,5$
4	Normális abszolút értéke	$m = 0, \sigma = 1$
5	Student-féle t	$df = 4$
6	Negatív binomiális	$r = 1, p = 1/11$
7	Indikátor	$p = 0,9$
8	Gamma	$\alpha = 0, 1, \lambda = 100$

1. táblázat. A felhasznált eloszlások és paramétereik



2. ábra. Az egyes eloszlású változók hisztogramjai

Az eloszlások között találhatóak diszkrétek és folytonosak, ill. 2 zero-inflated is (vagyis adott % eséllyel a változó 0-t, egyébként pedig az eredeti eloszlásból vesz fel értéket). A táblázatban lefelé haladva az eloszlások növekvő tendenciát mutatnak az erős kiugró értékek keletkezésében, amiről ismert, hogy problémás lehet. A [8] forrás szerzői a Cook-féle távolságot használták az, egy kritikus értéket meghaladó, adatpontok átlagos arányának meghatározására (az i. megfigyeléshez tartozó Cook-távolság: $D_i = \frac{\sum_{j=1}^{n}(\hat{Y}_j - \hat{Y}_{j(i)})^2}{k\sigma^2}$, ahol $\hat{Y}_{j(i)}$ a függő változó j. megfigyelésének becslése abban az esetben, ha a regresszióból kihagyjuk az i. megfigyelést, k a magyarázó változók száma). A számítások menete a következő:

1. Egy-egy adott eloszlásból, adott nagyságú, független minta generálása a függő és magyarázó változóknak.

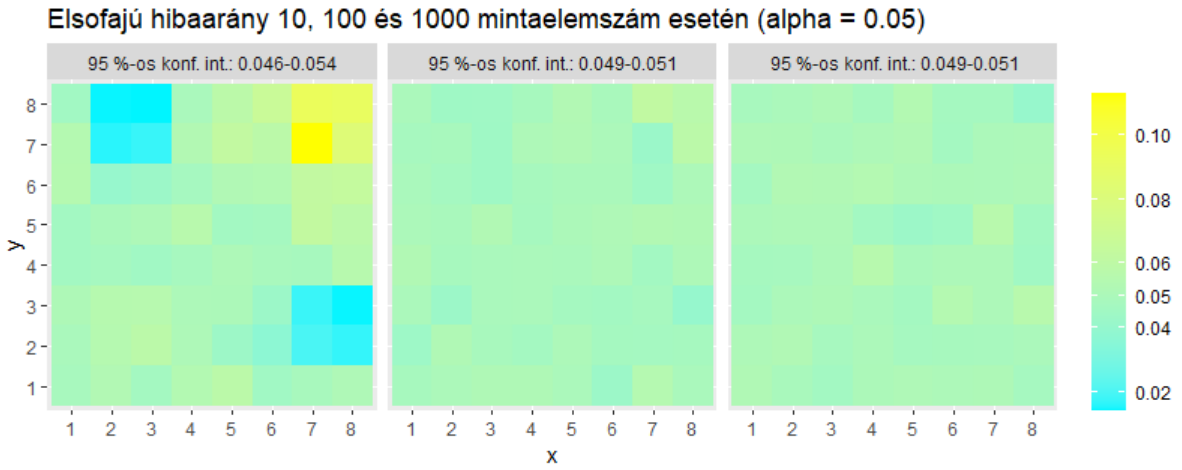
2. A két minta függetlenségének vizsgálata: lineáris regressziós modell illesztése az adatokra, t-próba az együtthatók tesztelésére.

3. A p-értékek elmentése.

4. Az előbbi lépések elvégzése összesen 5000-szer.

5. A korábbi lépések elvégzése 10, 100 és 1000 mintaelemszámra is.

Azt várjuk, hogy egy-egy adott mintaelemszám esetén a p-értékek egyenletes eloszlást kövessenek, tehát az elsőfajú hibaarány adott α érték mellett (az esetek hányad részében jött ki szignifikáns eredmény, azaz a megadott α szintnél kisebb p-érték) megegyezzen a szignifikancia szinttel. Legyen $\alpha = 0,05$, ekkor az elsőfajú hibaarányok a 3. ábráról olvashatók le.

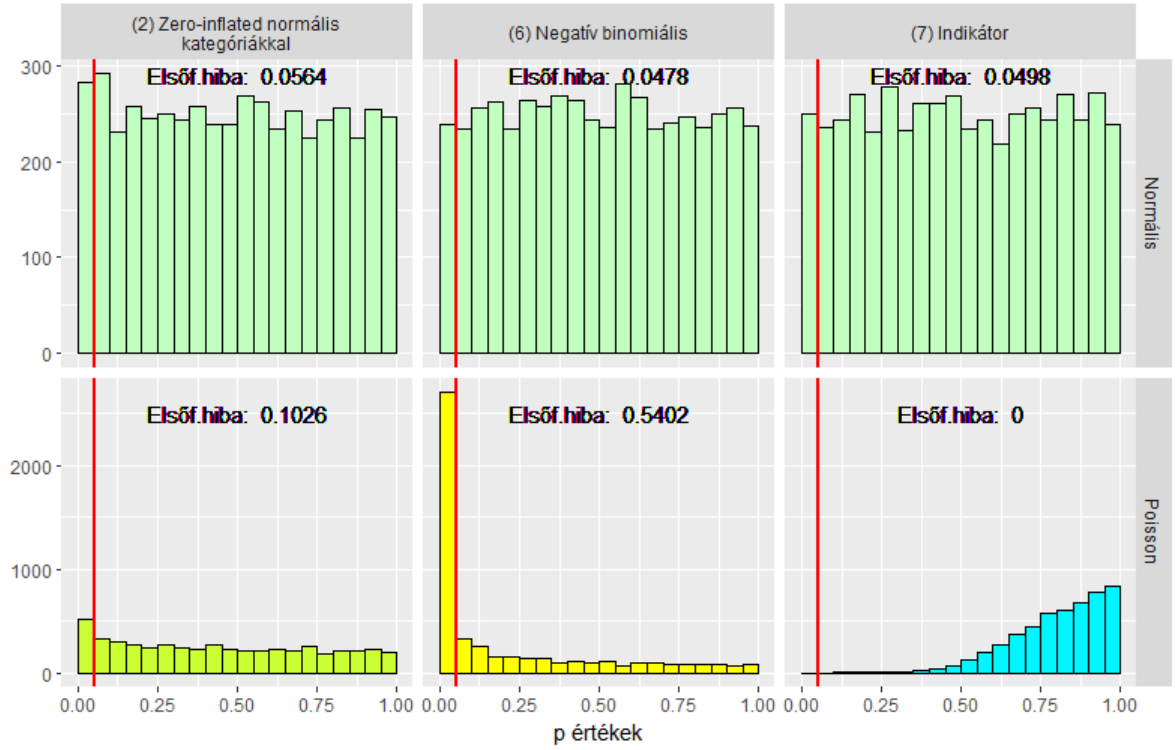


3. ábra. Elsőfajú hibaarányok $\alpha = 0,05$ mellett

A szimulációk elvégzése után látható, hogy a várt és a kapott értékek nincsenek nagyon messze egymástól, a mintaelemszám növelésével pedig egyre szűkebb tartományba összpontosultak a kapott elsőfajú hibaarányok. A legnagyobb eltérés a 10 mintaelemszámos futtatásoknál figyelhető meg, háromszor 4 párosítás esetén. Ezek mindegyikében szerepel a 7-es (indikátor) és vagy a 8-as (gamma) eloszlás. Nem meglepő, hogy ezek okozták a legnagyobb eltérést a várt értéktől, hiszen ahogy már korábban említettem a kiugró értékek problémásak lehetnek. Ezen két változó kombinációi egymással és önmagukkal mind 0,05-nél nagyobb értékeket produkáltak (kb 0,08 és 0,11 közöttieket), vagyis ezekben az esetekben a vártnál többször jött ki szignifikáns eredmény, miszerint a függő és a magyarázó változó nem független egymástól (pedig tudjuk, hogy azok). A másik véglet, amikor ez a két változó a 2-es (zero-inflated normális, kategóriákkal) és a 3-as (zero-inflated normális) eloszlású változókkal párosulnak. Ekkor 0,02 alatti értékek jöttek ki, amik alacsonyabbak a vártnál, tehát kevesebbszer kaptunk szignifikáns eredményeket, mint amire számíthattunk.

Az előző kísérletből láthattuk, hogy a hibák normális eloszlását feltételező modellek lényegesen képesek ellenállni az eloszlási feltétel megsértésével szemben. Ellenben, ha pl. Poisson eloszlást feltételező modellt alkalmazunk, akkor a kapott eredmények jóval szélsőségesebbek lehetnek. Az előzőekben használt változók közül megismételtem a kísérletet a diszkrét eloszlásúakra (mivel a Poisson eloszlás diszkrét, a folytonos változókra nem lehet alkalmazni a modellt), vagyis a 2-esre (zero-inflated normális, kategóriákkal), a 6-osra (negatív binomiális) és a 7-esre (indikátor). A 4. ábrán láthatóak a p-értékek eloszlása a különböző esetekben. Mindegyik szimulációnál a függő és a magyarázó változó azonos eloszlásból származnak, amit az oszlopok jeleznek, a sorok a feltételezett hibaeloszlást

mutatják meg, valamint egy piros egyenes jelzi a használt $\alpha = 0,05$-ös szignifikancia szintet.



4. ábra. A p-értékek eloszlása normális és Poisson hibát feltételező modellek esetén

A hisztogramokon feltüntetett elsőfajú hibaarányokon jelentős különbség figyelhető meg egy-egy eloszlás esetén. Amíg normális eloszlásúnak feltételeztük a hibák eloszlását, addig a p-értékek nagyjából egyenletes eloszlást követtek, közel 0,05-ös elsőfajú hibaaránnyal (a p-értékek egyenletes eloszlásának tesztelésekor (Kolmogorov-Smirnov próbával) 0,1517, 0,5119 és 0,6987 p-értékeket kaptam, amik nagyobbak 0,05-nél, így nem tudjuk elvetni a nullhipotézist, miszerint a szimulációk p-értékei egyenletes eloszlást követnek). Amikor viszont ugyanazt a szimulációt azzal az egy különbséggel végeztem el, hogy a hibák eloszlását Poisson eloszlásúnak feltételeztem, már jóval másabb eredmények jöttek ki. A p-értékek eloszlása a negatív binomiális és az indikátor eloszlású változók esetén egyáltalán nem tekinthető egyenletes eloszlásúnak, ezekben az esetekben az elsőfajú hibaarányok is messze állnak a 0,05-ös szinttől (0,54 és 0), de a 2-es számú változó esetében is a vártnál kétszer akkora értéket kaptunk (a p-értékek egyenletes eloszlásának tesztelésekor mindhárom esetben 0-t kaptam p-értékként, amik kisebbek nemcsak 0,05-nél, hanem szinte bármilyen kicsi terjedelemnél, így elvetjük a nullhipotézist, a szimulációk p-értékei nem egyenletes eloszlást követnek).

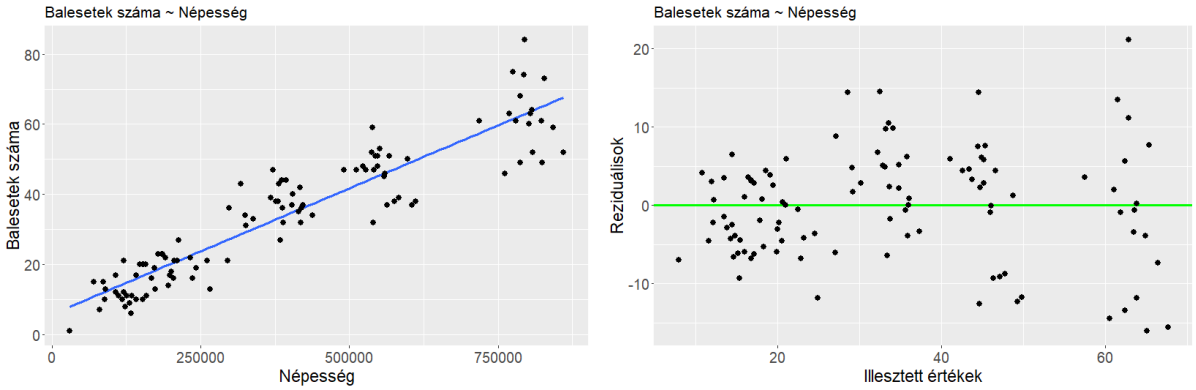
A Poisson-eloszlást használó számításoknál a figyelembe nem vett túlszóródás (a variancia meghaladja az átlagot) jelentős mértékben képes torzítani a p-értékeket, míg a normális hibát feltételező modellek robusztusságát magyarázhatja a centrális határeloszlástétel közelítő teljesülése. Tehát az összehasonlításból azt a következtetést vonhatjuk le, hogy a normalitási feltétel megsértése kevésbé lehet rossz, mint ha egy másik hibaeloszlást feltételező modellnél sérül az eloszlási feltétel.

4.2. Homoszkedaszticitás

Az alfejezetet a [9] forrás alapján dolgoztam ki.

A gyakorlatban sokszor előfordul, hogy egy regresszióelemzés során a reziduálisok nem azonos szórással rendelkeznek, ekkor heteroszkedaszticitásról beszélünk. Ennek számos oka lehet, de a leggyakoribb magyarázat az, hogy a hibák szórása arányosan változik egy bizonyos tényezővel, ami lehet egy hatás, egy jelenség vagy éppen konkrétan az egyik változó. Jellemzően olyan adatoknál lehet ilyen probléma, ahol a megfigyelt értékek széles tartományban helyezkednek el, mivel a nagyobb értékekhez nagyobb szórású reziduálisok társulhatnak.

Vegyük példának a városok népességét és az előforduló balesetek számát. Nyilván egy nagyobb városban rendszerint több baleset fordul elő, mint egy kisebb településen, azonban minél nagyobb egy település, annál változékonyabb az előforduló balesetek száma. (Ennél a modellnél akár alkalmas lehet, az előző fejezetben vizsgált, Poisson eloszlású reziduálisokat feltételezni.) A következő szimulációkhoz a [9] forrásbéli (fiktív) adatokat használtam fel. A 5. ábrán az alapadatokon végzett regressziószámítás eredménye látható. Megfigyelhető, hogy az adatpontokra illesztett regressziós értékek növekedésével a reziduálisok szórása is nő. A heteroszkedaszticitás jelenlétét Breusch-Pagan próbával teszteltem, ahol p-értéknek $1,499 \cdot 10^{-5}$-t kaptam, ami határozottan elutasítja a homoszkedaszticitást állító nullhipotézist.



(a) Az adatpontok és a regressziós egyenes

(b) A reziduálisok eloszlása

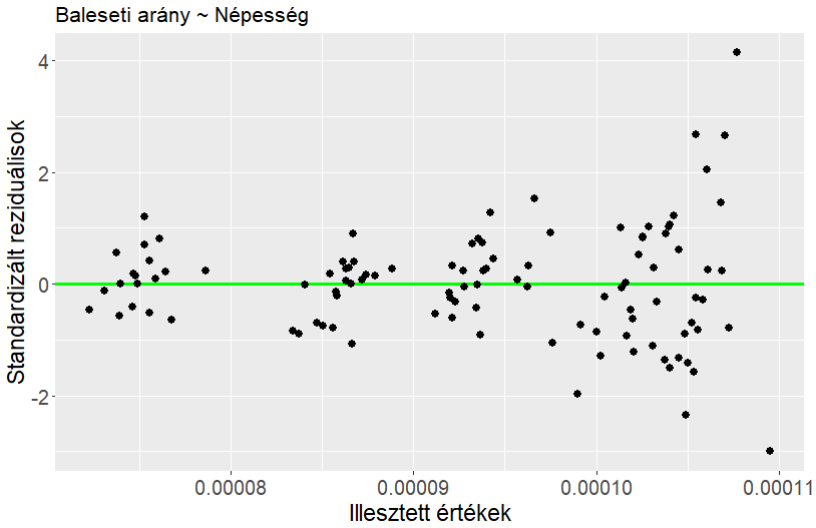
5. ábra. Az alapadatokon végzett regresszió

Miért probléma a heteroszkedaszticitás? Ha a hibák szórása nem konstans, akkor kevésbé megbízható eredményeket kaphatunk. Amellett, hogy kevésbé lesz pontos a paraméterbecslés, kisebb p-értékek jöhetnek ki a kelleténél. A heteroszkedaszticitás növeli az együtthatóbecslések szórását, azonban a legkisebb négyzetek módszere nem veszi figyelembe ezt a növekedést, így a t-értékek kiszámítására, a valódinál alacsonyabb varianciaösszeggel kerül sor. Ennek hatására szélsőséges esetben előfordulhat, hogy statisztikailag szignifikáns eredményt kapunk, pedig valójában az a bizonyos hatás nem is szignifikáns erejű.

A heteroszkedaszticitás csökkentésére többféle módszer is létezik, de hogy melyik a legjobb, az minden esetben a használandó adatoktól függ. Ha nem konstans a reziduálisok szórása, akkor a következőket tehetjük:

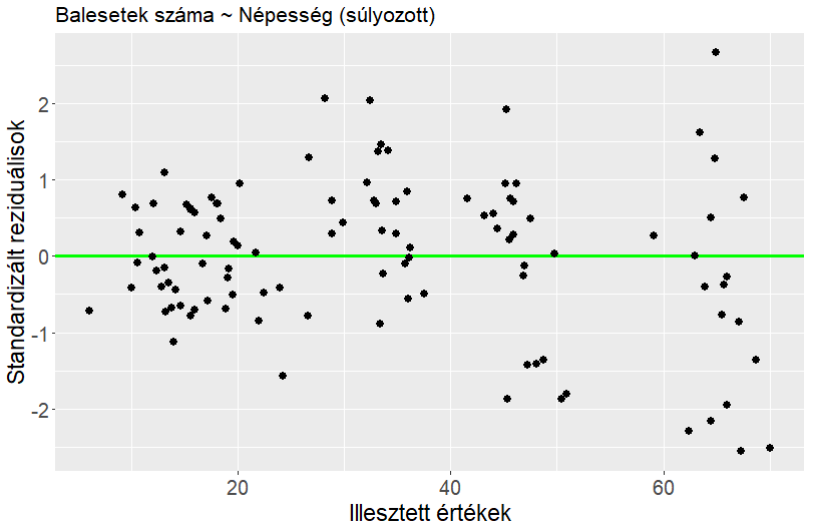
- **A változók újradefiniálása.** Ha egy változó egyes megfigyelései nagyon kicsik, míg mások nagyon nagyok, akkor a különbségek csökkentése érdekében használha-

tunk az eredeti értékek helyett arányokat. Visszatérve a példára, a balesetek száma helyett megvizsgálhatjuk a baleseti arányt a népesség tekintetében (balesetek száma/népesség). Ebben az esetben az reziduálisok még mindig heteroszkedaszticitást mutatnak, de az elsőhöz képest egy kicsit nagyobb p-értékkel ($1,028 \cdot 10^{-4}$).



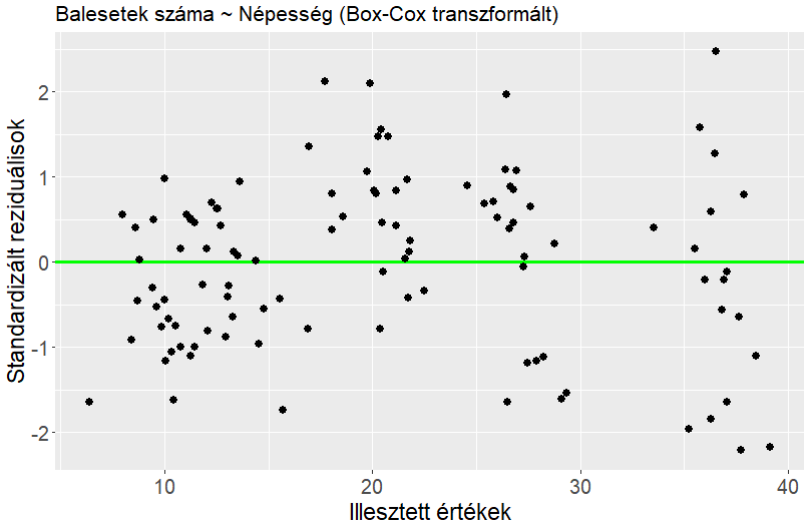
6. ábra. Reziduálisok eloszlása a baleseti arány becslése esetén

- **Súlyozott regresszió.** Ezen módszer használatakor minden egyes adatponthoz egy súlyt rendelünk, méghozzá a súlyozás nélküli modell reziduálisai alapján. Minél nagyobb varianciával rendelkezik egy megfigyelés, annál kisebb súlyt kap, így csökken a reziduálisok négyzetösszege. A példa folytatásaként, mivel a reziduálisok szórása növekszik a népesség növekedésével, ezért a súlyok legyenek a népességi adatok reciprokai úgy módosítva, hogy a súlyok összege megegyezzen a mintaelemszámmal. Ez a megoldás már kicsit jobb az előzőnél, az ábra alapján csökkent a heteroszkedasztikusság mértéke, azonban még mindig szignifikáns, $2,461 \cdot 10^{-3}$-nal egyenlő p-értéket mutat a Breusch-Pagan teszt.

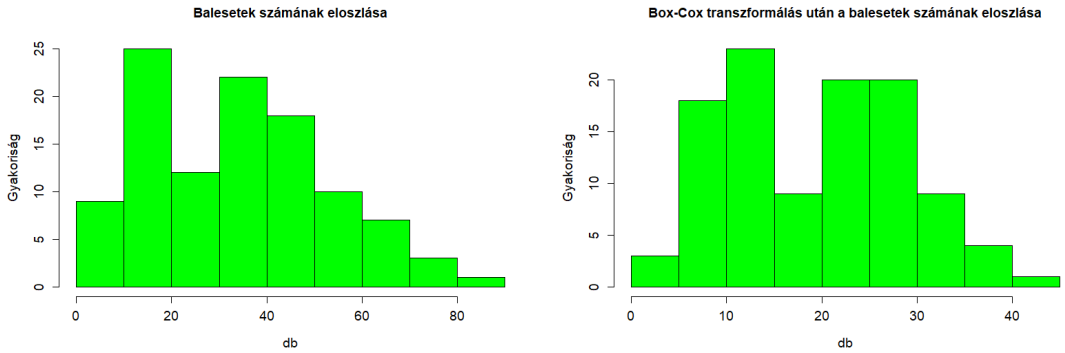


7. ábra. Reziduálisok eloszlása súlyozott regressziót használva

- **A változók transzformációja.** Megoldás lehet a homoszkedaszticitás elérésére, ha a változókat átalakítjuk, ezzel csökkentve a megfigyelések értékei közötti hatalmas különbségeket. Végül alkalmazzunk a függő változóra Box-Cox transzformációt. Ebben az esetben $\lambda = 0,8$ lett. Az előző megoldáshoz hasonló ábrát kaptunk, p-értékként pedig $1,927 \cdot 10^{-3}$ jött ki.



8. ábra. Reziduálisok eloszlása Box-Cox transzformációt használva

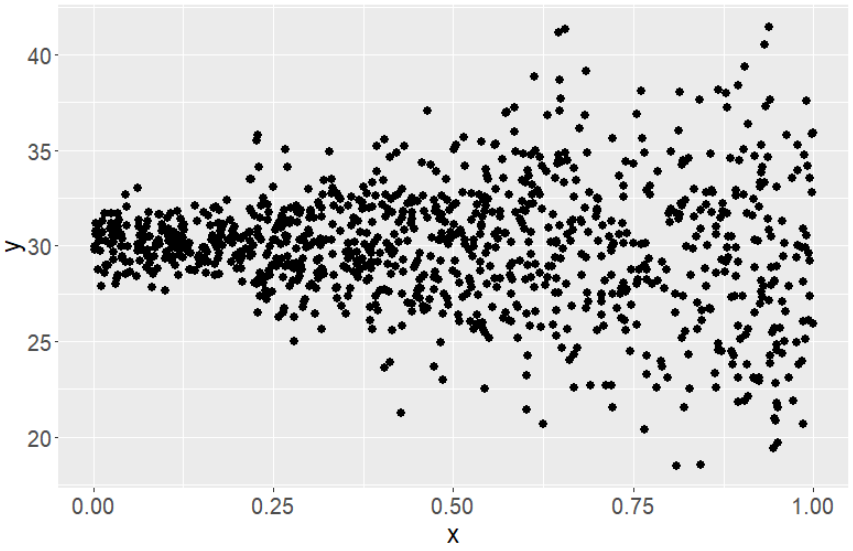


9. ábra. A balesetek számának eloszlása transzformálás előtt és után

A súlyos heteroszkedaszticitás okozhat problémát a számítások során, azonban nincs szükség mindig korrekcióra. A példa esetében nem sikerült homoszkedaszticitást elérni, de csökkent a heteroszkedaszticitás mértéke. Azonban mind a négy variáció nagyon alacsony regressziós p-értéket ($3,03 \cdot 10^{-46}$, $3,13 \cdot 10^{-5}$, $6,97 \cdot 10^{-49}$, $6,61 \cdot 10^{-46}$) produkált, a reziduálisok eloszlásától függetlenül szignifikáns lett az eredmény. Továbbá megfigyeltem az egyes regressziók reziduálisainak szórását (visszatranszformálás után, hogy összehasonlíthatóak legyenek): az alapadatokkal végzett számításoknál 7,05, a baleseti arány becslésénél 6,82, a súlyozott regressziónál 7,16, a Box-Cox transzformált adatok esetén pedig 7,30 jött ki. A különböző módszerek csak pár tizedesjeggyel csökkentették, ill. növelték meg a reziduálisok szórásának nagyságát. Tehát a becslések pontosságában csak kis mértékű változást idéztek elő.

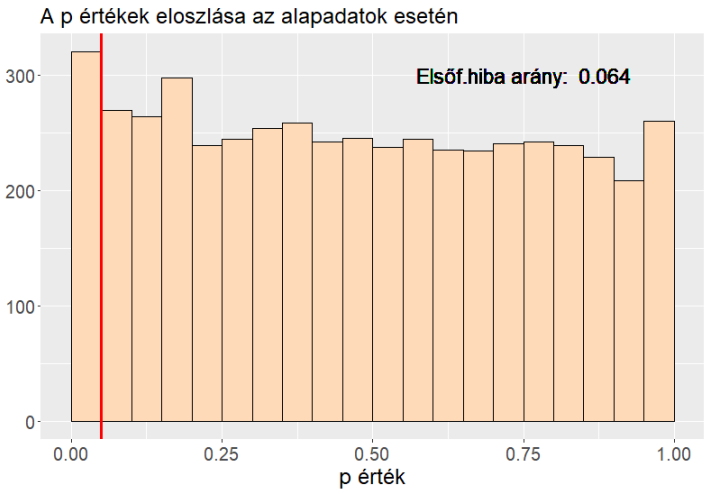
Saját szimuláció

Megvizsgáltam a Box-Cox transzformáció és a súlyozott regresszió hatását a p-értékekre, egymástól független mintákra. Egy-egy szimuláció során a függő változóhoz 1000 db, egymástól független adatpontot generáltam, 30 várható értékű és 1, 2, 3, 4 ill. 5 szórású normális eloszlásból (mindegyik szórással 200-200 db-ot). A magyarázó változóhoz pedig független 1000 db, növekvő sorrendbe állított, [0, 1] intervallumbéli egyenletes eloszlású adatpontot használtam fel. Egy ilyen adathalmaz a 10. ábrán látható.

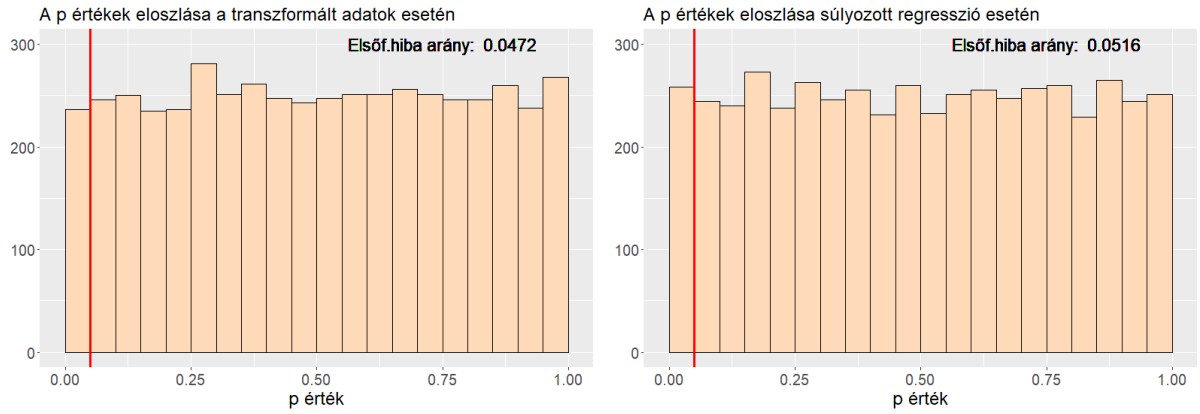


10. ábra. Y: növekvő szórású normális, X: egyenletes eloszlású adatpontok

A generált mintákon háromféle regressziószámítást hajtottam végre: először az eredeti adatokat, másodszor a függő változó Box-Cox transzformáltját, harmadszor pedig súlyokat használva. Súlyoknak a szórásnégyzetek reciprokát vettem. Összesen 5000-szer végeztem el a szimulációkat, majd megnéztem, hogy az egyes esetekben a p-értékek hányadrésze esett 0,05 alá, vagyis mekkora a szignifikáns elsőfajú hibák aránya. Mivel a függő és a magyarázó változó valójában független egymástól, így a kapott p-értékektől azt várjuk, hogy egyenletes eloszlást kövessenek 0 és 1 között, vagyis a vizsgált arányszámnak 0,05-tel kell megegyeznie. A p-értékek eloszlása a 11. és a 12. ábrán láthatóak



11. ábra. p-értékek eloszlása az alapadatok esetén



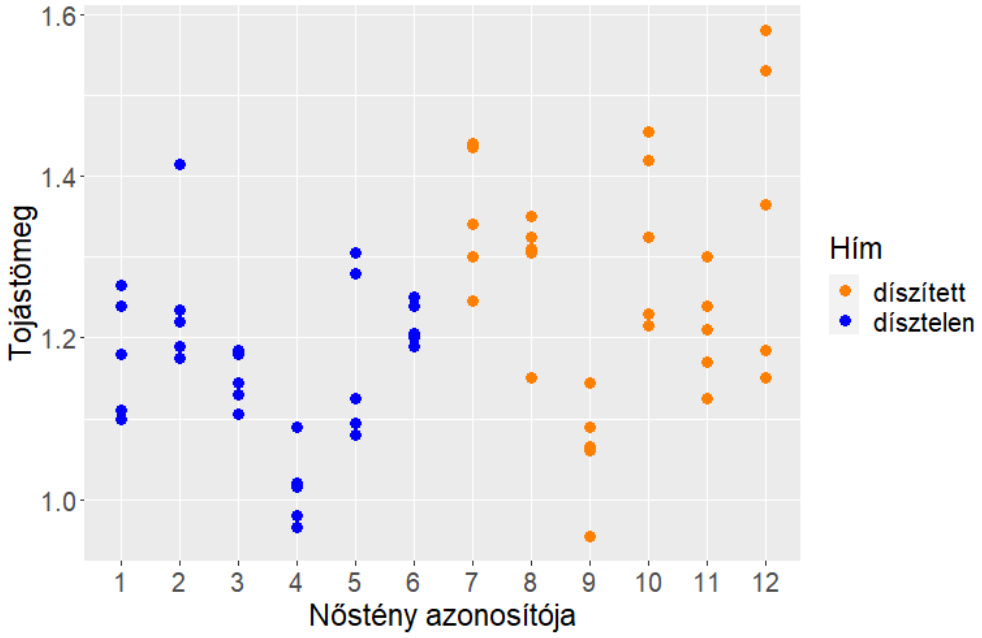
12. ábra. p-értékek eloszlása a transzformált adatok és a súlyozott regresszió esetén

Az első esetben a szignifikáns eredmények arányára 0,064 jött ki, a második esetben 0,0472, míg harmadszor 0,0516. A generált nyers adatokkal végzett számításokkor a vártnál többször kaptunk 0,05-nél kisebb p-értéket. Az elméleti értéktől való eltérés a második és harmadik számolásnál 0,0028, ill. 0,0016, amik sokkal kisebbek, mint az első esetben, ahol ez a szám 0,014. A várható értékre konstruált 95 %-os megbízhatósági szintű konfidencia intervallum: [0,042, 0,058], ami az első értéket nem, de a másodikat és a harmadikat már tartalmazza. A szimulációkhoz tartozó p-értékek egyenletes eloszlását vizsgáló tesztelés (Kolmogorov-Smirnov próba) során az első esetben $1,14 \cdot 10^{-4}$, a második esetben már 0,483-at, harmadszor pedig 0,999-et kaptam p-értéknek. Tehát mondhatjuk, hogy az alapadatok által szolgáltatott p-értékek nem egyenletes eloszlást követnek, míg a transzformált és súlyozott adatok esetében már nem tudjuk ezen nullhipotézist elvetni. Ez a példa alátámasztja azt az állítást, miszerint a heteroszkedaszticitás a valódiaknál alacsonyabb p-értékeket produkálhat, de az ilyen helyzetekre lehet alkalmazni különböző korrekciós módszereket, mint például a Box-Cox transzformáció vagy a súlyozott regresszió. A reziduálisok szórását elhanyagolható mértékben változtatták meg a különböző módszerek: az első esetben 3,3146, a második esetben 3,3148, a harmadik esetben pedig 3,3154 lett a reziduálisok átlagos szórása a szimulációk során. Így a becslések pontossága szinte változatlan maradt.

4.3. Függetlenség

Az adatpontok függetlenségének megsértése okozhatja a legnagyobb problémát a p-értékek szempontjából. Ha függetlenként kezelünk olyan megfigyeléseket, amelyek valójában nem azok, akkor túlságosan alacsony p-értékeket kaphatunk, és ezáltal hamis következtetéseket vonhatunk le. Amikor az adatpontok egymástól nem függetlenül keletkeznek, hanem azokat bizonyos szempontból csoportokba lehet osztani, akkor pszeudoreplikációról beszélünk. Ilyen jelenséget okozhat például a mérések területi, ill. időpontbeli közelsége, vagy éppen az is, ha több mérés ugyanazon egyedtől vagy közeli rokonságból származó egyedektől származik.

Elvégeztem egy vizsgálatot a [10] forrásbélihez hasonlóan, mely a következő kérdést veti fel: Függ-e a madarak tojásmérete a hím tollazatának díszítettségétől? A számításokhoz fiktív megfigyeléseket használtam fel, magába foglalva 60 tojásméretet 12 különböző tojótól (mindegyikhez 5-5 tojás tartozik) és 6-6 díszített ill. dísztelen hímtől. Az adatok a 13. ábrán láthatóak.



13. ábra. A fiktív kísérletben szereplő tojások tömegei

Ha a 30-30 tojást állítjuk szembe egymással, és a regressziószámításnál csupán a hím díszítettségét vesszük figyelembe a tojásméret előrejelzésénél, akkor ezen adatok 0,0022-es p-értéket szolgáltatnak, ami szignifikáns eredmény. Azonban hamis következtetés lenne azt mondani ezek után, hogy a két változó összefügg, hiszen nem vettük figyelembe, hogy több tojás is egy nősténytől származik. Mivel egy adott nőstényre lehet jellemző egy egyfajta tojásméret, ezért az előző számítás akkor lenne helyes, ha 12 helyett 60 különböző tojótól származnának a tojások.

A jelenlévő pszeudoreplikációt többféleképpen is ki tudjuk küszöbölni. Ismételjük meg a számításokat, de most vegyük az egyes nőstények tojásainak átlagát függő változónak. Tehát most nem 30-30, hanem 6-6 értéket vetünk össze. Ebben az esetben a p-érték 0,101 lett, ami már nem szignifikáns. Egy másik lehetséges módszer, ha vesszük az összes megfigyelést, de véletlen tengelymetszet modellt alkalmazunk. Ekkor a függő változót szintén az eredeti tojásméretek alkotják, a magyarázó változó a díszítettség megléte és véletlen hatás az egyes nőstények azonosítója, mivel mindegyik nőstényhez egy egyéni tengelymetszet tartozhat. Ezen eljárás mellett 0,076-ot kaptam p-értéknek, ami szintén nagyobb 0,05-nél.

Ez a példa jól mutatja, hogy mindig ügyelni kell a megfigyelések független mivoltára vagy helyesen kezelni a nem-függetlenség forrásait, hiszen alapvetően befolyásolja a regressziószámítás eredményét.

4.4. Multikollinearitás

Az alfejezetet a [14] forrás alapján dolgoztam ki.

Multikollinearitásról akkor beszélünk, ha a regressziószámításhoz használt magyarázó változók között vannak olyanok, melyek korrrelálnak egymással. Ha erős kapcsolat van két vagy több változó között, akkor az együtthatóbecslés nagyon érzékennyé válik a modellbéli apró változásokra. A becsült együtthatók pontosságának csökkenésével, gyengül a modell statisztikai ereje és a p-értékek is kevésbé lesznek megbízhatóak.

Az, hogy mekkora probléma a multikollinearitás, függ annak mértékétől és hogy pontosan milyen célból végezzük el a vizsgálatot. A multikollinearitás csak a korreláló változókat érinti, a többire vonatkozó számításokra nincs hatással. Az előrejelzések készítésében nem okoz gondot ezen előfeltétel megszegése, így ha nem lényeges az egyes magyarázó változók egyedi hatása, akkor nem szükséges vizsgálni, hogy jelen van-e az adatokban a multikollinearitás. Fontos megjegyezni azonban, hogy az együtthatókat és a p-értékeket alapvetően befolyásolja az egyes változók közötti korreláció mértéke.

Megismételtem a [14] forrásban szereplő vizsgálatokat, melyek a combcsont csontsűrűsége és néhány magyarázó változó közötti kapcsolat feltárására irányulnak. Az első modellillesztés során az imént említett függő változó becslésére a következő magyarázó változókat használtam: fizikai aktivitás, testzsír százalék, testtömeg és az utóbbi kettő szorzata. A regresszió p-értékeit és VIF értékeit a 2. táblázat tartalmazza. Ekkor csak a testzsír százalék hatása nem szignifikáns. Azonban a VIF értékek közül csak a fizikai aktivitáshoz tartozó kisebb 5-nél, tehát a többinél szükség van korrekcióra a p-értékek helyes értelmezéséhez (nem meglepő módon).

	Aktivitás	testzsír százalék	testtömeg	(testzsír százalék)·testtömeg
p-érték	0,003	0,176	$2,19 \cdot 10^{-6}$	0,005
VIF	1,05	14,93	33,95	75,06

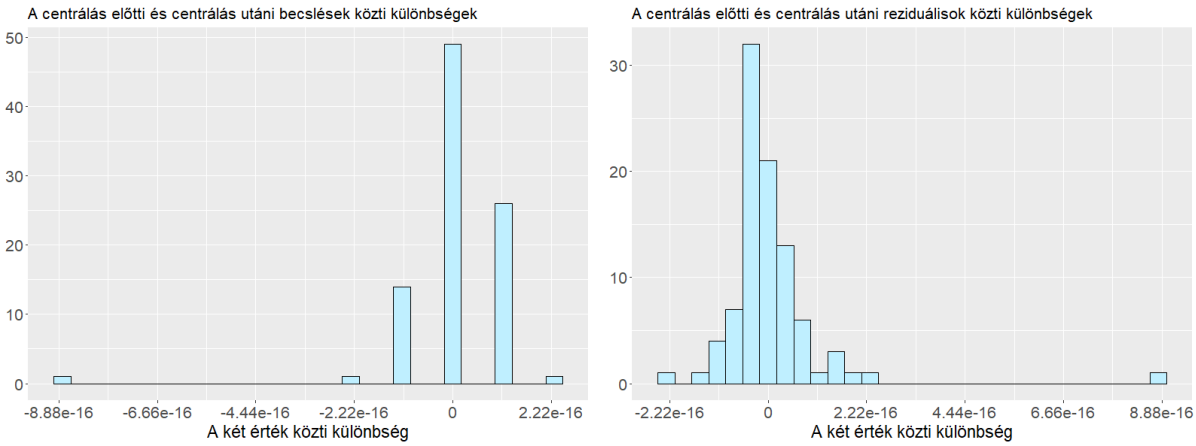
2. táblázat. Az első modellillesztés eredménye

A változók szorzatából adódó multikollinearitás megszüntetésének egy gyors és egyszerű módja az érintett változók centrálása (folytonos esetben). Visszatérve a példához, a centrálás utáni VIF értékek mind kisebbek 5-nél. Tehát sikerült kiküszöbölni a magyarázó változók közötti jelentős korrelációt. Ezúttal p-értékeknek pedig jóval 0,05 alatti számok jöttek ki. Most már biztonsággal lehet őket értelmezni, mindegyik elutasítja a nullhipotézist.

	Aktivitás	testzsír százalék	testtömeg	(testzsír százalék)·testtömeg
p-érték	0,003	0,003	$1,1 \cdot 10^{-11}$	0,005
VIF	1,05	3,32	4,75	1,99

3. táblázat. A centrálás utáni modellillesztés eredménye

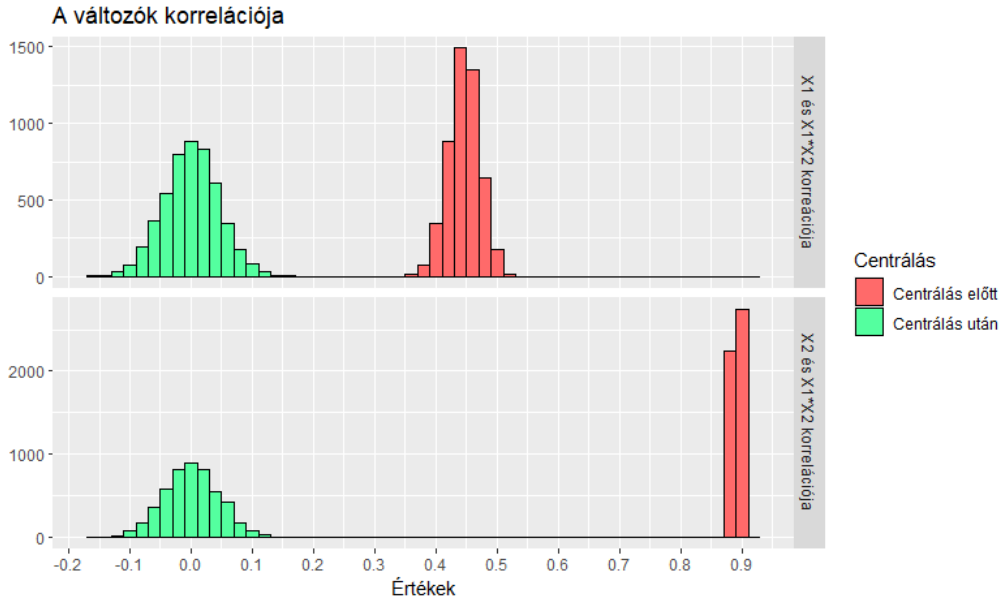
Összehasonlítva a két regressziószámítást a testzsír százalék p-értékeiben látható a legnagyobb különbség, jelentős mértékben módosítja azt az erős multikollinearitás. A fizikai aktivitást pedig egyáltalán nem befolyásolja (az együttható becslést sem), hiszen ez a magyarázó változó nem korrelál a többivel. A modellt átfogóan jellemző értékek ugyanazok maradtak mindkét számítás során: a reziduálisok szórása = 0,0705, az R^2 = 0,5623, a korrigált R^2 = 0,5422 és a függő változó valamint az összes magyarázó változó közötti függetlenség vizsgálatára vonatkozó F-statisztika, ill. p-érték 27,95, ill. $6,242 \cdot 10^{-15}$. A két regresszió együttható becsléseinek és reziduálisainak különbségét a 14. ábra hisztogramjai mutatják. Látható, hogy minden becslés- és reziduálispár közti különbség szinte 0. Ez is igazolja, hogy a multikollinearitás a regressziós előrejelzést nem, csak a változók egyedi hatását befolyásolja.



14. ábra. A centrálás előtti és utáni becslések és reziduálisok közti különbségek.

Centrálás generált adatokra

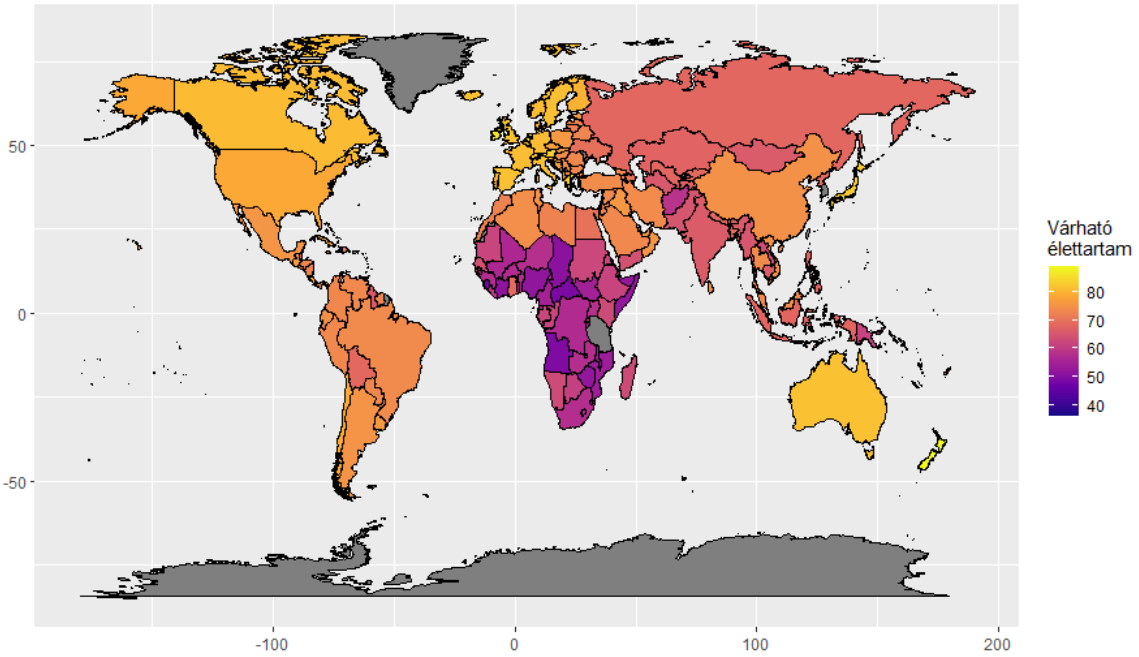
Megvizsgáltam a centrálás hatását a szerkezeti korrelációra. Ehhez két változót generáltam; X_1 normális eloszlású 10 várható értékkel és 1 szórással, X_2 pedig szintén normális eloszlású, de várható értéke 5, szórása pedig 1. Vettem a két változó korrelációját a szorzatukkal, centrálás előtt és centrálás után is. Összesen 5000-szer végeztem el a szimulációt. A korrelációs értékek hisztogramja a 15. ábrán látható. A centrálás egyértelműen csökkentette a változók és szorzatuk közötti korrelációt. Eredetileg a korrelációk átlaga 0,45 és 0,89 volt, centrálás utáni értékeik pedig jóformán nullák lettek (ezt vártuk): $-5,7 \cdot 10^{-4}$, ill. $2,8 \cdot 10^{-4}$. A korrelációk szórása ugyancsak megváltozott: 0,026-ról és 0,0068-ról szinte ugyanakkorára 0,044-re, ill. 0,045-re módosultak.



15. ábra. A generált változók korrelációja a szorzatukkal

5. A várható élettartam modellezése

Az R program segítségével végeztem néhány regressziószámítást, a várható élettartamra vonatkozóan. Az általam felhasznált adatbázis [2] [3] tartalmazza a 2010-es évhez tartozó, az egyes országokban élő emberek várható élettartamát, több magyarázó változóval egyetemben. A szimulációk során megvizsgálom a lineáris regresszió feltételeinek teljesülését, módosítom őket, ha szükséges és bemutatom, hogy hogyan képesek egyes tényezők, mint a regresszióban szereplő változók jellemzője vagy a mintaelemszám, befolyásolni a p-értékeket.

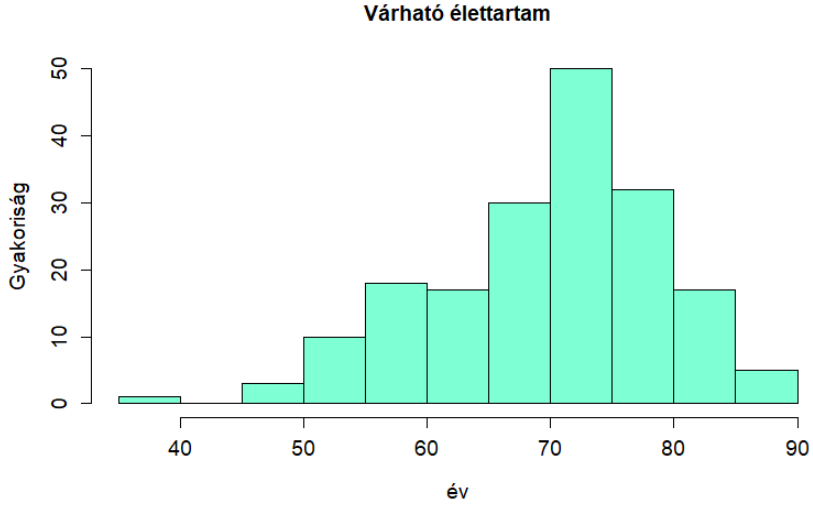


16. ábra. Az egyes országokban élő emberek várható élettartama

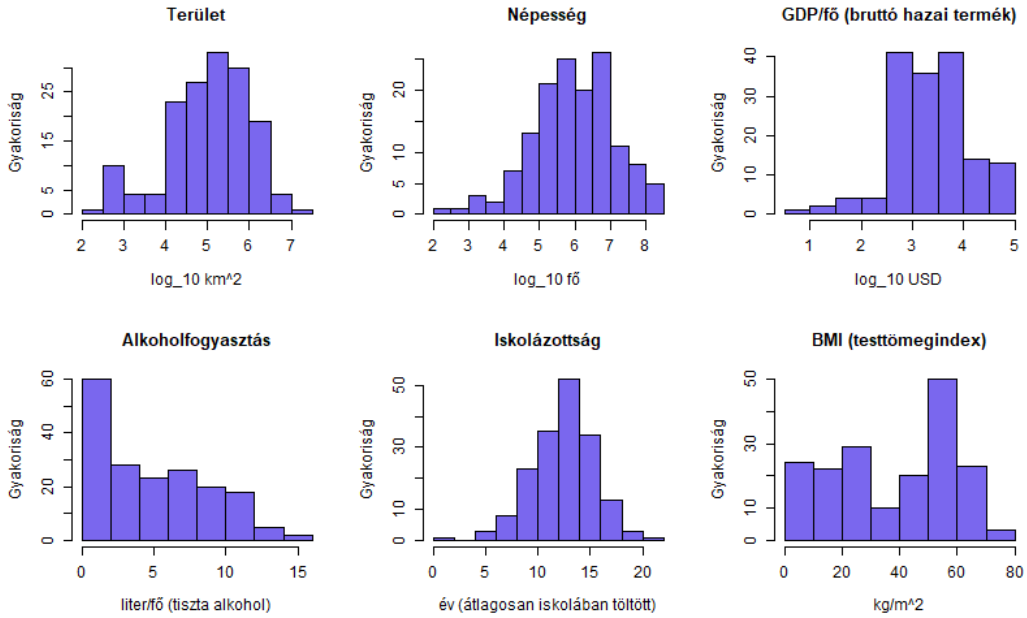
A regressziószámítások során a várható élettartam töltötte be a függő változó szerepét (17. ábra), mellette pedig összesen hat magyarázó változót vettem figyelembe, melyek a következők: az egyes országok területe, népessége, egy főre jutó GDP-je, alkoholfogyasztása, átlagos iskolázottsága és átlagos testtömegindexe. Az első három magyarázó változóhoz tartozó adathalmazt $\log_{10}$ skálára transzformáltam, mivel eredetileg ezek az adathalmazok nagyon nagy kiugró értékeket tartalmaztak, az értékek többsége pedig hozzájuk képest kicsik voltak. A transzformáció után ezek az adathalmazok már nem térnek el túl nagy mértékben a normális eloszlástól. A magyarázó változók hisztogramjai a 18. ábrán láthatóak.

A várható élettartam, mint függő változó és a magyarázó változók közötti kapcsolatot lineáris regresszióbéli hipotézisvizsgálattal elemeztem, vagyis t-próbával teszteltem az egyes együtthatókat. Ekkor a nullhipotézis minden egyes magyarázó változóra az, hogy a hozzá tartozó együttható (b_i) egyenlő nullával, tehát a várható élettartamot nem befolyásolja az ország adott jellemzője.

[2] https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who
[3] https://data.worldbank.org/indicator/AG.LND.TOTL.K2?view=chart



17. ábra. A várható élettartamok hisztogramja



18. ábra. A magyarázó változók hisztogramjai

Háromszor három futtatást végeztem: első körben a felsorolásban szereplő első három, másodszor az első négy és végül az összes magyarázó változó felhasználásával, mindet az adatok harmadára, kétharmadára és összességére. Az adatok harmadának és kétharmadának vizsgálatára egy bootstrap technikát alkalmaztam: a megfigyeléseket véletlenszerűen választottam ki és minden szimulációt 100-100 alkalommal végeztem el. Leteszteltem a lineáris regresszió feltételeinek teljesülését azon három regresszióra, amihez az összes adatot felhasználtam: a reziduálisok homoszkedasztikusságának és normalitásának vizsgálatára Breusch-Pagan és Shapiro-Wilk tesztet alkalmaztam, a magyarázó változók multikollinearitásának feltérképezéséhez pedig a VIF értékeket néztem meg. Felteszem a reziduálisok függetlenségét, mivel soros hatások keresésének nincs relevanciája (az országok ABC sorrendben vannak), a csoportba rendezhetőség ellenőrzése pedig túlmutat ezen a szakdol-

gozaton. Az eredményeket a 4.-6. táblázatok foglalják magukba, a reziduálisok grafikái pedig a 19.-20. ábrákon láthatóak. .

Feltétel	**Vizsgálati módszer**	**Eredmény**
Homoszkedaszticitás	Breusch-Pagan teszt	p-érték: 0.087
Normalitás	Shapiro-Wilk teszt	p-érték: 0.004
Multikollinearitás	VIF értékek	Terület: 1,39; Népesség: 1,39; GDP/fő: 1,00

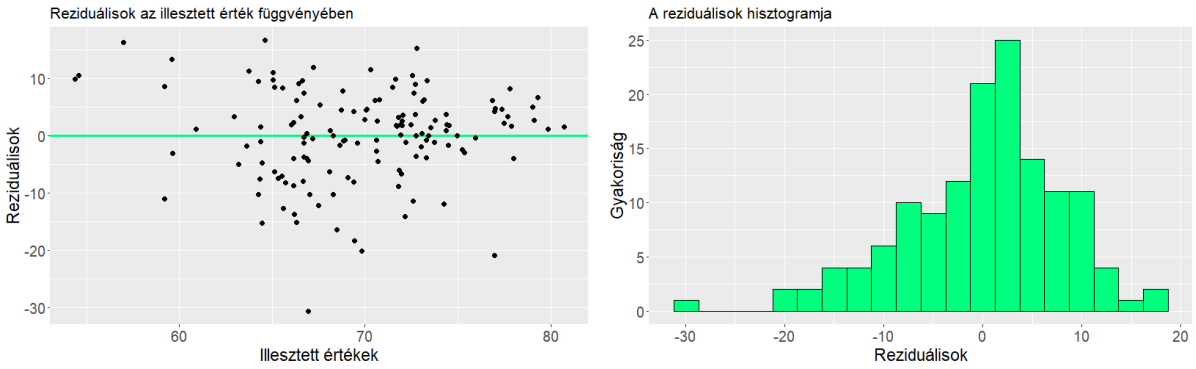
4. táblázat. Az első három magyarázó változót felhasználó regresszió tesztjei.

Feltétel	**Vizsgálati módszer**	**Eredmény**
Homoszkedaszticitás	Breusch-Pagan teszt	p-érték: 0,142
Normalitás	Shapiro-Wilk teszt	p-érték: $1,70 \cdot 10^{-5}$
Multikollinearitás	VIF értékek	Terület: 1,39; Népesség: 1,39; GDP/fő: 1,32; Alkoholfogyasztás: 1,31

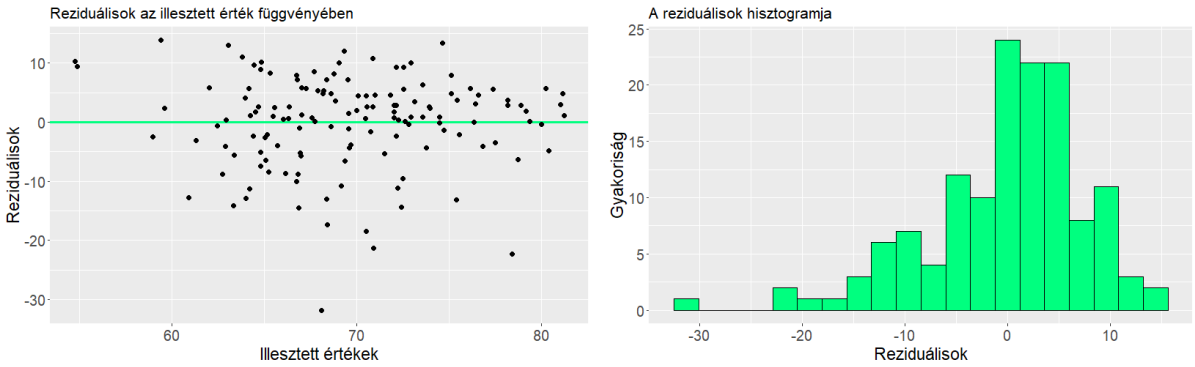
5. táblázat. Az első négy magyarázó változót felhasználó regresszió tesztjei.

Feltétel	**Vizsgálati módszer**	**Eredmény**
Homoszkedaszticitás	Breusch-Pagan teszt	p-érték: 0.056
Normalitás	Shapiro-Wilk teszt	p-érték: 0.004
Multikollinearitás	VIF értékek	Terület: 1,41; Népesség: 1,38; GDP/fő: 1,70; Alkoholfogyasztás: 1,85 Iskolázottság: 2,58; BMI: 1,61

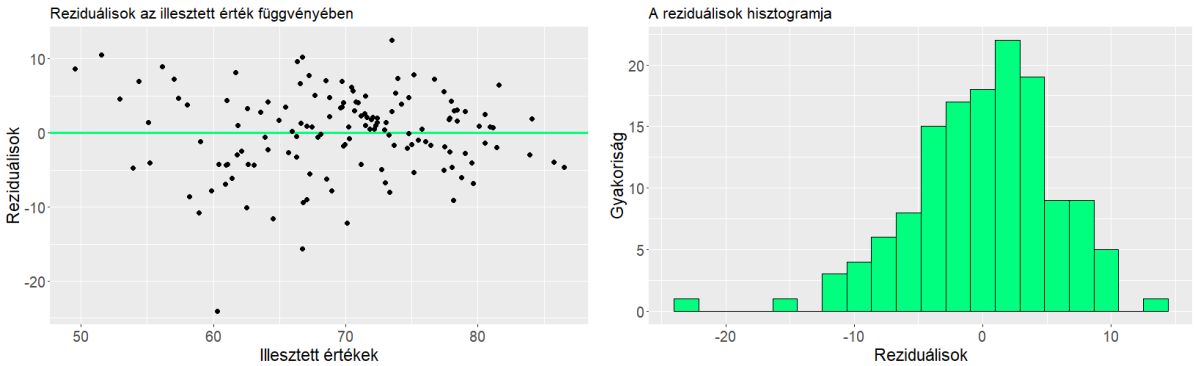
6. táblázat. Az összes magyarázó változót felhasználó regresszió tesztjei.



19. ábra. Az első három magyarázó változót felhasználó regresszió reziduálisai



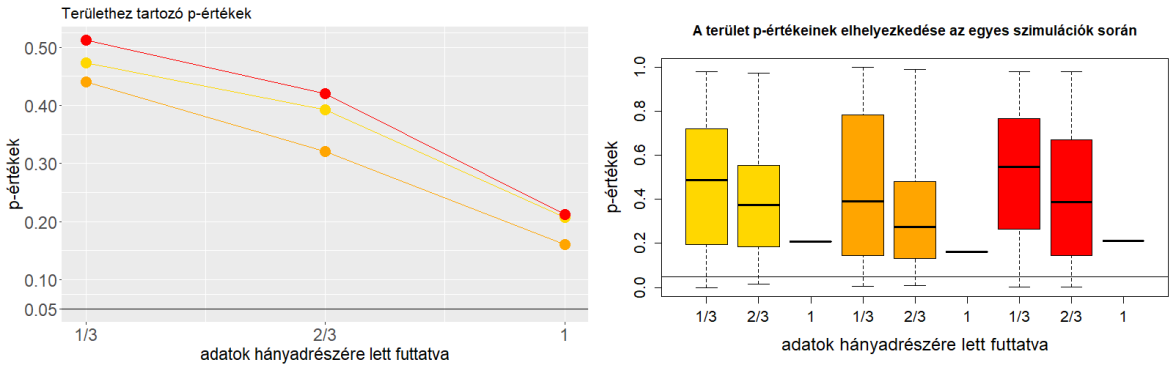
20. ábra. Az első négy magyarázó változót felhasználó regresszió reziduálisai



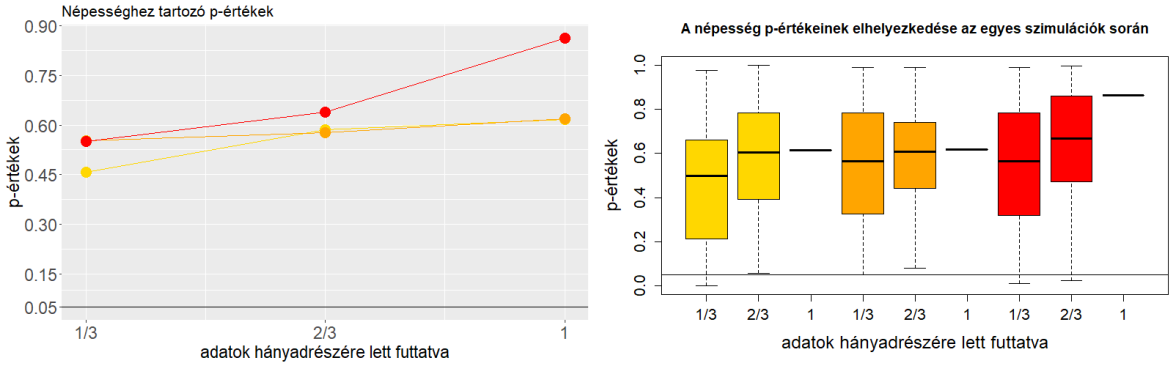
21. ábra. Az összes magyarázó változót felhasználó regresszió reziduálisai

A homoszkedaszticitásra vonatkozó próbák p-értékei mind nagyobbak 0,05-nél, azonban (főleg az összes magyarázó változót felhasználó regressziónál) nem sokkal haladják meg ezt a határt, más terjedelem mellett el is lehetne vetni a nullhipotézist. Ez az eredmény nem utal súlyos heteroszkedaszticitásra, a kis mértékű pedig, ahogy a 4.2 fejezetben említettem, nem okoz problémát. A reziduálisok hisztogramjait tekintve mondhatjuk, hogy a reziduálisok eloszlása hasonlít a normális eloszláshoz, habár a Shapiro-Wilk próba mindhárom esetben a nemnormalitást támasztja alá. Tudjuk, hogy a Gauss hibát feltételező modellek erős ellenállóképességgel rendelkeznek a feltétel megsértésével szemben (4.1 fejezet), a jelen eset pedig nem tűnik súlyos szabályszegésnek, így nem tartom fontosnak a kezelését. Végül a multikollinearitás vizsgálatánál pedig egyértelműen kisebb minden regresszió minden magyarázó változójához tartozó VIF érték 5-nél, így biztosan nem áll fenn kritikus multikollinearitás.

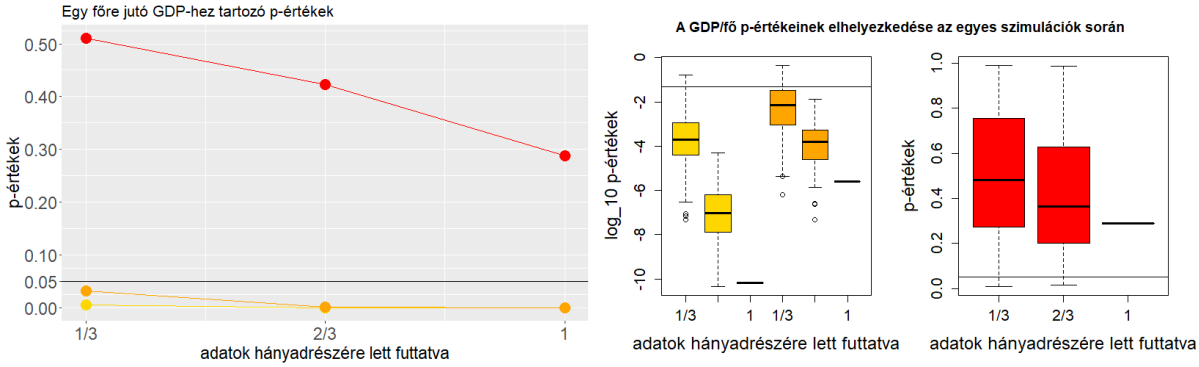
Most már lehet vizsgálni a függő és az egyes magyarázó változók közti kapcsolatra vonatkozó p-értékeket, melyek a 22.-27. ábrákon figyelhetők meg. Bal oldalon a szimulációk p-értékeinek átlaga a mintaelemszám függvényében, jobb oldalon pedig az összes szimuláció során kijött p-értékek boxplot ábrája áll. Arany színnel a három magyarázó változós, narancssárgával a négy magyarázó változós, piros színnel pedig a hat magyarázó változós futtatás eredményei láthatóak, valamint egy fekete egyenes jelzi a 0,05-ös szignifikancia szintet.



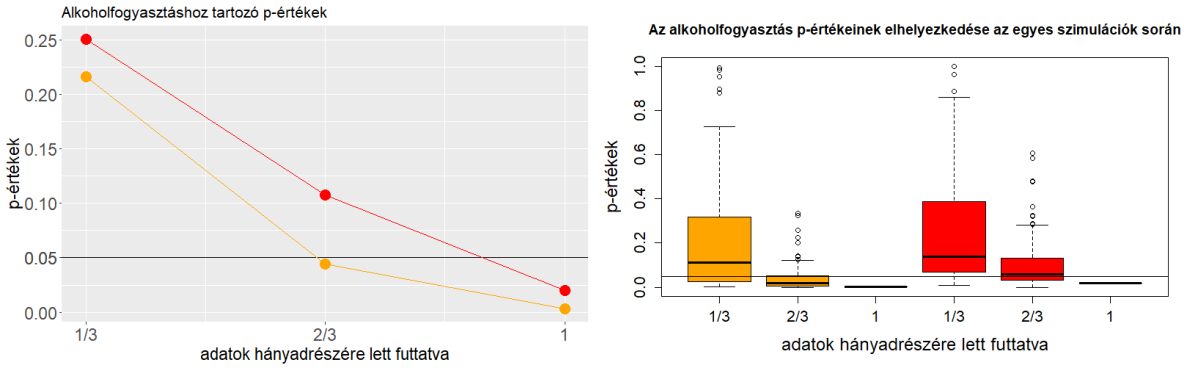
22. ábra. A szimulációk területhez tartozó p-értékeinek átlaga és boxplot ábrája



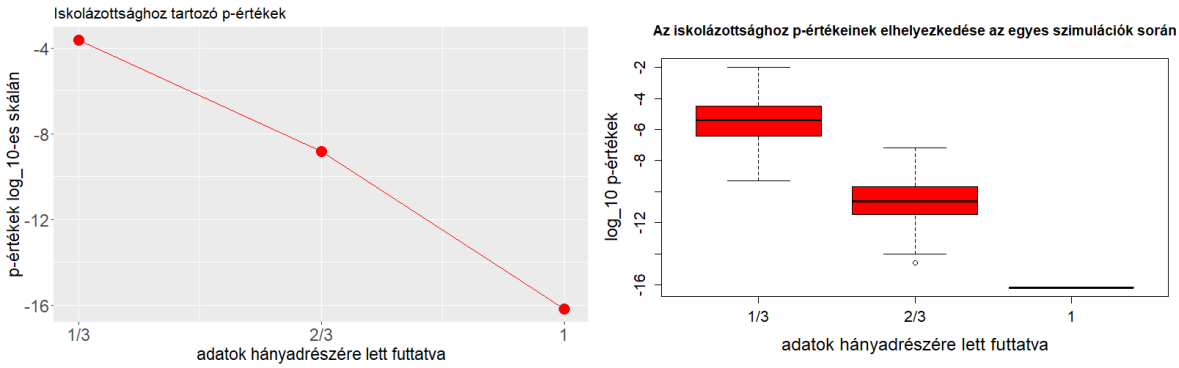
23. ábra. A szimulációk népességhez tartozó p-értékeinek átlaga és boxplot ábrája



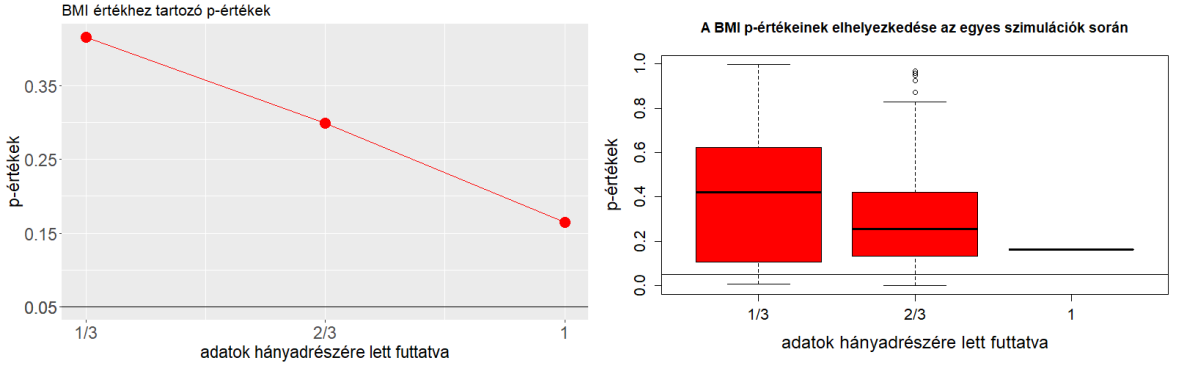
24. ábra. A szimulációk GDP/fő-höz tartozó p-értékeinek átlaga és boxplot ábrája



25. ábra. A szimulációk alkoholfogyasztáshoz tartozó p-értékeinek átlaga és boxplot ábrája



26. ábra. A szimulációk iskolázottsághoz tartozó p-értékeinek átlaga és boxplot ábrája



27. ábra. A szimulációk BMI-hez tartozó p-értékeinek átlaga és boxplot ábrája

A grafikonokon jól látható, hogy az eredményt nagyban befolyásolhatja a minta elemszáma és a regresszióban szereplő változók mivolta is. Általánosságban elmondható, hogy több adat felhasználásával az interkvartilis terjedelmek (a sorbarendezett adatok középső 50 %-a, az ábrákon a színes téglalapok (dobozok) jelzik) csökkennek, tehát a mintaelemszám növelésével egyre kisebb intervallumra koncentrálódnak a kapott p-értékek.

A területhez tartozó p-értékek átlaga minden esetben bőven 0,05 felett maradtak, sőt az interkvartilis terjedelmek is, csak néhány esetben jöttek ki 0,05 alatti számok (a p-értékek skálája majdnem lefedi a teljes [0,1] intervallumot, ha nem vesszük figyelembe az összes megfigyelést). Habár megfigyelhető egy csökkenő tendencia a mintaelemszám növelése mellett, mégsem tudunk összefüggést kimutatni egy ország területe és az ott élő emberek várható élettartama között.

A népességhez tartozó p-értékek még nagyobbak, mint a terület esetében, bár itt is szinte az egész [0,1] intervallumból jöttek ki értékek, amikor nem vettük be a regresszióba az összes megfigyelést. Mind az átlagok, mind az interkvartilis terjedelmek magasan a 0,05-ös szint felett helyezkednek el, sőt a mintaelemszám növelésével egyre nagyobb értékeket kapunk (a hat változó közül ez csak itt figyelhető meg). Tehát mondhatjuk, hogy egy ország lakosainak száma és a várható élettartamuk között nincs számottevő lineáris kapcsolat.

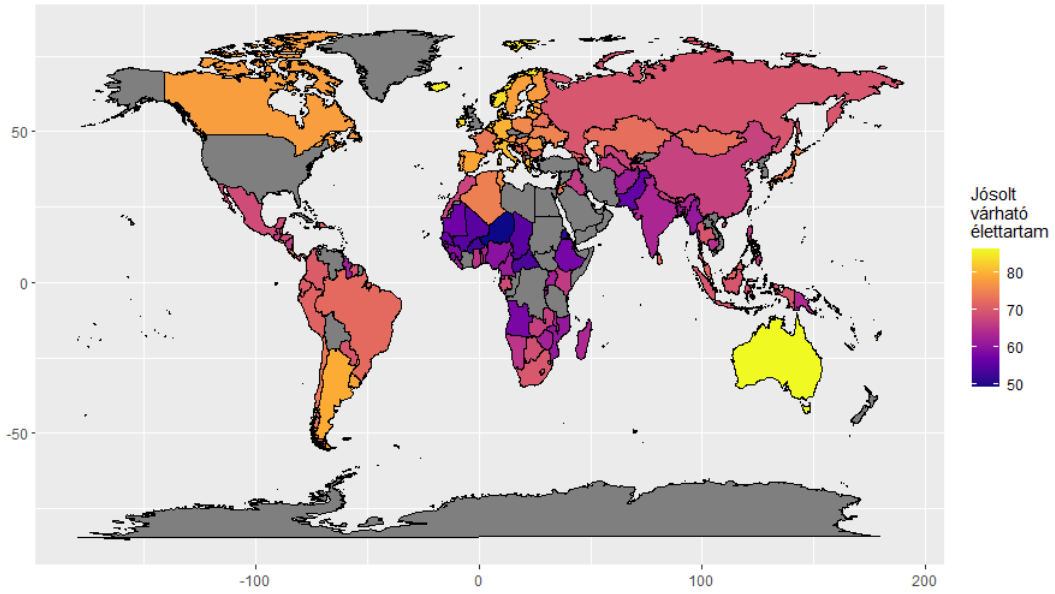
A GDP esete az előző kettőtől nagyban különbözik. A három, ill. négy magyarázó változós futtatás során a p-értékek átlaga és az interkvartilis terjedelmek 0,05 alatt maradtak, csak néhány esetben jöttek ki annál magasabb értékek. Azonban amikor már mind a hat magyarázó változó szerepelt a regresszióban, a p-értékek megugrottak és egy-két esettől eltekintve jóval a 0,05-ös szint fölé kerültek. Az egyes szimulációk során kapott p-értékek

skálája is kiszélesedett az előző változókhoz hasonlóan, míg amikor még a nullhipotézist elutasító eredményt írtak le, ez az intervallum is jóval kisebb volt. Ez egy remek példája annak, hogy a regresszióban szereplő egyes változók hatásai képesek elnyomni egymást, a jelenlegi vizsgálatban látható, hogy az iskolázottság mutat egyértelmű lineáris kapcsolatot a várható élettartammal, amely kiszoríthatta a GDP hatását. Mindemellett a változók számától függetlenül az egy főre jutó GDP p-értékei is csökkennek a mintaelemszám növelésével.

Az alkoholfogyasztásnál szintén megfigyelhetőek lineáris kapcsolatot támogató és ellenező p-értékek is. Ebben az esetben nem maguk a változók, hanem a vizsgált minta elemszáma idézte elő a jelentős változást. A p-értékek egyre alacsonyabban helyezkednek el, minél több megfigyelést veszünk be a regresszióba. A négy magyarázó változós futtatások során már az adatok kétharmadának figyelembe vételével alacsonyabb lett a p-értékek átlaga, mint 0,05, az összes megfigyelés beszámításával pedig már az összes változós regresszió p-értéke is kisebb a szignifikancia szintnél. Pár kiugró esetet leszámítva nem csak az értékük, hanem a terjedelmük is jóval lecsökkent.

Az iskolázottsághoz nagyon alacsony, a mintaelemszám növelésével egyre kisebb, nullához közeli p-értékek tartoznak. A kapott értékek teljes terjedelme a 0,05-ös szint alatt helyezkedik el, és a többi változóhoz képest nagyon szűk is. Ebben az esetben elutasítható a nullhipotézis, miszerint az egy országra jellemző iskolázottság és a várható élettartam között nincs lineáris kapcsolat.

Végül jönnek a testtömegindex p-értékei. Ezen változó esete szinte teljesen megegyezik a terület esetével. Hasonló nagyságokban helyezkednek el a p-értékek, hasonló terjedelemben és ugyancsak csökkenő tendenciát mutatva. Jelen esetben semmi sem cáfolja azt az állítást, hogy a várható élettartam nem mutat lineáris kapcsolatot a BMI értékkel.



28. ábra. A várható élettartam jósolt értékei

A 28. ábrán látható a legutolsó modell (6 magyarázó változó, összes adat) által jósolt várható élettartamok (a szürkével jelzett országokra valamely hiányzó adatuk miatt nem készült becslés). A valós és a jósolt értékek különbségének átlaga (ahogy vártuk) 0, míg a szórásuk 5,6 év lett. A legnagyobb eltérés kicsit több mint 24 év volt: Haiti esetében, ahol a modell által előrejelzett érték 60,33 év lett, míg az adatbázis szerint ez csupán 36,3 év.

Valójában a modell nem ad rossz előrejelzést, mivel 2009-ben és 2011-ben is 60 év körül volt a várható élettartam, azonban 2010-ben egy nagy erejű földrengés rengeteg áldozatot követelt, ezzel jelentős mértékben lecsökkentve az átlagos halálozási életkort. A modellbéli magyarázó változók együtthatóit a 7. táblázat tartalmazza, ahol a tengelymetszet 39,34 év lett. A korrigált R^2 értékére pedig 0,64 jött ki, ami nem túl magas, de persze még nagyon sok változót figyelembe lehetett volna venni, mint például különböző betegségeket.

	log_{10}(Terület)	log_{10}(Népesség)	log_{10}(GDP)
Együtthatók:	-0,784	0,084	0,916
	Alkoholfogyasztás	Iskolázottság	BMI
Együtthatók:	-0,398	2,529	0,042

7. táblázat. Az utolsó regressziós modell együtthatói.

6. Összefoglalás

Szakdolgozatom megírása során a p-érték játszott központi szerepet. A szakirodalom feldolgozása betekintést engedett abba a kérdéskörbe, hogy a modern statisztika milyen problémákkal küzd a hipotézisvizsgálat témakörében. A kutatások és tanulmányok sokszor nem elég óvatosak vagy éppen szándékosan manipuláltak egy kedvezőbb eredmény elérése érdekében. Ezen ok miatt egyre több statisztikus és statisztikai területen dolgozó szólal fel a változtatás mellett, többek között nagyobb precizitásra, több nyíltságra sarkallva azokat, akik ilyen jellegű kutatásokat végeznek.

Az elméletet a gyakorlatba az R programban végzett szimulációkon keresztül ültettem át. A cikkekben szereplő és általam generált adatokon keresztül vizsgáltam, hogy a lineáris regresszió előfeltételeinek nem teljesülése mennyire képes befolyásolni a regressziós elemzések eredményeit, elsősorban a p-értékeket.

A reziduálisok normalitásának vizsgálatára egymástól valóban független adatokat generáltam, ahol a feltétel hol kisebb hol nagyobb mértékben sérült. A szimulációk azt mutatták meg, hogy a Gauss-féle hibákat feltételező modellek robusztusak a feltétel nem teljesülésével szemben, csak kis mintaelemszám esetén okoznak gondot, elsősorban a nagy és több kiugró értékkel rendelkező adathalmazok.

A reziduálisok homoszkedaszticitását egy cikkbéli adathalmazon vizsgáltam először, a probléma kezelésére pedig három módszert próbáltam ki: a változók újradefiniálását, súlyozott regressziót és Box-Cox transzformációt. A heteroszkedaszticitás mértéke csökkent picit, de nem sikerült megszüntetni egyik eljárással sem, azonban mivel volt egyértelmű lineáris kapcsolat a függő és a magyarázó változó között, így a p-értékek szempontjából ez nem volt probléma. Ezt követően általam generált, független adatokon figyeltem meg a súlyozott regresszió és a Box-Cox transzformáció hatását. Mivel valóban független adatokat vizsgáltam, várhatóan a p-értékeknek egyenletes eloszlást kellett volna követniük, ám a heteroszkedaszticitás miatt kisebb p-értékek jöttek ki, így kissé ferdült az eloszlás. A súlyozott regresszió (az elméletileg legjobb súlyokat használva) és a Box-Cox transzformáció azonban jól teljesítettek, mindkét esetben kiegyenlítődtek a p-értékek.

A reziduálisok függetlenségére egy cikkhez hasonló vizsgálatot végeztem. Ebben a kísérletben a feltétel sérülését az okozta, hogy a megfigyeléseket csoportokba lehetett rendezni. Ha nem vettem figyelembe a klaszterezhetőséget akkor nagyon kicsi, a nullhipotézist elutasító p-értéket kaptam. Ha a csoportok átlagára végeztem el a regresszió számítást akkor pedig már jóval nagyobb p-érték jött ki. Végül alkalmaztam a véletlen tengelymetszet modellt, amely szintén hatásosnak bizonyult, ebben az esetben sem lehetett elutasítani a nullhipotézist. A példában a reziduálisok közötti kapcsolat jelentősen befolyásolta a regressziószámítások eredményét.

Az utolsó feltétel a magyarázó változók multikollinearitása, amit egy forrásbéli adathalmazon vizsgáltam meg. A regresszióban szerepelt két magyarázó változó szorzata is, mely nem meglepő módon erősen korrelált a szorzat tagjaival. A multikollinearitás kimutatására a varianciainflációs tényezőt használtam, a kezelésére pedig a változók centrálását. Volt olyan változó melyhez a centrálás előtt még 0,176-os, centrálás után pedig már 0,003-as p-érték tartozott. A centrálás hatását a változók és szorzatuk korrelációjára külön is megvizsgáltam, generált adatokon. Eredményképp tényleg jóval lecsökkentek, a korrelációk átlaga egyik esetben 0,45-ről $-5,7 \cdot 10^{-4}$-re, másik esetben pedig 0,89-ről $2,8 \cdot 10^{-4}$-re csökkent le, vagyis szinte nullára. A változók korrelációjának szórása centrálás előtt 0,026, ill. 0,0068 volt, míg centrálás után már hasonló értékeket vettek fel: 0,044-re, ill. 0,045-re

változtak.

Végül a tapasztalataimat valós adatokon, a várható élettartam modellezésénél is alkalmaztam, ahol szintén sikerült további érdekességeket felfedezni. A regresszióban szereplő magyarázó változók tulajdonsága és a mintaelemszám nagysága is képes volt egy-egy magyarázó változóhoz tartozó p-értéket a 0,05-ös határ egyik, ill. másik oldalára tolni. Az egy főre jutó GDP esetében egyértelműen látszott lineáris kapcsolat a várható élettartammal, amíg mellette a terület, a népesség és az alkoholfogyasztás szerepelt a regresszióban. Azonban mikor már az iskolázottságot (és a BMI értéket) is bevettem a modellbe a p-értékek megugrottak és már nem lehetett elutasítani a nullhipotézist. Az alkoholfogyasztás p-értékeit pedig a mintaelemszám befolyásolta jelentősen. Amikor az adatok egyharmadára végeztem el a számításokat, akkor még 0,2-0,25 körüli p-értékeket kaptam, míg az összes megfigyelés bevételével már kisebbek lettek, mint 0,025. Továbbá a vizsgált magyarázó változók közül voltak olyanok (mint a népesség), melyek nem mutattak lineáris kapcsolatot a várható élettartammal, de például az átlagos iskolázottságról már állíthatjuk, hogy jelentős a kapcsolata a várható élettartammal.

Természetesen a téma nem lezárt, számtalan kísérletet lehet még végezni eddig nem vizsgált eltérések súlyosságának feltárására és még több eljárás hatékonyságának összehasonlítására.

Úgy érzem a dolgozat megírása során hasznos ismeretekkel bővült az egyetemen megszerzett tudásom, mint például hogyan lehet kellőképpen körültekintően megvizsgálni különböző változók kapcsolatát, hogy a megfelelően alátámasztott eredményeket bárki biztonsággal tudja értelmezni, melyet szeretnék későbbi tanulmányaim alatt, ill. a való életben is kamatoztatni.

Hivatkozások

[1] Amrhein, V., Greenland, S. & McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, 567, 305-307.

[2] Kuffner, T. A., Walker, S. G. (2019). Why are p-Values Controversial? *The American Statistician*, 73:1, 1-3.

[3] Wasserstein, R. L., Lazar, N. A. (2016). The ASA Statements on p-Values: Context, Process and Purpose. *The American Statistician*, 70:2, 129-133.

[4] Hurlbert, S. H, Levine, R. A. & Utts J. (2019). Coup de Grâce for a Tough Old Bull: "Statistically Significant" Expires. *The American Statistician*, 73:sup1, 352-357.

[5] Wasserstein, R. L., Schirm A. L. & Lazar N. A. (2019). Moving to a World Beyond "p $<$ 0.05". *The American Statistician*, 73:sup1, 1-19.

[6] Greenland, S. (2019). Valid P-values Behave Exactly as They Should: Some Misleading Criticisms of P-values and Their Resolution with S-values. *The American Statistician*, 73:1, 106-114.

[7] Anderson, A. (2019). Assessing Statistical Results: Magnitude, Precision and Model Uncertainity. *The American Statistician*, 73:sup1, 118-121.

[8] Knief, U., Forstmeier, W. (2021). Violating the normality assumption may be the lesser of two evils. *Behav Res*, 53, 2576–2590.

[9] Frost, J. (2023.03.25.). Heteroscedasticity in Regression Analysis. *Statistics By Jim.* https://statisticsbyjim.com/regression/heteroscedasticity-regression/

[10] Forstmeier, W., Wagenmakers, E.-J. & Parker T. H. (2017). Detecting and avoiding likely false-positive findings–a practical guide. *Biol. Rev. Cambridge Philos. Soc.*, 92, 1941–1968.

[11] Rossiter, D. G. (2019). Box-Cox transformation. *Cornell CALS.* https://www.css.cornell.edu/faculty/dgr2/_static/files/R_html/Transformations.html

[12] Vaghefi, R. M. (2023.05.02.). Weighted Linear Regression. *Towards Data Science.* https://towardsdatascience.com/weighted-linear-regression-2ef23b12a6d7

[13] Pillinger, R. (2023.04.22.). Random intercept models. *University of Bristol.* http://www.bristol.ac.uk/cmm/learning/videos/random-intercepts.html

[14] Frost, J. (2023.04.22.). Multicollinerity in Regression Analysis: Problems, Detection, and Solutions. *Statistics By Jim.* https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/

[15] Az Investopedia csapata (2023.04.22.). Variance Inflation Factor (VIF). *Investopedia.* https://www.investopedia.com/terms/v/variance-inflation-factor.asp

[16] Iacobucci, D., Schneider, M.J., Popovich, D.L. & Bakamitsos, G. A. (2016). Mean centering helps alleviate "micro" but not "macro" multicollinearity. *Behav res*, 48, 1308–1317.

[17] Bohrnstedt, G. W., & Goldberger, A. S. (1969). On the exact covariance of products of random variables. *Journal of the American Statistical Association*, 64(328), 1439-1442.

[18] Oleszak, M. (2023.05.03.). Regularization in R Tutorial: Ridge, Lasso and Elastic Net. *DataCamp*. https://www.datacamp.com/tutorial/tutorial-ridge-lasso-elastic-net

[19] van Wieringen, W. N. (2023.05.16.). Lecture notes on ridge regression, Version 0.40 (2021). https://arxiv.org/pdf/1509.09169.pdf

[20] Kerns, G. J. (2010). Introduction to Probability and Statistics Using R. https://www.atmos.albany.edu/facstaff/timm/ATM315spring14/R/IPSUR.pdf