# Kívülálló értékek hatása a lineáris modellben

Szakdolgozat

Írta: Huszárik Anna Flóra

Elemző matematikus BSc

Témavezető:

Németh László, PhD hallgató
Valószínűségelméleti és Statisztika Tanszék
Eötvös Loránd Tudományegyetem, Természettudományi Kar


Eötvös Loránd Tudományegyetem
Természettudományi Kar
Budapest, 2018

# Köszönetnyilvánítás

Elsősorban szeretnék köszönetet mondani témavezetőmnek, Németh Lászlónak, akinek témaötlete és segítőkészsége nélkül a dolgozat nem jöhetett volna létre. Hálával tartozom a rengeteg konzultációért és mérhetetlen türelméért is.

Továbbá köszönöm a családomnak és a barátaimnak, hogy az egyetemi évek alatt végig biztattak, motiváltak és nyugodt környezetet biztosítottak.

# Tartalomjegyzék

# 1. Bevezetés

A statisztikai modellek rendkívül hasznos eszközök egy adatkészlet jellemzőinek megismeréséhez és megértéséhez. Gyakran előfordul azonban, hogy ezeket a modelleket jelentősen befolyásolja egy, vagy néhány szélsőséges mintaelem. Az ilyen mintaelemek általában olyan pontok, amelyek szokatlan értékeket vesznek fel a többi pont értékeihez képest és emiatt kívül állnak az adathalmaz nagy részétől.

Ezeknek az elemeknek az azonosítása az élet minden területén fontos lehet. Az egészségügyben például ezen elemek segítségével gyorsabban lehet a daganatos sejteket detektálni és megkönnyíti a kezelés módszerének pontos meghatározását. Az informatika területén a szélsőséges értékekkel felderíthetőek az illetéktelen belépések a számítógép hálózatban, vagy a csalók jogtalan hitelkártya és mobiltelefon használata.

A kívülálló pontok azonosítása alkalmas az adathibák kiszűrésére is, így ezeknek a javítása vagy törlése után pontosabb modellt kapunk egy statisztikai elemzés során. A dolgozatban ilyen pontoknak a lineáris regresszión belül hatását vizsgálom, hogy egy vagy több kiugró érték milyen mértékben mozdítja el a modellt, és hogyan változtatja meg annak paramétereit.

A dolgozat első felében az egyszerű regressziós modellek elméleti hátterét ismertetem. A lineáris regresszió általános modellje után bemutatom a modell mérőszámait, a hipotézisvizsgálatát, majd az együtthatóinak közönséges és súlyozott legkisebb négyzetek becslését. Ezt követően többváltozós esetben foglalkozom a lineáris regresszióval.

A második fejezetben ismertetem a kiugró értékek fogalmát és bemutatom, hogy hatásuk szerint milyen két osztályba soroljuk őket. Kétváltozós esetben szemléltetem, hogy milyen befolyással lehetnek a becslés pontosságára a regressziós modellben. Ebből következtetve látszik, hogy nem minden megfigyelésnek van ugyanolyan jelentősége a legkisebb négyzetek regressziójában. Továbbá megmutatom a kiugró értékek kiszűrésére leggyakrabban használt eljárásokat, és két regressziós módszert, amelyek kevésbé érzékenyek ezekre az értékekre.

A dolgozatban a regressziós alapismeretek bemutatása után egy valós ingatlan értékesítési adathalmazon is szemléltetem a kívülálló értékek lehetséges detektálásait és

kezelését. Ehhez az R programnyelvet és az RStudio programot használom.

Zárásképp megállapításra kerül, hogy hogyan érdemes azonosítani és kezelni a kívülálló értékeket különböző esetekben.

# 2. Lineáris regresszió

## 2.1. Matematikai alapfogalmak

A regressziós módszerek bemutatása előtt ismertetek néhány valószínűségszámítási, illetve statisztikai alapdefiníciót és jelölést.

**2.1. Definíció.** A $\Omega$ halmazt nevezzük eseménytérnek, az $\omega \in \Omega$ elemeket pedig elemi eseménynek. Ekkor az $\eta(\omega) : \Omega \longrightarrow \mathbb{R}$ függvényt valószínűségi változónak nevezzük, ha a $P(a \leq \eta \leq b)$ valószínűség létezik bármely $a < b$, $a, b \in \mathbb{R}$ esetén.

**2.2. Definíció.** A minta valamely valószínűségi változóra vonatkozó véges számú független kísérlet vagy megfigyelés eredménye: véges sok azonos eloszlású valószínűségi változó.

A klasszikus és többváltozós lineáris regresszió bemutatásához felhasznált irodalmak: [1], [2], [10].

### Klasszikus lineáris regresszió

A klasszikus lineáris regresszió egy statisztikai módszer, amely segítségével tanulmányozhatunk és összegezhetünk két folytonos változó közötti lineáris kapcsolatot. A két változó elnevezései:

X: független változó, magyarázó változó

Y: függő változó, eredményváltozó, célváltozó

A lineáris kapcsolat a következőképpen fejezhető ki:

$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon,$$

ahol $\beta_1$ a magyarázó változóhoz tartozó együttható, $\beta_0$ a tengelymetszet, $\epsilon$ pedig a hibatag vagy maradékváltozó. A lineáris regresszió esetén feltételezzük, hogy a magyarázó változó van hatással a célváltozóra és nem pedig fordítva. A hibatagra feltesszük, hogy normális eloszlású és a várható értéke 0.

### Többváltozós lineáris regresszió

Többváltozós lineáris regresszióról akkor beszélünk, amikor 2 vagy több magyarázó változóval rendelkezünk és ezeknek a célváltozóra nézett együttes hatását vizsgáljuk. A lineáris modell $n$ darab független változó esetén:

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \ldots + \beta_n \cdot X_n + \varepsilon,$$

ahol $\beta_1, \beta_2, \ldots, \beta_n$ a magyarázó változókhoz tartozó együtthatók, $\beta_0$ a tengelymetszet, $\varepsilon$ pedig a hibatag vagy maradékváltozó. A hibatagra itt is feltesszük, hogy normális eloszlású és a várható értéke 0.

### Hipotézisvizsgálat

A hipotézis a sokaság eloszlására vagy valamely paraméterére vonatkozó állítás, amelynek az igazságát vizsgálni szeretnénk. A minta alapján döntést hozunk a hipotézisről. Nullhipotézisnek nevezzük azt a feltevést, amelynek a helyességét a statisztikai próba során feltételezzük:

$$H_0 : \mu = m_0$$

A nullhipotézis elfogadásáról statisztikai próba segítségével döntünk. A feltételezés igaz volta mellett felállítható a mintára, illetve a minta alapján számolt próbafüggvényre egy elfogadási tartomány, amin belüli értéket nagy valószínűséggel kaphatunk. Amennyiben ezen belül eső értéket tapasztalunk, elfogadjuk, ha pedig kívül eső értéket kapunk, akkor elutasítjuk a nullhipotézist. A nullhipotézis helyett ekkor egy úgynevezett ellenhipotézis létezését kell elfogadnunk. Ez általában :

$$\underline{H_1 : \mu \neq m_0}$$

Esetenként lehet egyoldali, vagy egy konkrétan megadott, $H_0$–tól különböző feltételezés. A legismertebb statisztikai próbák: u-, t- és F-próba

### T-próba

Adott egy normális eloszlású független minta, ahol a várható érték és a szórás ismeretlen paraméterek, de a várható értékről azt feltételezzük, hogy egy adott értékkel egyenlő. A próbastatisztikát, jelölje $T_0$, a mintából számoljuk és Student-féle t-eloszlást követ a nullhipotézis igaz volta mellett.

A kétmintás t-próba azt vizsgálja, hogy két normális eloszlású független mintában egy-egy valószínűségi változó átlagai egymástól szignifikánsan eltérnek-e.

### P-érték

A p-érték annak a valószínűsége, hogy igaz $H_0$ hipotézis mellett milyen valószínűséggel kaphatunk $|T_0|$–nál nagyobb statisztikai értéket. Nagy p-érték mellett a megfigyelésünk tipikus, kis p-érték mellett pedig szokatlan lenne. Hipotézisvizsgálati feladat esetén általában 0.05–nél kisebb p-érték esetén a nullhipotézist a minta alapján nem tartjuk igaznak.

### F-próba

Legyen adott két normális eloszlású független minta, ahol a várható értékek és a szórások is ismeretlen paraméterek. Az F-próba a szórásnégyzetek egyezését teszteli. Igaz $H_0$ mellett a próbastatisztika Fisher-féle kétparaméteres F eloszlású lesz.

## 2.2. Klasszikus lineáris regresszió

Az egyszerű lineáris regresszió két folytonos változó közötti korrelációt írja le. Ha $n$ darab megfigyelésünk van, akkor a következő egyenletek írják le a magyarázó változó

$(X)$ és az eredményváltozó $(Y)$ közötti lineáris kapcsolatot:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 \cdot x_1 + e_1 \\ y_2 &= \beta_0 + \beta_1 \cdot x_2 + e_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 \cdot x_n + e_n \end{aligned}$$

ahol $y_i$ az $Y$ változóból kapott mintaelemek, míg $x_i$ az $X$ magyarázó változóból kapott mintaelemek. A modell mátrixos alakja:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} e_0 \\ e_1 \\ \vdots \\ e_n \end{bmatrix}$$

azaz,

$$y = X\beta + e$$

### Reziduálisok

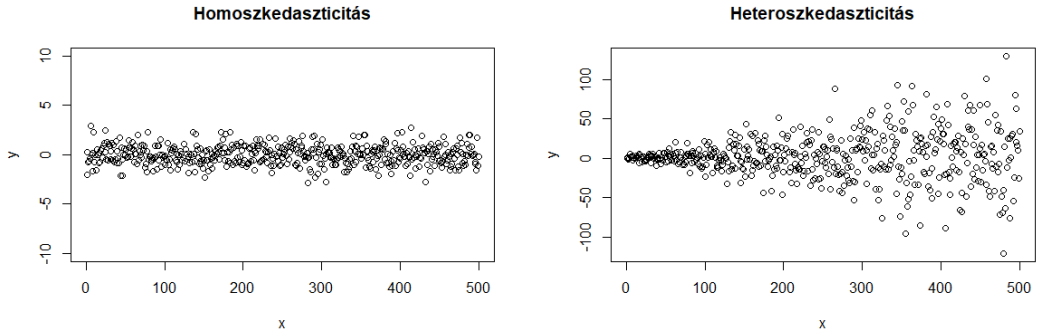
Az eredményváltozó $(y)$ és a becsült célváltozó $(\hat{y})$ különbségét reziduálisnak vagy a modell hibájának $(e)$ nevezzük [3]. A hibák normális eloszlást követnek.

$$e = y - \hat{y}$$

**Állítás.** A reziduálisok várható értéke 0.

*Bizonyítás.* $E[e] = E[y - \hat{y}] = E[\beta_0 + \beta_1 x - (\hat{\beta}_0 + \hat{\beta}_1 x)] = 0,$
mivel $E[\hat{\beta}_0] = \beta_0$ és $E[\hat{\beta}_1] = \beta_1$ □

A lineáris regresszióban feltételezzük a homoszkedaszticitást. Ez azt jelenti, hogy a reziduálisok varianciája konstans $(Var(\varepsilon) = \sigma^2)$ és annak nagysága nem függ a magyarázó változó nagyságrendjétől. Ellenkező esetben a heteroszkedaszticitás jelenik meg.




1. ábra. A hibatag lehetséges szórásai szemléletesen

## 2.3. Mérőszámok a lineáris regresszió pontosságára

A lineáris regresszió pontosságát jól jellemző mérőszám a determinációs együttható, ehhez bemutatom az őt felépítő mennyiségeket és a köztük lévő kapcsolatot. A determinációs együttható függ a becsült és a valós értékek különbségeitől, és a becsült és a valós értékeknek a mintaátlagtól való távolságától.

### Reziduális négyzetösszeg

A reziduális négyzetösszeg (Sum of Squares for Error):

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

$SSE = 0$ pontosan akkor, ha a becsült érték és a valós minden pontban megegyezik, azaz tökéletes az illeszkedés. Más esetben pedig a négyzetre emelés miatt $SSE > 0$.

### Totális négyzetösszeg

A totális négyzetösszeg (Sum of Squares Total) a megfigyelt $y_i$ értékek és a vízszintes $\overline{y}$ egyenes (= nem regressziós egyenes) távolságát méri. Vegyük minden távolságnak a négyzetösszegét:

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

**Regressziós négyzetösszeg**

A regressziós négyzetösszeg (Sum of Squares for Regression):

$$SSM = SS_R = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$$

**Állítás.**

$$\sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$SS_T = SS_R + SS_{RES}$$

*Bizonyítás.*

$$\sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$\sum_{i=1}^{n} (y_i^2 - 2y_i\overline{y} + \overline{y}^2) = \sum_{i=1}^{n} (\hat{y_i}^2 - 2\hat{y}_i\overline{y} + \overline{y}^2) + \sum_{i=1}^{n} (y_i^2 - 2y_i\hat{y}_i + \hat{y_i}^2)$$

$$-\sum_{i=1}^{n} 2y_i\overline{y} = \sum_{i=1}^{n} \hat{y_i}^2 - \sum_{i=1}^{n} 2\hat{y}_i\overline{y} - \sum_{i=1}^{n} 2y_i\hat{y}_i + \sum_{i=1} \hat{y_i}^2$$

$$-\sum_{i=1}^{n} 2y_i\overline{y} = \sum_{i=1}^{n} 2\hat{y_i}^2 - \sum_{i=1}^{n} 2\hat{y}_i\overline{y} - \sum_{i=1}^{n} 2y_i\hat{y}_i$$

Tudjuk, hogy $y_i$ felírható a becsült érték és a hiba összegeként, vagyis $y_i = \hat{y}_i + e_i$ és ezt behelyettesítve:

$$-\sum_{i=1}^{n} 2(\hat{y}_i + e_i)\overline{y} = \sum_{i=1}^{n} 2\hat{y_i}^2 - \sum_{i=1}^{n} 2\hat{y}_i\overline{y} - \sum_{i=1}^{n} 2(\hat{y}_i + e_i)\hat{y}_i$$

$$-\sum_{i=1}^{n} 2(\hat{y}_i + e_i)\overline{y} + \sum_{i=1}^{n} 2\hat{y}_i\overline{y} = \sum_{i=1}^{n} 2\hat{y_i}^2 - \sum_{i=1}^{n} 2(\hat{y}_i + e_i)\hat{y}_i$$

mivel a hiba várható értéke 0, ezért a mennyiségek kiejtik egymást. □

**Determinációs együttható ($R^2$)**

A determinációs együttható megmutatja, hogy $X$ változékonysága mekkora százalékban magyarázza jól $Y$ változékonyságát:

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{RES}}{SS_T}$$

A lineáris kapcsolat erős ha az $R^2$ értéke 1-hez közeli, és gyenge, ha 0-hoz van közelebb. Ha $R^2 = 1$, akkor az azt jelenti, hogy az adatpontjaink tökéletesen illeszkednek az egyenesünkre. $R^2 = 0$ esetén az illesztett egyenesünk vízszintes, mert $X$ megváltozása nem játszik szerepet $Y$ változásában.

## 2.4. Többváltozós lineáris regresszió

Ha 2 vagy több magyarázóváltozóval rendelkezünk, akkor többváltozós lineáris regresszióról beszélünk. Ebben az esetben már érdekes a változók egyenkénti és az együttes hatása is. A megfigyelés alapján írjuk fel az $n$ egyenletet itt is mátrixos alakban:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} e_0 \\ e_1 \\ \vdots \\ e_n \end{bmatrix}$$

vagyis

$$y = X\beta + e,$$

ahol $y$ egy $n \times 1$–es szlopvektor, $X$ mátrix dimenziója $n \times (p+1)$–s, ha az $y$ tengelymetszet nem 0 és $n \times p$ az ellenkező esetben, $\beta$ egy $(p+1) \times 1$–es és $e$ egy $n \times 1$–es vektor.

## Mérőszámok a többváltozós lineáris regresszióban

A többváltozós regresszió mérőszámai hasonlóak az egyszerű lineáris regresszió mennyiségeihez.

Az átlagos négyzetösszeg (Mean of Squares for Model) a regressziós négyzetösszeg leosztva a becslés magyarázó változóinak számával:

$$MSM = \frac{SS_R}{p-1} = \frac{SS_R}{q}$$

A reziduálisok átlagos négyzetösszegénél (Mean of Squares for Error) a reziduális négyzetösszeget osztjuk a mintaelemszám és a magyarázó változók különbségével:

$$MSE = \frac{SS_{RES}}{n-p}$$

The sample variance of the residuals.

### Determinációs együttható ($R^2$)

A determinációs együttható ugyanaz mint a kétváltozós lineáris regressziónál (a tengelymetszet $\beta_0$-t is beleértve):

$$R^2 = \frac{SS_R}{SS_T}$$

### Korrigált determinációs együttható ($R^2_{adj}$)

A determinációs együttható a szabadságfokkal korrigálva figyelembe veszi, hogy hány független változót vontunk be a modell illesztésekor. Jelölje $n$ az összes magyarázó változót és $q$ a bevont magyarázó változók számát:

$$R^2_{adj} = 1 - \frac{n-1}{n-q-1}(1-R^2)$$

## 2.5. A módszer hipotézisvizsgálata

Hipotézisvizsgálatnál arra vagyunk kíváncsiak, hogy $x$ és $y$ szignifikáns kapcsolatban állnak-e egymással, ezért a $\beta$ paraméterek értékére teszünk feltevéseket [5]. Az egyváltozós lineáris regresszió hipotézisvizsgálata általában a következő:

$$H_0 : \beta_1 = 0$$

$$\underline{H_1 : \beta_1 \neq 0}$$

Tehát, ha a nullhipotézis igaz, akkor $y = \beta_0 + \beta_1 x$–ből látjuk, hogy $x$–nek nincs hatása $y$–ra, ezzel szemben az alternatív hipotézis szerint $x$ változása összefügg $y$ megváltozásával.

### 2.5.1. T-próba lineáris regresszió együtthatóinak vizsgálatára

A t-próba megmutatja, hogy a $\beta_i$ regressziós együtthatók külön-külön szignifikánsak-e a többváltozós lineáris regresszióban. A $\beta_i$ együtthatóra vonatkozó hipotézisvizsgálat:

$$H_0 : \beta_i = 0$$

$$\underline{H_1 : \beta_i \neq 0}$$

$H_1$–ben feltételezzük, hogy nincs lényeges hatása $\beta_i$–nek, ezért egy erős statisztikai ellenérvet keresünk t-próbával.

A próbastatisztika azon alapszik, hogy a paraméter becslése elosztva a becsült szórásával (standard error) t-eloszlást követ igaz $H_0$ esetén, $n - 2$ szabadsági fokkal, ahol $n$ a minta elemszáma. Vagyis:

$$T_0 = \frac{\hat{\beta}_i}{se(\hat{\beta}_i)}$$

A nullhipotézist elfogadjuk, ha a statisztika a következő intervallumban esik: $-t_{n-2,1-\alpha/2} < T_0 < t_{n-2,1-\alpha/2}$. Ennek következménye, hogy ha $T_0$ kívül esik a tartományon, akkor a $H_1$ hipotézist fogadjuk el, azaz lényeges hatása van a változónak a lineáris modellen belül. Egyoldali próbánál $H_1 : \beta_i > 0$ az ellenhipotézisünk és nullhipotézist $T_0 > t_{n-2,1-\alpha}$ esetén fogadjuk el.

A t-próbához tartozó p-értékek a számítógépes elemzés során segítenek eldönteni, hogy a magyarázó változók egyenként szignifikánsak-e a modellben vagy sem.

### 2.5.2. Hipotézisvizsgálat az együtthatók együttes hatására

A $\beta_i$ együtthatók együttes szignifikanciáját mérhetjük úgy, hogy összehasonlítjuk a jelenlegi többváltozóst modell és az együtthatók nélküli konstans modell illeszkedését [12]. A hipotézisvizsgálat $i$. feladata ebben az esetben:

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

$H_1: \beta_i \neq 0$, legalább egy $i$–re

Vizsgáljuk meg tesztstatisztikánk, $F_0$ eloszlását $H_0$ igaz volta mellett, ehhez felhasználom a már korábban definiált hányadosokat ($MSM$–t és $MSE$–t):

$$F_0 = \frac{MSM}{MSE} = \frac{SS_R/q}{SS_{RES}/n-p} = \frac{(SS_T - SS_{RES})/q}{SS_{RES}/n-p}$$

**Állítás.** A $H_0$ hipotézis teljesülése esetén $F_0$ standard F eloszlást követ, ezért F-próba alkalmazásával tesztelhető a hipotézis.

*Bizonyítás.* A $k$ szabadságfokú khí-négyzet eloszlás $k$ darab független standard normális eloszlás valószínűségi változóinak a négyzetösszege.

A khí-négyzet eloszlás definíciója szerint a független khí-négyzet változók összege is khí-négyzet eloszlású (additivitás). Speciálisan, ha $X_1, \ldots, X_n$ független khí-négyzet eloszlású változók $k_1, \ldots, k_n$ szabadságfokkal, akkor $Y = X_1 + \cdots + X_n$ is khí-négyzet eloszlásúak $k_1 + \cdots + k_n$ szabadságfokkal.

Ha $V_1 \sim \chi^2_{(m_1)}$ és $V_2 \sim \chi^2_{(m_2)}$ és $V_1$ és $V_2$ függetlenek, akkor

$$F = \frac{V_1/m_1}{V_2/m_2} \sim F_{(m_1,m_2)}$$

A modell F statisztikájának számítása a következő hányadossal történik:

$$F_0 = \frac{(SS_T - SS_{RES})/q}{SS_{RES}/n-p} = \frac{(\sum_{i=1}^n (y_i - \overline{y})^2 - \sum_{i=1}^n (y_i - \hat{y}_i)^2)/q}{\sum_{i=1}^n (y_i - \hat{y}_i)^2/(n-p)},$$

ahol $n$ az elemszám és $p = 1$+magyarázó változók száma.

$e_i = y_i - \hat{y}_i$–re, vagyis a hibára feltettük, hogy normális eloszlású, $y_i$–re szintén feltettük, hogy normális eloszlásúak, így $y_i - \overline{y}$ is az lesz. Tehát $(y_i - \overline{y})^2$ és $(y_i - \hat{y}_i)^2$ is $\chi^2$ eloszlású.

Ezért $\sum_{i=1}^n (y_i - \overline{y})^2$ és $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ is khí-négyzet eloszlású. Itt a szabadsági fok [13] alapján: $(n - p + q) - (n - p) = q$.

$$F_{q,n-p} = \frac{\chi^2_q/q}{\chi^2_{n-p}/n-p},$$

ahol $\chi^2_q$ és $\chi^2_{n-p}$ egymástól független Khí-négyzet eloszlásúak.

$SS_R$ és $SS_{RES}$ khí négyzet eloszlást követnek, megfelelő $q$ és $n-q$ szabadsági fokokkal, ezért $F_0$ éppen F eloszlású lesz, ha igaz a nullhipotézis.

A modell $F_0$ statisztikája az $F$ eloszlást követi $p-1$ és $n-p$ szabadsági fokokkal. A nullhipotézist elutasítjuk, ha $F_0 > F_{p-1,n-p}$. □

## 2.6. Legkisebb négyzetek módszere

A lineáris regresszió együtthatóinak a becslésére többféle módszer is létezik. Ezek közül a legelterjedtebb a legkisebb négyzetek módszere. Az eljárás célja, hogy minimalizálja a megfigyelt adatok és az illesztett modell különbségének a négyzetösszegét, vagyis a reziduális négyzetösszeget ($SS_{RES}$). Innen kapta a legkisebb négyzetek elnevezést (Ordinary Least Squares, OLS). A következő két fejezet vázát Nielsen (2013) [8] című könyve adja.

A hiba, azaz $y-X\beta$ négyzetösszegét megkapjuk ha a transzponálttal balról szorzunk. A későbbi számolás miatt szorozzunk be 1/2-el. A kapott mennyiséget nevezzük $\varepsilon$–nak. Ennek a minimalizálása ekvivalens a négyzetösszegek minimalizálásával. A levezetés abban az esetben érvényes, ha $X$ teljes rangú, ekkor $\det X \neq 0$, valamint ekkor lesz $X^\top X$ pozitív definit, és tudjuk invertálni.

$$\begin{aligned}
\varepsilon &= 1/2(y - X\beta)^\top (y - X\beta) \\
&= 1/2(y^\top y - y^\top X\beta - \beta^\top X^\top y + \beta^\top X^\top X\beta) \\
&= 1/2(y^\top y - 2\beta^\top X^\top y + \beta^\top X^\top X\beta)
\end{aligned}$$

A gyakorlatban előfordulhat, hogy $X$ nem teljes rangú, ekkor összefüggnek a magyarázó változók (tehát az oszlopok), de ilyen esetben lecsökkenthetjük a magyarázó változók számát és egy kisebb dimenziós feladatot kapunk. Ha sorok függnek össze, akkor ugyanolyan paraméterek tartoznak két vagy több megfigyeléshez.

Az egyenlet deriváltja $\beta$ szerint:

$$\frac{\partial \varepsilon}{\partial \beta} = -X^\top y + X^\top X\beta$$

Az egyenlet második deriváltja $\beta$ szerint:

$$\frac{\partial \varepsilon^2}{\partial \beta^2} = X^\top X$$

Mivel ez a kifejezés pozitív definit, ezért az első derivált zérushelyén minimumot kapunk.

A $\partial\varepsilon/\partial\beta = 0$ egyenletet átrendezve megkapjuk a $\beta$ magyarázóváltozónk becslését, melyet $\hat{\beta}_{OLS}$–al fogok a továbbiakban jelölni.

$$X^\top X\hat{\beta}_{OLS} = X^\top y$$

$$\hat{\beta}_{OLS} = (X^\top X)^{-1}X^\top y$$

Ha $X^\top X$ mátrix nem invertálható, akkor a becslést a Moore-Perose féle általánosított inverzzel átszorozva kaphatjuk meg. A becslésünk $y$–ra, $\hat{y}$–al jelölve:

$$\hat{y} = X\hat{\beta}_{OLS} = X(X^\top X)^{-1}X^\top y = Hy$$

ahol a $H = X(X^\top X)^{-1}X^\top$ mátrix az úgynevezett "kalap mátrix", mivel az transzformálja át $y$-t $\hat{y}$–ba, más szóval $H$ teszi a kalapot $y$–ra.

**2.3. Tétel** (Gauss-Markov)**.** *Lineáris regresszió esetén a legkisebb négyzetek módszere által adott becslés a legjobb lineáris torzítatlan becslés.*

## 2.7. Súlyozott legkisebb négyzetek módszere

A legkisebb négyzetek módszernél feltételezzük az adathalmazról, hogy nincsenek benne kívülálló értékek és nem sérül a homoszkedaszticitás tulajdonsága, mert ezek jelentősen befolyásolhatják az eljárás eredményét. A kiugró értékekre és a heteroszkedaszticitásra egy kevésbé érzékeny regressziós eljárás a súlyozott legkisebb négyzetek módszere (Weighted Least Squares, WLS) [8], [4].

A módszer a paraméter becslése során az illesztett egyenestől távolabb eső pontokat kisebb súllyal, míg az egyeneshez közeli pontokat nagyobb súllyal veszi figyelembe, így nem ad hamis képet az adatsorról.

Legyen a súlymátrix

$$W = \begin{bmatrix} w_1 & 0 & \dots & 0 \\ 0 & w_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w_n \end{bmatrix}$$

ahol az optimális súlyokat a rezduálisok szerint fejezhetjük ki:

$$w_i = \frac{1}{e_i^2}$$

Így a nagy reziduálisú pontokat kisebb súllyal veszi figyelembe, míg a a regressziós egyeneshez közeli pontoknak nagyobb lesz a jelentősége.

$$\begin{aligned} \varepsilon &= 1/2(y - X\beta)^\top W(y - X\beta) \\ &= 1/2(y^\top Wy - y^\top WX\beta - \beta^\top X^\top Wy + \beta^\top X^\top WX\beta) \\ &= 1/2(y^\top Wy - 2\beta^\top X^\top Wy + \beta^\top X^\top WX\beta). \end{aligned}$$

Az egyenlet $\beta$ szerint deriválva:

$$\frac{\partial \varepsilon}{\partial \beta} = -X^\top Wy + X^\top WX\beta$$

A második derivált:

$$\frac{\partial \varepsilon^2}{\partial \beta^2} = X^\top WX$$

Ez pozitív definit, így minimumot kaptunk $\varepsilon$–ra. A levezetés itt is abban az esetben érvényes, ha $X$ teljes rangú, ekkor $\det X \neq 0$, valamint ekkor lesz $X^\top X$ pozitív definit, és tudjuk invertálni.

Az első derivált egyenletét átrendezve megkapjuk a $\beta$ magyarázóváltozónk becslését, melyet $\hat{\beta}_{WLS}$–al fogok a továbbiakban jelölni.

$$X^\top WX\hat{\beta}_{WLS} = X^\top Wy$$

$$\hat{\beta}_{WLS} = (X^\top W X)^{-1} X^\top W y$$

A becslésünk a függő változóra, $\hat{y}$–al jelölve:

$$\hat{y} = X\hat{\beta}_{WLS} = X(X^\top W X)^{-1} X^\top W y = Hy$$

A kalap mátrix most $H = X(X^\top W X)^{-1} X^\top W$, mivel ez transzformálja át $y$-t $\hat{y}$–ba.

# 3. Kívülálló értékek

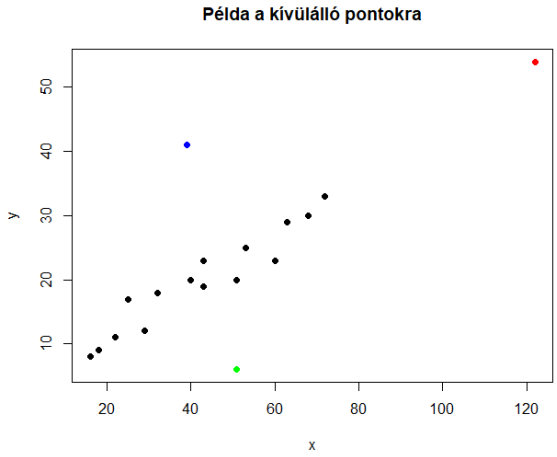
## 3.1. Különbség a kívülálló értékek között és azonosításuk

A kívülálló értékek olyan adatpontok, amelyek távolabb fekszenek a többi ponthoz képest. Ez többféleképpen is előfordulhat és nem is minden esetben jelentenek problémát. Amikor regressziós egyenest illesztünk egy adatkészletre, nem biztos, hogy foglalkozni kell a kiugró pontokkal annak ellenére, hogy egy szokatlan értékpár jelenik meg. Itt az x és y koordináta értékek is szokatlanok (túl magasak vagy alacsonyak és így kívül állnak a többi megfigyeléstől), de attól még az illesztett lineáris modellünk mentén fekszenek és nem törik meg a lineáris trendet. Előfordulhat azonban, hogy az x értéke az y koordináta értékének függvényében külön-külön átlagosnak számítanak, de azok együtt nem olyan kapcsolatban állnak, ezáltal messze esnek az adatfelhőnktől és megtörik a mintát.

### 3.1.1. Outlierek és nagy hatóerejű pontok

A kívülálló pontokon belül megkülönböztetjük a kiugró pontokat (outlierek) és azokat az extrém értékeket, amiknél nagy a hatóerő (leverage). Az outlierek olyan pontok, ahol a válasz ($y$ értéke) nem követi a trendet. A kívülálló pontok nagy hatóerővel általában szokatlan $x$ és $y$ érték kombinációja. Ezeknek a detektálására több módszer is létezik.

A táblázatból látszik, hogy a fekete ponthalmazra illesztett modell paraméter becslései eltérnek a többi modellétől. Más eredményt kapunk ha a kívülálló értékeket is hozzávesszük egyenként, vagy mindhármat egyszerre. A kék és a zöld pontok maguk




2. ábra. A kívülálló pontok változatai

felé húzzák az illesztett egyenest. Az is látszik, hogy a piros színű ponttal kapott becslések vannak a legközelebb a csak fekete pontokéból állók paraméter becsléseihez, mert az jól követi a lineáris trendet.

| Bevett pontok | $\beta_0$ | $\beta_1$ | $R^2$ |
|---|---|---|---|
| Minden pont | 3.7 | 0.3 | 0.652 |
| Fekete és piros pontok | 2.3 | 0.4 | 0.96 |
| Fekete és zöld pontok | 3.2 | 0.3 | 0.652 |
| Fekete és kék pontok | 5.08 | 0.38 | 0.558 |
| Csak fekete pontok | 3.05 | 0.39 | 0.912 |

## 3.2. A kalap-mátrix és a hatóerő kapcsolata

A kalap-mátrix és a hatóerő kapcsolatának vizsgálatához a [9] jegyzetet használom fel. Emlékeztetőül $\hat{y} = X\hat{\beta}$, amit az együtthatók becslésével továbbalakítva megkaptuk a becsült célváltozók egy alternatív felírását: $\hat{y} = X(X^\top X)^{-1}X^\top y$. Ez igazából:

$$\hat{y} = Hy,$$

ahol $H = X(X^\top X)^{-1}X^\top$ a kalap-mátrix volt, aminek az értékei csak a magyarázó változóktól függenek. Ennek a módszernek a segítségével a rendhagyó, kívülálló $x$ értékeket határozhatjuk meg. A képlet alapján felírhatjuk az $\hat{y}_i$–t, az $i$. becsült eredményváltozót az $y_1, \ldots, y_n$ és $H$ kalap-mátrix $i$. sorának a lineáris kombinációjaként:

$$\begin{aligned}
\hat{y}_1 &= h_{11}y_1 + h_{12}y_2 + \ldots + h_{1n}y_n \\
\hat{y}_2 &= h_{21}y_1 + h_{22}y_2 + \ldots + h_{2n}y_n \\
&\vdots \\
\hat{y}_n &= h_{n1}y_1 + h_{n2}y_2 + \ldots + h_{nn}y_n
\end{aligned}$$

Ha általános alakban írjuk az $i$. megfigyelésünkre, akkor adódik, hogy $y_i$ súlya $h_{ii}$ lesz, vagyis $h_{ii}$–vel mérni tudjuk, hogy mekkora ráhatása van a megfigyelt $y_i$–nek a saját megjósolt $\hat{y}_i$ értékére.

$$\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + \ldots + h_{ii}y_i + \ldots + h_{in}y_n$$

Ha $h_{ii}$ elem kicsi, akkor $y_i$–nek csak kis szerepe $\hat{y}_i$ alakulásában, ha $h_{ii}$ nagy, akkor azt mondjuk, hogy nagy hatóereje van (az angol szakirodalomban szerinti megnevezés: *high leverage observations*). Általában $h_{ii}$ akkor számít nagynak, ha az értéke legalább kettőszöröse a súlyok átlagának:

$$h_{ii} > 2\sum_{j=1}^{n}\frac{h_{jj}}{n}$$

**Állítás.** A kalap-mátrix nyoma a magyarázó változók száma: $trH = p$

*Bizonyítás.* $trH = tr(X(X^\top X)^{-1}X^\top) = tr(X^\top X(X^\top X)^{-1}) = trI_p = p$ □

Mivel a $h_{ii}$ súlyok összege a magyarázó változók száma, vagyis $trH = p$, így a hányadost átalakítva $\sum_{i=1}^{n}\frac{h_{ii}}{n} = \frac{p}{n}$ szerint az alsó határra egy új képletet kapunk:

$$h_{ii} > 2 \cdot \frac{p}{n}$$

tehát a kalap-mátrixszal detektálhatjuk a szokatlan $x$ értékű pontokat.

## 3.3. Studentizált reziduálisok

A studentizált reziduálisokkal azonosíthatjuk a rendellenes $y$ értékeket [7], [15].

Az $i$. megfigyelésre vonatkozó közönséges (nem standardizált) reziduális a megfigyelt és a jósolt érték különbsége:

$$e_i = y_i - \hat{y}_i$$

Ha egy adott $x$ érték mellett az $y$ értéke $\hat{y}$–hoz képest kiugró, akkor a hiba nagy lesz és kívülálló pontnak tekinthető, de mivel a kiugró pontok maguk felé húzzák a regressziós egyenest, ezért a hiba nem minden esetben lesz nagy.

Standardizáljuk a reziduálisokat a terjedelem miatt úgy, hogy a hibák átlaga 0 és szórásuk 1 legyen. Így a közönséges reziduálist a standard hibával leosztva megkapjuk az $i$. studentizált reziduálist:

$$r_i = \frac{e_i}{se(e_i)} = \frac{e_i}{\sqrt{MSE(1-h_{ii})}} = \frac{e_i}{\sigma\sqrt{1-h_{ii}}}$$

A standard hiba becslésével osztunk, pont mint a Student-féle t-próbánál, ezért $t$-eloszlást követ $n-k$ szabadsági fokkal, ahol $n$ az elemszám, $k$ a magyarázó változók száma. Ha egy megfigyelés studentizált reziduálisa abszolút értékben nagyobb mint 2 (vagy 3), attól függően, hogy milyen szignifikancia szintet választottunk, akkor azt mondhatjuk, hogy az kívülálló pont. A t-eloszlás határait 5%–os szignifikanciaszint mellett kiszámolva $\pm 1.96$-ot kapunk, ezt kerekítjük 2–re és így kaptuk az alsó határt.

### 3.3.1. Törölt studentizált reziduálisok

A studentizált reziduálisokban nem tűnik ki minden szokatlan $y$ érték, mivel a lehetséges kívülálló értékek maguk felé húzzák az egyenest és nem lesz mindig kiugró a hiba. Hogy megtaláljuk ezeket a lehetséges outliereket is, használjuk a törölt studentizált reziduálisokat [10], melynek működése a következő: a modell illesztésekor hagyjuk ki az $i$. adatpontot a mintából, és számoljuk ki így minden pontra az eltérést. Jelölje $y_i$ az $i$. megfigyelés válaszát, $\hat{y}_i^{(-i)}$ pedig az $i$. becsült válasz úgy, hogy az $i$. pontot már töröltük a $\beta$ paraméterek becslésénél. A pont kihagyásával a hiba értéke:

$$e_i^{(-i)} = y_i - \hat{y}_i^{(-i)}$$

Ha a valódi és a becsült érték különbségét leosztjuk a standard szórással, akkor megkapjuk törölt studentizált reziduális:

$$t_i = \frac{e_i^{(-i)}}{se(e_i^{(-i)})} = \frac{e_i^{(-i)}}{\sqrt{MSE^{(-i)}(1-h_{ii})}} = \frac{e_i^{(-i)}}{\sigma^{(-i)}\sqrt{1-h_{ii}}}$$

A studentizált reziduálisoknál az összes megfigyelés alapján osztottunk a reziduálisok négyzetösszegével ($MSE$) és a kalap-mátrix $i$. diagonálisával. Itt a mean square error ($MSE^{(-i)}$) egy olyan modellből jött létre, amiből már kitöröltük az $i$. pontot, viszont $H$ számításakor az $X$ mátrixban az összes megfigyelés szerepel. A törölt studentizált reziduálisok $t$-eloszlást követnek $(n-1)-k-1 = n-k-2$ szabadsági fokkal. Az előzőhöz hasonlóan általában itt is 2 a határ (abszolút értékben), ami felett már lehetséges, hogy a megfigyelés egy kívülálló pont.

## 3.4. Cross-validation és hiba korrekció

A Cross-validation (CV) egy az adatbányászatban is gyakran használt algoritmus, amely jól méri a prediktív modell teljesítményét. A $k$-szoros Keresztvalidáció $k$ egyenlő részhalmazra bontja a megfigyelések halmazát, ahol $k-1$ darab tanítóhalmazon építjük fel a modellt, majd a maradék egy úgynevezett teszthalmazon mérjük a modell pontosságát. Ezt minden halmazra alkalmazva $k$ darab modellt alkotunk és ezeknek a $k$ darab teljesítményét átlagolva kapjuk meg a becslést.

### Leave-One-Out

A Cross-validation egy speciális esete a "Hagyj-ki-egyet Keresztvalidáció" ("Leave-one-out Cross-validation", LOOCV) [9]. Itt $k = n$, ahol $n$ mintaelemszám, tehát egyetlen pontot használunk ellenőrzésre, a többi $n-1$ pont pedig a tanítóhalmaz. Ebben az esetben a modellt $n$–szer illesztjük, ami nagy mintaelemszám esetén drága.

Kérdés: Hogyan változik a modellünk, ha kihagyjuk az $i$. pontot? Becsüljük meg az együttható vektort, az $i$. pont törlése esetén ezt jelölje $\hat{\beta}^{(-i)}$.

**Leave-More-Then-One-Out**

Előfordulhat, hogy több kívülálló pont is van az adathalmazunkban. Ebben az esetben, ha kiveszünk egy ilyen pontot, akkor az nem fog nagy változást eredményezni a modellillesztésnél. Az összes kiugró pont eltávolításával jó modellt kapnánk, de ez még a Leave-One-Out–nál is költségesebb lenne, mivel $\binom{n}{k}$ variációt kéne megfontolni.

## 3.5. Cook távolság

A Cook távolság számítására a pont $x$ és $y$ értéke is hatással van és gyakran használják a legkisebb négyzetek becslésnél, hogy azonosítsák a befolyásosabb pontokat [9]. A Cook távolság fogalmát R. D. Cook 1977-ben vezette be [14]. A módszerben $\hat{\beta}^{(-i)}$ távolságát mérjük $\hat{\beta}$–tól, ahol $\hat{\beta}^{(-i)}$ az a becslés amiből eltávolítottuk az $i$. pontot, $\hat{\beta}$ pedig az összes megfigyelésből számolt paraméterbecslés. Azt vizsgáljuk, hogy az $i$. megfigyelés elhagyása mennyire mozdítja el a regressziós egyenest. Ha ez a távolság nagy, akkor a megfigyelésünk befolyása is nagy és kívülálló pontnak tekinthető. A Cook távolság az $i$. pontra:

$$D_i = \frac{(\hat{\beta} - \hat{\beta}^{(-i)})^\top (\hat{\beta} - \hat{\beta}^{(-i)})}{(p+1)\sigma^2}$$

Ha a Cook-távolság nagyobb mint 0.5, akkor már elképzelhető, hogy a kihagyott pont egy outlier. Ezzel ekvivalens, ha a reziduálisok ($e_i = y_i - \hat{y}_i$) és a kalap mátrix ($h_{ii}$) kombinációjával fejezzük ki a Cook távolságot:

$$D_i = \frac{(y_i - \hat{y}_i)^2}{p \cdot MSE} \left[ \frac{h_{ii}}{(1 - h_{ii})^2} \right]$$

## 3.6. Mahalanobis távolság

A Mahalanobis távolság segítségével a többdimenziós outliereket detektálhatjuk a válasz változó nélkül úgy, hogy egy adott pont és a többi pont eloszlásának távolságát vizsgáljuk [16], [6]. Ha egy pontnak sokkal nagyobb a Mahalanobis távolsága a többi ponthoz képest, akkor az kívülálló pont. Jelölje $\mu$ az átlagot, $\sigma$ szórást. Fejezzük ki az

$x_i$ pont és a normális eloszlású adatfelhőnk közepének a távolságát a következőképpen:

$$\frac{x_i - \overline{x}}{\sigma}$$

Ha standardizáljuk a szórással $i$. pont távolságát az átlagtól, akkor megőrzi az eredeti eloszlás alakját, de a szórása 1 és az átlag 0 lesz, ezt nevezzük Z-pontszámnak. A hányados megmutatja, hogy $x_i$ hány szórásnyira van az átlagtól,és ennek az értéke akkor jó, ha adott $x_i = k$ esetén a hányados kisebb vagy egyenlő lesz mint $1/k^2$. Emeljük négyzetre távolságot:

$$\frac{(x-\mu)^2}{\sigma^2} = (x-\mu)(\sigma^2)^{-1}(x-\mu)$$

A mérést többváltozós adatokon végezzük és az outlierek azonosításához számoljuk ki mind az $n$ megfigyelésre a Mahalanobis távolságot ($MD_i$). Legyen a kovariancia mátrix:

$$\Sigma = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x})^\top$$

$$MD_i = \sqrt{(x-\mu)\Sigma^{-1}(x-\mu)}$$

## 3.7. Robusztus regressziós módszer

A robusztus regressziós eljárást akkor érdemes alkalmazni, ha a legkisebb négyzetek módszernél nem teljesül mindegyik feltétel [11], [10]. Ez egy olyan alternatíva a lineáris regresszióban, ami kevésbé szigorúbb feltételeket kér és tompítani tudja a kívülálló pontok hatóerejét a modellillesztéskor. Mivel a kiugró pontoknak nagy a befolyásuk, ezért lekicsinyíti a hatásukat, hogy a pontok döntő többségére illesszen jobb modellt. Emlékeztetőül, az $n$ elemű többváltozós mintánkra illesztett modell hibája:

$$\varepsilon_i = y_i - x_i^\top \beta$$

A legkisebb négyzeteknél a négyzetre emelés miatt még nagyobbak lesznek a reziduálisok, és így a kívülálló pontok még jobban maguk felé húzzák a regressziós egyenest. Emlékeztetőül a $\beta$ paraméter közönséges legkisebb négyzetek becslése:

$$\beta_{OLS} = \arg\min \sum_{i=1}^{n}(\varepsilon_i)^2$$

A robusztus regresszióban többféle módszert is használhatunk az együtthatók becslésére, ehelyett minimalizálhatjuk például (least absolute deviation):

$$\beta_{LAD} = \arg\min \sum_{i=1}^{n} |\varepsilon_i|$$

Ha a $\beta$ együtthatók becslésénél a hiba abszolút értékét minimalizáljuk, akkor a kisebb hiba miatt kisebb befolyása lesz a kívülálló pontoknak és az egyenes jobban fogja a valódi trendet követni.

Egy másik gyakori becslés az M-becslés, amely a reziduálisok egy olyan függvényének összegét minimalizálja, amely kevésbé gyorsan növeli a hiba összeget:

$$\beta_M = \arg\min \sum_{i=1}^{n} \rho(\varepsilon_i)$$

A legkisebb négyzetek módszerében a $\rho$ a négyzetes függvény.

## 3.8. Rezisztens regressziós módszer

A rezisztens statisztika olyan statisztika, amely "kikerüli" a kiugró értékek befolyását. Például egy szélsőséges pont az átlag értékére hat, de a mediánéra nem. Innen jön a rezisztens regresszió elnevezés, amely egy olyan regressziós módszer, amiben nincs hatásuk az outliereknek a $\beta$ együtthatók becslésére. Ezt úgy próbálja elérni például, hogy az adathalmaznak levágja az alsó és felső 25%–át, és a maradékra illeszti a modellt. A paraméterek becslése hasonlóan történik a robusztus regresszióéhoz [10].

Végezzük el a legkisebb négyzetek módszerét úgy, hogy csak a $k \leq n$ legkisebb reziduálisú pontok négyzetösszegét minimalizáljuk (Least Trimmed Sum of Squares):

$$\beta_{LTS} = \arg\min \sum_{i=1}^{k} (\varepsilon_i)^2$$

Így ha kivesszük a legnagyobb hibájú pontokat, kisebb hatása lesz a kívülálló pontoknak. $k = n$ esetén a közönséges legkisebb négyzetek becslést kapjuk.

Ha pedig csak a $k \leq n$ legkisebb reziduálisok abszolút értékének összegét minimalizáljuk (Least Trimmed Sum of Absolute Deviations):

$$\beta_{LTA} = \arg\min \sum_{i=1}^{k} |\varepsilon_i|$$

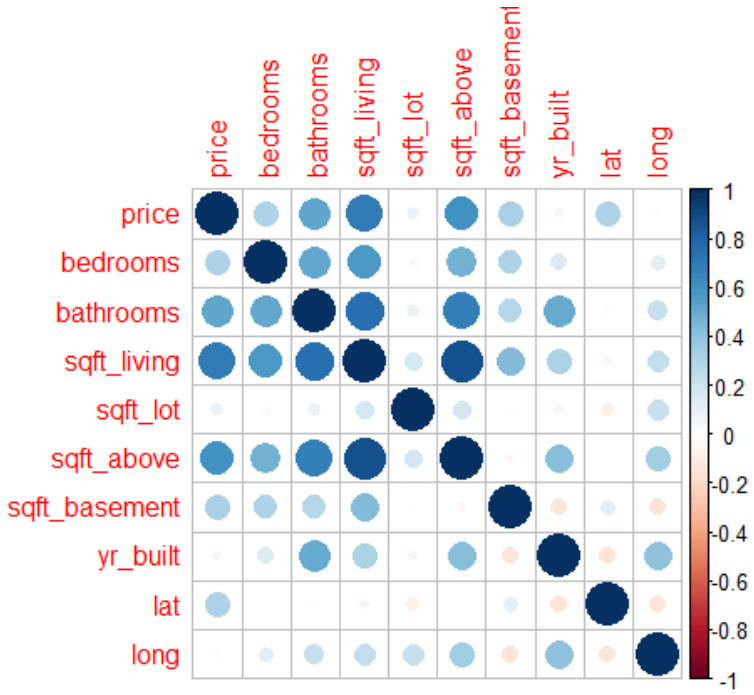
Itt $k = n$ esetén simán $\beta_{LAD}$ becslését kapjuk.

## 3.9. Hogyan kezeljük a kívülálló értékeket?

A kívülálló értékeket hatását többféleképpen is kezelhetjük. Ha az adatpont azért kiugró, mert hibás, akkor töröljük vagy javítsuk. Ha a pont értéke nem hiba miatt problémás, akkor végezzük el az elemzést az outlierrel együtt és anélkül is. A bemutatott mérésekkel könnyen detektálhatjuk a kiugró pontokat. Ezután kivehetjük az outliert a modellépítés előtt, de akkor azt a dokumentációban fel kell tüntetni, hogy melyik pontot vettük ki és miért. Ha nem vesszük ki az ilyen szélsőséges értékeket, akkor a paramétereket becsülhetjük kevésbé érzékeny regressziós eljárásokkal, például robusztus vagy rezisztens regresszióval. A következő fejezetben ezeket a módszereket egy gyakorlati példán keresztül szemléltetem.

# 4. Szimulációk R-ben

Az adathalmaz az amerikai King megyében eladott ingatlanok paramétereit tartalmazza 2014–2015 között. Összesen 21613 darab házat foglal magába. Az adatkészlet forrása: *https://www.kaggle.com/harlfoxem/housesalesprediction*. A változók: ár (price), hálószobák száma (bedrooms), fürdőszobák száma (bathrooms), ingatlan belterülete (sqft_living), kert területe (sqft_lot), szintek száma (floors), tengerszint feletti magasság (sqft_above), pince területe (sqft_basement), építés éve (yr_built), elhelyezkedés szélességi foka (lat), hosszúsági fok (long).



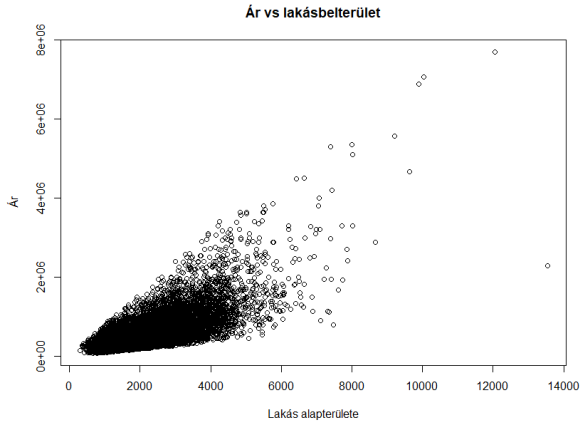

3. ábra. A korreláció nagysága az egyes változók között

Legyen az ár (price) a célváltozónk. Az árral legjobban korreláló magyarázóváltozók a fürdőszobák száma (bathrooms), a ház belterülete (sqft_living) és a ház tengerszint feletti magassága (sqft_above), de mivel ezek a változók egymással is korrelálnak, ezért először a lakás ára és a ház belterülete közötti lineáris kapcsolatot vizsgálom.

```
Call:
lm(formula = price ~ -1 + sqft_living, data = kc_house_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1282228  -154542   -35315    92984  4529775

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
sqft_living 263.0892     0.7839   335.6   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 262000 on 21612 degrees of freedom
Multiple R-squared:  0.839,     Adjusted R-squared:  0.839
F-statistic: 1.126e+05 on 1 and 21612 DF,  p-value: < 2.2e-16
```

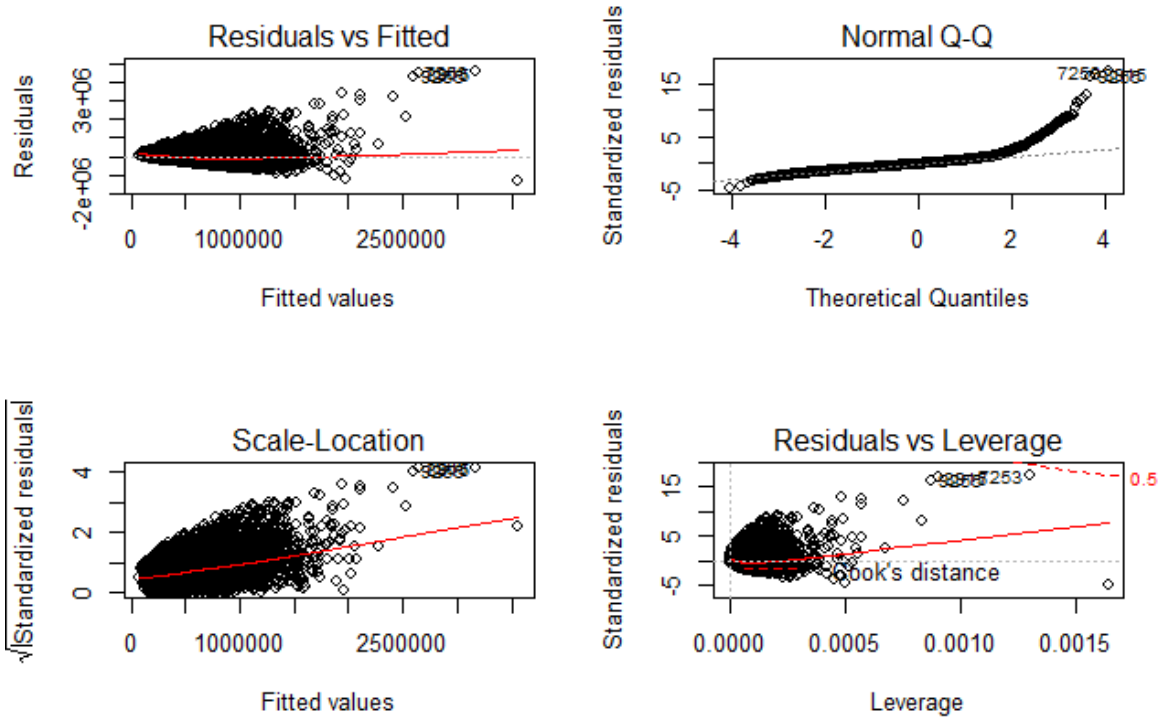
4. ábra. A klasszikus lineáris regresszió outputja

Az egyszerű lineáris regresszió eredményéből látszik, hogy egységnyi változás a belterülteben 263.08$–nyi változást jelent az árban. Az $R^2$ értéke elég magas, az ingatlan alapterületének a változékonysága 83.9%–ban magyarázza az ár változékonyságát. Tehát szignifikáns szerepe van az ár változásában, ezt a '***' is jelöli.



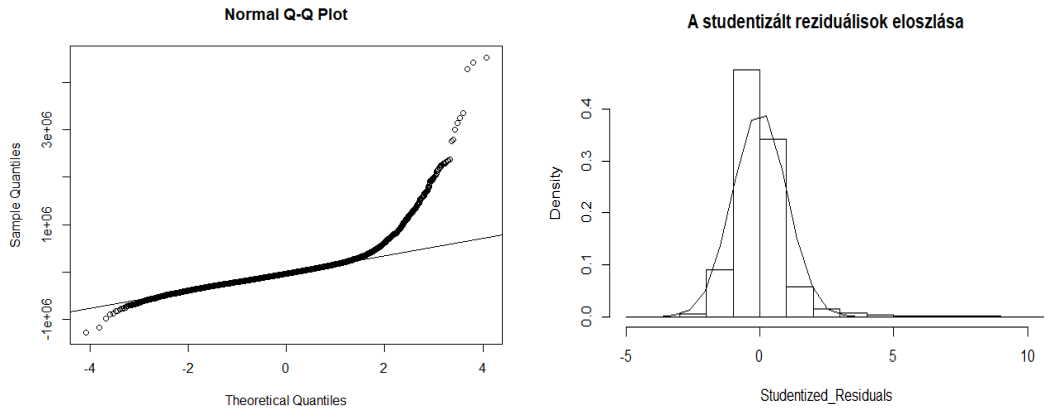

5. ábra. Az ár és a lakásbelterület kapcsolata

Az 5. ábrában megjelenik a reziduálisoknál tárgyalt heteroszkedaszticitás.




6. ábra. Reziduálisok

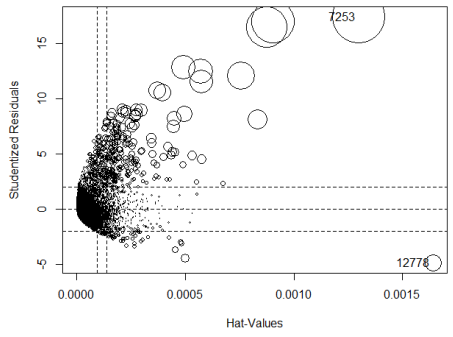
A 6. ábra szerint a 7253, 9255 és a 3915-ös pontok lehetnek problémásak. Az első ábrán a pontok egy vízszintes vonal mentén szóródnak, így biztosan nem nem-lineráis kapcsolat van a magyarázó és a válasz változó között. Mivel nem egyenletesen szóródnak, ezért sérül az egyik feltétel. A második ábra akkor lenne jó, ha a reziduálisok a szaggatott egyenes mentén helyezkednének el, így lehet, hogy nem normális eloszlásból származnak. A harmadik ábrával ellenőrizhetjük a homoszkedaszticitást. Ez akkor fogadható el, ha az adatok egy vízszintes mentén feküdnének. A mi esetünkben egyre jobban és jobban szóródnak, ez a heteroszkedaszticitást jelenti, vagyis a szórás nem állandó. A negyedik ábra segítségével kiugró pontokat detektálhatunk, a Cook távolság alapján. Azokkal a pontokkal lehet baj, ami az a jobb alsó vagy felső sarokban vannak (a szaggatott vonalon kívül). Mivel nem esik ki egyik pont sem a szaggatott vonalon kívül,



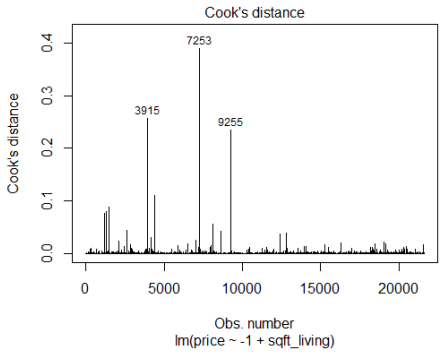

7. ábra. A reziduálisok Q-Q Plotja és a studentizált reziduálisok eloszlása

így a Cook távolság alapján mindegyik megfelelő. Ez nem azonosít minden szélsőséges értéket!

A 7. ábrák segítségével a reziduálisok normalitását vizsgálom meg jobban. A normalitás feltétele nem igaz az ábrák szerint, mert a pontok erősen eltérnek az egyenestől és a jobb felső sarokban sok kívülálló pont jelenik meg, így a további eredményeket feltételekkel kezelem.




8. ábra. A studentizált reziduálisok és a kalap-mátrix kapcsolata




9. ábra. A megfigyelések Cook távolsága

A 9. ábra Cook távolságok szemléletesen, megszámozza a lehetséges kívülálló pontokat. Ha az így kiszűrt pontokat kivesszük az adathalmazból, akkor megváltoznak az illesztett lineáris modell paraméterei is. Az eredmény így:

```
Call:
lm(formula = price ~ -1 + sqft_living, data = house_cd)

Residuals:
     Min       1Q   Median       3Q      Max
-1265047  -151731   -33105    95106  3365148

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
sqft_living 261.8203     0.7696   340.2   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 256800 on 21609 degrees of freedom
Multiple R-squared:  0.8427,    Adjusted R-squared:  0.8427
F-statistic: 1.158e+05 on 1 and 21609 DF,  p-value: < 2.2e-16
```

10. ábra. A lineáris regresszió outputja a Cook távolsággal azonosított kívülálló pontok nélkül

Tehát a 3915, 7253, 9255-s pontok kivételével $\beta = 263.08$–ból $\beta = 261.82$ lesz, és megnő az $R^2$ értéke is, az eredeti 0.839–ről 0.842–re emelkedett.

```
Call:
lqs.formula(formula = price ~ -1 + sqft_living, data = kc_house_data,
    method = "lms")

Coefficients:
sqft_living
      221.7

Scale estimates 175516 161986
```

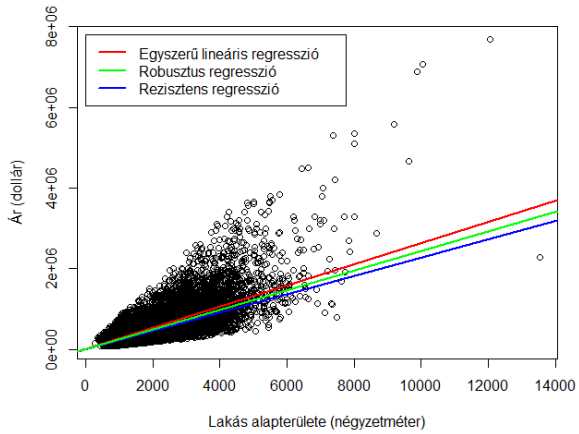
11. ábra. A rezisztens regresszió outputja

A legkisebb négyzetek módszerében a kívülálló pontok nagy súlyt kapnak, míg a rezisztens regresszió kikerüli ezek hatását. Így az outputból $\beta_1 = 221.7$ kaptunk, ami sokkal kisebb az eredeti $\beta_1 = 263.08$–hoz képest. A robusztus regresszió lekicsinyíti a kívülálló pontok befolyását, az együttható becslésére ezzel $\beta_1 = 243.21$ kaptunk.

```
Call: rlm(formula = price ~ -1 + sqft_living, data = kc_house_data)
Residuals:
     Min       1Q   Median       3Q      Max
-1019250  -114370     2876   128035  4769256

Coefficients:
            Value    Std. Error t value
sqft_living 243.2153   0.5612   433.3799

Residual standard error: 179100 on 21612 degrees of freedom
```

12. ábra. A robusztus regresszió outputja

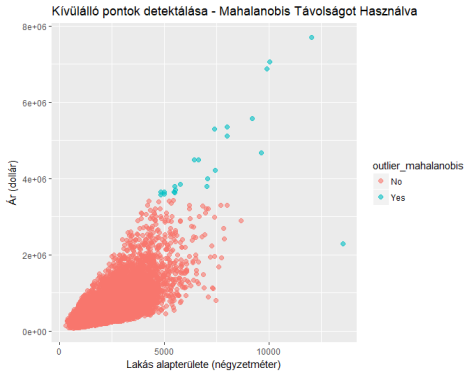
A 13. ábrában a rezisztens, a robusztus és sima lineáris regresszió egyeneseit szemléltetem. A robusztus és rezisztens regresszió egyenesei az egyszerű legkisebb négyzetek módszerével kapott modell egyenese alatt helyezkednek el, szemmel láthatólag is, a kívülálló pontok főleg felül helyezkednek el és a két módszer kevésbé érzékeny a szélsőséges értékekre.




13. ábra. A különböző regressziós egyenesek szemléletesen

| Regressziós módszer | $\beta_1$ | $R^2$ |
|---|---|---|
| Klasszikus regresszió | 263.08 | 0.839 |
| Robusztus regresszió | 243.10 | - |
| Rezisztens regresszió | 221.7 | 0.981 |
| Mahalanobis távolsággal kivett pontok | 259.11 | 0.85 |
| Cook távolsággal kivett pontok | 261.82 | 0.842 |

Látszik, hogy vannak befolyásosabb pontok ezért kaptunk más értékeket a paraméterre. Mindegyik eljárás máshogy kezeli a messzebb eső pontokat, de mivel a kívülálló pontoknak sincs pontos definíciója arra, hogy mikor lesz egy pont kiugró, ezért nincs legjobb módszer a detektálásukra.




14. ábra. Outlierek azonosítása a Mahalanobis távolsággal

A Mahalanobis távolság az adatok eloszlása alapján vizsgálja a kívülálló értékeket. Összesen 23 ilyen értéket talált ez a módszer. Ezeket a pontokat kivéve a lineáris modellünk paraméterei:

```
Call:
lm(formula = price ~ -1 + sqft_living, data = house_new)

Residuals:
     Min       1Q   Median       3Q      Max
-1138157  -146219   -28011    99175  2296183

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
sqft_living  259.112      0.739   350.6   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 245500 on 21589 degrees of freedom
Multiple R-squared:  0.8506,    Adjusted R-squared:  0.8506
F-statistic: 1.23e+05 on 1 and 21589 DF,  p-value: < 2.2e-16
```

15. ábra. A lineáris regresszió outputja a Mahalanobis távolsággal azonosított kívülálló pontok nélkül

A $\beta$ paraméterünk most is csökkent, az új értéke $\beta_1 = 259.11$, $R^2$ értéke pedig javult, az eredeti 0.839 helyett most már 0.85 lett, tehát a csökkentett adathalmazunkkal 1%–kal jobban tudjuk magyarázni az ár változását az alapterület változásával.

A többváltozós lineáris modellben továbbra is az ingatlan árat ($price$) fogom jósolni. A magyarázó változók közé a ház belterületét ($sqft\_living$), elhelyezkedésének szélességi fokát ($lat$) és az építés évét ($yr\_built$) fogom bevenni. A 3. ábrában jól látszik, hogy a tengerszint feletti magasság ($sqft\_above$), fürdőszobák száma ($bathrooms$) és az alapterület ($sqft\_living$) változók között erős korrelációs kapcsolat van, így nincs értelme mindegyiket egyszerre bevenni a modell építésekor. A többváltozós modell paraméterei:

```
Call:
lm(formula = price ~ -1 + sqft_living + lat + yr_built, data = kc_house_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1717755  -132142   -19851    98269  4018991

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
sqft_living    307.406      1.928  159.46   <2e-16 ***
lat         112731.577   2361.803   47.73   <2e-16 ***
yr_built     -2770.465     57.554  -48.14   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 248500 on 21610 degrees of freedom
Multiple R-squared:  0.8552,    Adjusted R-squared:  0.8552
F-statistic: 4.254e+04 on 3 and 21610 DF,  p-value: < 2.2e-16
```

16. ábra. A többváltozós lineáris regresszió outputja

A kapott regressziós modell egyenlete:

$$price = 307.41 \cdot sqft\_living + 112731.57 \cdot lat - 2770.47 \cdot yr\_built$$

Tehát az alapterület és a szélességi fok növelésével az ár is nőni fog, míg az évek fordítottan hatnak az árra, minél újabb, annál jobban csökken az ingatlan ára. Az F-statisztikánál a p-érték nagyobb mint a szignifikancia szint, ezért elutasíthatjuk a 2.5.2 fejezet nullhipotézisét, vagyis a többváltozós modell illeszkedése jobb, mintha kivennénk valamelyik magyarázó változót és úgy illesztenénk modellt. A három változó változékonysága jelentősen magyarázza az ár célváltozó változékonyságát, a determinációs együtthatóra $R^2 = 0.8552$–t, azaz jobb értéket kaptunk mint az egyváltozós regressziónál.

# 5. Összefoglalás

A kívülálló pontok a mindennapi életben is jelentősek. Létezésük plusz információt is adhat az adatról, de hibás adatfelvétel miatt is megjelenhetnek, ezért kell foglalkozni az azonosításukkal és a kezelésükkel.

Egy statisztikai elemzés során könnyen befolyásolhatják a modellt, rossz képet adva az adathalmazról. A dolgozat során olyan eljárásokat mutattam meg, amelyekkel detektálni lehet a szélsőséges pontokat, hogy azok ne torzítsák az eredményt az elemzés elkészítése során. Ezek az eljárások távolság (pl. Cook távolság) vagy eloszlás (pl. Mahalanobis távolság) alapján azonosítják a kiugró pontokat. Nagy mintaelemszámnál költséges ezeknek a pontoknak az azonosítása, ezért használhatunk alternatív regressziós módszereket. Az egyszerű lineáris regresszióhoz képest két kevésbé érzékeny regressziós módszer, a robusztus és a rezisztens regressziók abban az esetben is hasznosak, ha túl sok van, vagy nehéz azonosítani a kívülálló pontokat, így nem kell kivenni az extrém értékeket, mert ezekben a regressziókban nincs nagy befolyásuk a modellre.

Az utolsó fejezetben ingatlanárakat jeleztem előre egy valós amerikai adathalmazon, és az abban lévő lehetséges kívülálló pontokat szűrtem ki. A regressziós egyenes változói az ár és a ház alapterülete voltak, és a kiugró adatpontok azonosítását a bemutatott eljárások segítségével végeztem. A ritka és drága luxusvillák így nem húzzák el regressziós egyenest, és az átlagos vásárlók számára jobban lehet az ingatlan árát jósolni, ha van egy adott alapterület elképzelésük.

Összefoglalva tehát, a dolgozatban több módszert is vizsgáltam a szélsőséges adatok detektálására, hogy a befolyásosabb pontok ne húzzák el az regressziós egyenest és pontosabb becslést kapjunk a modell együtthatóira. Az ilyen pontokra nincs elfogadott, pontos definíció, ezért nincs legjobb módszer a detektálásukra és a kezelésükre, érdemes több eljárást is alkalmazni. Az általam vizsgált adathalmazra a rezisztens regresszió adta a legjobb megoldást.

# Hivatkozások


[1] Yan X. (2009), Linear Regression Analysis: Theory and Computing, World Scientific

[2] George A. F. S., Alan J. L. (2003), Linear Regression Analysis

[3] Cox D. R., Snell E. J. (1968), A general definition of residuals, Journal of the Royal Statistical Society

[4] Strutz T. (2016), Data Fitting and Uncertainty (A practical introduction to weighted least squares and beyond)

[5] Savin N.E., Multiple hypothesis testing

[6] Hodge V. J., Austin J. (2004), A Survey of Outlier Detection Methodologies, Artificial Intelligence Review

[7] Zrínyi M., Katona É., Szántó I., Páll D., A lineáris regressziót befolyásoló esetek diagnosztikája, Statisztikai Szemle, 90. évfolyam 7—8. szám

[8] Nielsen A. A. (2013), Least Squares Adjustment: Linear and Nonlinear Weighted Regression Analysis

[9] Shalizi C. (2015), Outliers and Influential Points

[10] Iain P., Laura S., Derek Y., Regression Methods
https://onlinecourses.science.psu.edu/stat501/node/250/

[11] Colin C., Robust Regression and Outlier Detection with the Robustreg Procedure

[12] Matt B. (2008), Multiple Hypothesis Testing: The F-test

[13] Ezequiel U., 4 Hypothesis testing in the multiple regression model

[14] Cook R. D. (1977), Detection of Influential Observations in Linear Regression

[15] Allen J. P. (1976), The statistics of residuals and the detection of outliers

[16] Maesschalck R., Jouan-Rimbaud D., Massart D.L. (2000), The Mahalanobis distance

[17] Iain P. (2006), Applied Regression Modeling: A Business Approach