ÖKONOMETRIAI ALAPOK

Dr. Sipos Tibor
Dr. Török Ádám
Szabó Zsombor

KUKG

BUDAPESTI MŰSZAKI ÉS GAZDASÁGTUDOMÁNYI EGYETEM
Közlekedésmérnöki és Járműmérnöki Kar

Ökonometria

- Statisztikai módszerek alkalmazása a gazdasági adatok elemzésében
- Lépései:
  - Empirikusan tesztelhető modell
  - Adatgyűjtés
  - Modell becslése, ellenőrzése
  - Előrejelzés, döntéselőkészítés

Forrás: Elek Péter, Bíró Anikó: Ökonometria (2010)

Tartalom

- Empirikusan tesztelhető modell
- Adatgyűjtés
- Modell becslése
- Előrejelzés, döntéselőkészítés
- Órai példa bemutatása

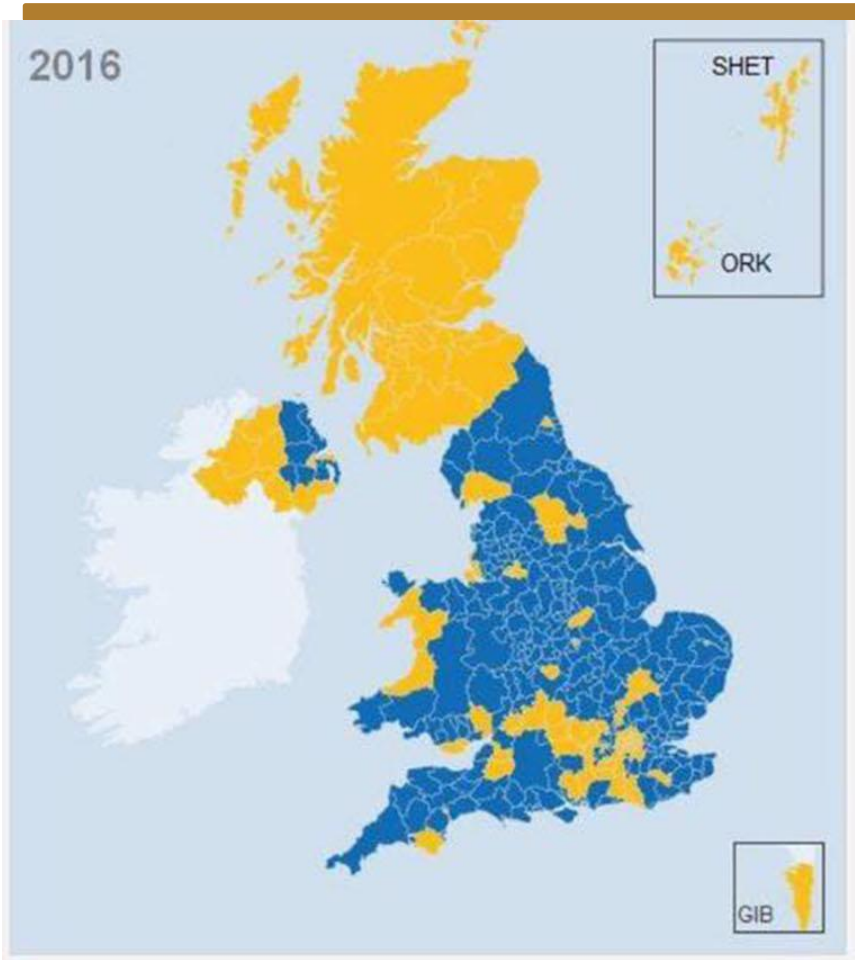
Tartalom

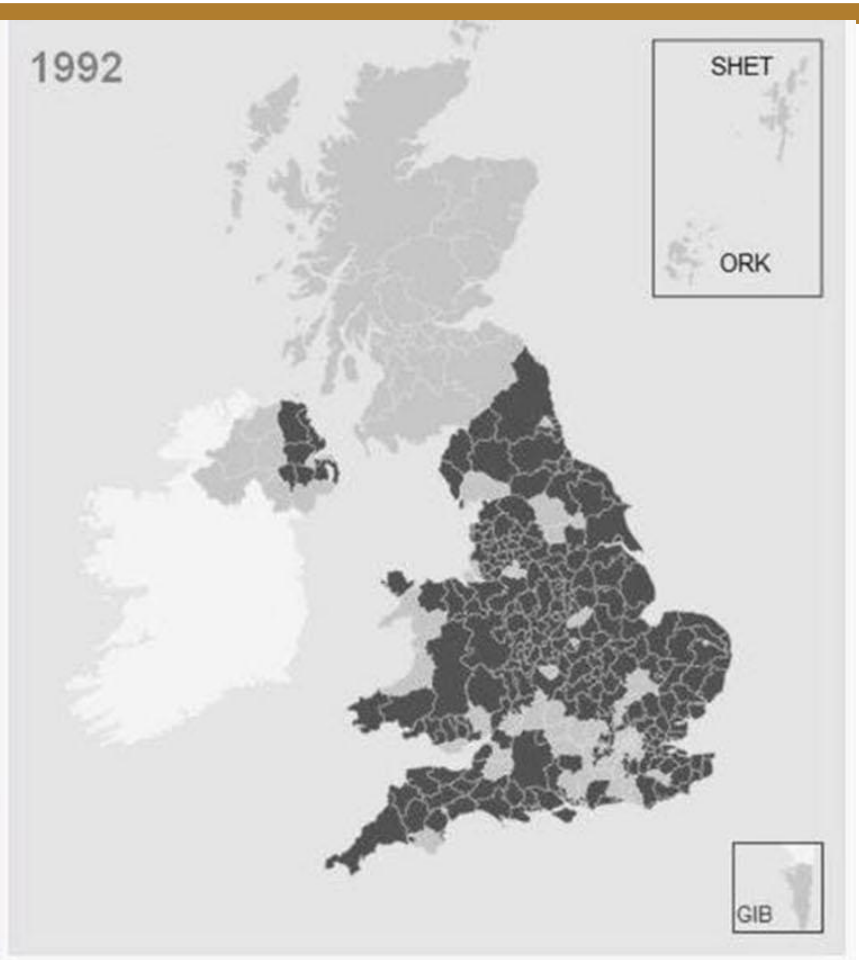
- Empirikusan tesztelhető modell
- Adatgyűjtés
- Modell becslése
- Előrejelzés, döntéselőkészítés
- Órai példa bemutatása

Modell felállítása

- Nagyon fontos a megfelelő modell felállítása
- Érdemes előre meghatározni a függő változót, majd ehhez keresni magyarázóváltozókat
- Legyen magyarázható a kapcsolat
- Modell építhető olyan összefüggésekre is, ami nincs egymással kapcsolatban
  - Fagylalt és bűnözés példája

Brexit és a kergemarhakór térképe





https://www.indy100.com/article/the-truth-about-the-map-comparing-brexit-and-mad-cow-disease--Zkm7IkclSb

Brexit és a kergemarhakór térképe

- Számtalan elmélet felállítható
  - Az EU nem megfelelően kezelte a kergemarhakórt, emiatt az ott élők a kilépésre szavaztak
  - A nagyvárosokban nem tartanak marhát így ott nem pusztított -> maradásra szavaztak
  - A kergemarha kór miatt az emberek is megkergültek
- Valóság: a két kép ugyanaz, egyszerűen csak elvették a színeket, és kicserélték a jelmagyarázatot

Tartalom

- Empirikusan tesztelhető modell
- Adatgyűjtés
- Modell becslése
- Előrejelzés, döntéselőkészítés
- Órai példa bemutatása

Adatok

- Keresztmetszeti adatok
  - Aggregált adatok (pld.: egy ország adatai egy időpontban)
  - Egyéni adatok (mikroadatok)
- Idősor adatok
  - Adatok időbeni alakulása
- Panel adatok
  - Keresztmetszei és időbeni adatok

Hasznos adatbázisok

- Statisztikai hivatalok
  - ksh.hu
  - https://ec.europa.eu/eurostat/web/main
- Nemzetközi szervezetek
  - https://data.worldbank.org/
- Népszámlálási adatok
  - http://www.ksh.hu/apps/hntr.kereses
  - http://webmap.lechnerkozpont.hu/webappbuilder/apps/foldgomb1708/
- Forgalmi adatok
  - http://kira.gov.hu/kira/
  - https://internet.kozut.hu/kozerdeku-adatok/orszagos-kozuti-adatbank/forgalomszamlalas/
  - https://www.ssc.sk/sk/cinnosti/rozvoj-cestnej-siete/dopravne-inzinierstvo/celostatne-scitanie-dopravy-v-roku-2015.ssc
  - https://www.gddkia.gov.pl/pl/2551/GPR-2015
- Egyéb
  - http://www.kti.hu/trendek/
  - http://www.hidadatok.hu/
  - http://online.winmenetrend.hu/

Dummy változók

- Amennyiben kategorikus változókat kívánunk figyelembe venni, a dummy változók alkalmazása a célszerű
- Ekkor a megfigyelt sokaságot n darab csoportra osztjuk, majd létrehozunk $n-1$ darab dummy változót D_i $\forall i \in 1..n-1$

$$D_i = \begin{cases} d & i.\text{ csoport esetén} \\ 0 & \text{egyébként} \end{cases}$$

Dummy változók

- Általában kétféle megközelítés
  - A csoportokra felírt függvények csak tengelymetszetükben térnek el

$$y = \begin{cases} \alpha_1 + \beta x + u & \text{az 1. csoportra} \\ \alpha_2 + \beta x + u & \text{a 2. csoportra} \end{cases}$$

$$y = \alpha_1 + (\alpha_2 - \alpha_1)D + \beta x + u$$

$$D = \begin{cases} 1 & \text{a 2. csoport esetén} \\ 0 & \text{az 1. csoport esetén} \end{cases}$$

    - Ebben az esetben vagy csak $n - 1$ dummy változót szabad alkalmazni, vagy pedig nem szabad konstans tagot alkalmazni
    - Ha mindkettőt alkalmazzuk, akkor teljes multikollinearitás lép fel

Dummy változók

- Általában kétféle megközelítés
  - A meredekségben térnek el

$$y = \begin{cases} \alpha + \beta_1 x + u & \text{az 1. csoport esetén} \\ \alpha + \beta_2 x + u & \text{a 2. csoport esetén} \end{cases}$$

$$y = \alpha + \beta_1 x + (\beta_2 - \beta_1) D + u$$

$$D = \begin{cases} x & \text{a 2. csoport esetén} \\ 0 & \text{az 1. csoport esetén} \end{cases}$$

    - Ebben az esetben is $n - 1$ dummy változót szabad alkalmazni, ellenkező esetben a $\beta_1 x$ tagot kell elhagyni

- Természetesen a két megközelítés egyszerre is alkalmazható több dummy változó alkalmazásával

Tartalom


Empirikusan tesztelhető modell
Adatgyűjtés
Modell becslése
Előrejelzés, döntéselőkészítés
Órai példa bemutatása

Módszertan

- Leggyakoribb módszertan a regressziós modellezés
- A függő változót ($\boldsymbol{y}$) becsüljük a független változók segítségével ($\boldsymbol{X}$)

$$\boldsymbol{y = a + bX}$$

$$\boldsymbol{y = a + bX + u}$$

- Becslés
  - Legkisebb négyzetek módszere (OLS)
  - Maximum likelihood függvény (ML)

OLS (kétváltozós eset)

- Hibatagok négyzetösszegének minimalizálása

$$\min \sum_{i=1}^{n} (y_i - a - b_i)^2$$

- Együtthatók

$$a = \bar{y} - b\bar{x}$$

$$b = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Maximum likelihood (ML) módszer

- Tfh adott egy $X(X_1, X_2, \dots, X_n)$ valószínűségi változó, adott ϑ paraméterű eloszlással
- Cél: ϑ becslése $x_1, x_2, \dots, x_n$ segítségével, ahol $x_1, x_2, \dots, x_n$ az X megfigyelt értékei
- Ekkor a likelihood függvény:

$$L(\vartheta) = \boldsymbol{P}(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) =$$

$$= f(x_1; \vartheta) f(x_2; \vartheta) \dots f(x_n; \vartheta) = \prod_{i=1}^{n} f(x_i; \vartheta)$$

- Több ϑ paraméter $(\vartheta_1, \vartheta_2, \dots, \vartheta_m)$ esetén

$$L(\vartheta_1, \vartheta_2, \dots, \vartheta_m) = \prod_{i=1}^{n} f(x_i; \vartheta_1, \vartheta_2, \dots, \vartheta_m)$$

ML normális eloszlás esetén

- ! $X(X_1, X_2, \dots, X_n) \sim \mathcal{N}(\vartheta; \sigma^2)$
- A normális eloszlás sűrűségfüggvénye (f)

$$f(x_i; \vartheta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\vartheta)^2}{2\sigma^2}}$$

$$L(\vartheta) = \prod_{i=1}^{n} f(x_i; \vartheta) = \sigma^{-n}(2\pi)^{-\frac{n}{2}} exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \vartheta)^2\right]$$

- Mivel a likelihood függvényt maximalizálni kell, vehetjük a logaritmusát, ugyanis a logaritmus függvény monoton

$$\ln L(\vartheta) = -n \ln \sigma - \frac{n}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \vartheta)^2$$

ML normális eloszlás esetén

- Egy függvény maximumpontját a deriváltfüggvény 0-vá tételével lehet meghatározni

$$\frac{d}{d\vartheta}\ln L(\vartheta) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \vartheta) = 0$$

$$\vartheta = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

- Tehát a ϑ paraméter a mintaátlaggal becsülhető

ML lineáris regresszió esetén

- ! $y_i = a + bx_i + u_i$, ahol $u_i \sim \mathcal{N}(0; \sigma^2)$
- Ebben az esetben a hibatag sűrűségfüggvénye

$$f(y_i; a, b, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i - a - bx_i - 0)^2}{2\sigma^2}}$$

$$L(a, b, \sigma) = \sigma^{-n}(2\pi)^{-\frac{n}{2}} exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - a - bx_i)^2\right]$$

$$\ln L(a, b, \sigma) = -n \ln \sigma - \frac{n}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - a - bx_i)^2$$

$$\frac{\partial}{\partial a}\ln L = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - a - bx_i) = 0$$

$$\frac{\partial}{\partial b}\ln L = \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i(y_i - a - bx_i) = 0$$

$$\frac{\partial}{\partial(\sigma^2)}\ln L = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - a - bx_i)^2 = 0$$

Gauss-Markov tétel

- $E(u_i) = 0$
- u_i, u_j függetlenek $\forall i \neq j$
- x_i, u_j függetlenek $\forall i, j$
- Nincs tökéletes kollinearitás
- $Var(u_i) = \sigma^2, \forall i$
- u_i normális eloszlású

- Első öt feltétel a Gauss–Markov-feltétel
- A hat összeségében a klasszikus lineáris modell feltételei
- Gauss–Markov-tétel: az együtthatókra vonatkozó torzítatlan lineáris becslések közül ez a legkisebb szórásnégyzetű, azaz hatásos

Multikollinearitás

- A multikollinearitás kérdésköre rendkívül szerteágazó és bonyolult témakör
- Általánosan elfogadott módszer a következő ökölszabály alkalmazása
  - Ha két változó között a korreláció négyzete nagyobb, mint 0,7, akkor érdemes tartózkodni egy azon modellben való alkalmazásuktól

Multikollinearitás

- Mi a teendő ha az r_{ij}^2 magas?
  - Kézenfekvő az adott változó elhagyása
    - Ez azonban megjelenik a hibatagban így nem a legjobb megoldás
  - Ridge regresszió
  - Főkomponens elemzés
  - Változók összevonása

Korrelációszámítás

- Korreláció: olyan mérőszám, amely a mennyiségi ismérvek közötti kapcsolat szorosságát (és irányát) mutatja meg
- Pearson-féle korrelációs együttható

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{SQ_x SQ_y}}$$

  - r: korrelációs együttható
  - x, y: a vizsgált változók ($\bar{x}$ illetve $\bar{y}$ az adott változókra vonatkoztatott átlag)
  - SQ_x, SQ_y: a vizsgált változókra vonatkoztatott eltérésnégyzet-összeg

Korrelációszámítás

- Természetesen a korreláció értéke is valószínűségi változó, így szignifikancia szintet is számítani kell hozzá
- Ezt t-próbával lehet megtenni, próbastatisztika értéke ($DF = n - 2$) mellett

$$t = \frac{r}{\sqrt{1 - r^2}} \sqrt{n - 2}$$

- A szignifikancia szint ismerete segíthet eldönteni, hogy a változók maradhatnak-e a modellben

Modellek eredménye

- A statisztikai programok három csoportra bontva szokták megadni az eredményeket
  - Regressziós statisztika
  - Varianciaanalízis
  - Koefficiensek

Regressziós statisztika

- R^2: Pearson féle korrelációs együttható négyzete a függőváltozó ($\boldsymbol{y}$), valamint a modell becsült értékei ($\tilde{\boldsymbol{y}} = \boldsymbol{a} + \boldsymbol{bX}$) között
- Korrigált R^2
  - A magyarázóváltozók számának növelésével az R^2 értéke nő, ami nem feltétlenül jelenti a modell javulását
  - Tartalmaz egy büntető paramétert (amely a magyarázóváltozók száma – q)

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

Varianciaanalízis

- ANOVA (ANalysis Of VAriance) tábla alkalmazása
- Ha az F-próba nem szignifikáns, akkor a modell csak a véletlentől függ

Variancia forrása	Eltérésnégyzet-összeg (SS)	Szabadság-fok (DF)	Átlagos eltérés-négyzetösszeg (MS)	F-érték	Szignifikancia
Regresszió	$ESS = \sum(\tilde{y}_i - \bar{y})^2$	q	$MSE = \frac{ESS}{DF}$	$F = \frac{MSE}{MSR}$	Szignifikancia-szint az F-próba táblázata alapján
Maradék	$RSS = \sum(y_i - \tilde{y}_i)^2$	$n - q - 1$	$MSR = \frac{RSS}{DF}$		
Összesen	$TSS = \sum(y_i - \bar{y})^2$	$n - 1$			

F-próba

- $!\, X(X_1, X_2, \dots, X_n) \sim \mathcal{N}(\mu_1; \vartheta_1)$
- $!\, Y(X_1, X_2, \dots, X_m) \sim \mathcal{N}(\mu_2; \vartheta_2)$
- $\vartheta_1, \vartheta_2 > 0$, de ismeretlen
- $H_0: \vartheta_1 = \vartheta_2$
- $H_1: \vartheta_1 \neq \vartheta_2$
- A próbastatisztika

$$f = \frac{s_{X,n}^{*2}}{s_{Y,m}^{*2}} \sim F_{n-1,m-1}$$

  - Ahol $s_{X,n}^{*2}$ és $s_{Y,m}^{*2}$ a valószínűségi változók empirikus szórásnégyzetei

Együtthatók

- Vegyük a következő modellt

$$y_i = a + \sum_{j=1}^{q} b_j x_{ij} + u_i \; \forall i \in 1..n$$

- Vezessük be a következő jelöléseket
  - $S_{jj} = \sum_i x_{ij}^2 - n\bar{x}_j^2 \; \forall j \in 1..q$
  - $S_{jl} = \sum_i x_{ij} x_{li} - n\bar{x}_j \bar{x}_l \; \forall j, l \in 1..q$
  - $S_{jy} = \sum_i x_{ij} y_i - n\bar{x}_j \bar{y} \; \forall j \in 1..q$
  - $S_{yy} = \sum_i y_i^2 - n\bar{y}^2$

Együtthatók

- A normálegyenletek általánosítása után a következő egyenleteket kapjuk

$$a = \bar{y} - \sum_j b_j x_j$$

$$S_{ly} = \sum_j b_j S_{lj} \quad \forall l \in 1..q$$

Együtthatók

- A bevezetett jelölésekkel a következő paraméterek is felírhatók

$$RSS = S_{yy} - \sum_j b_j S_{jy} = S_{yy}(1 - R^2)$$

$$R^2 = \frac{\sum_j b_j S_{jy}}{S_{yy}}$$

Együtthatók

- Standard hibák meghatározása

$$var(b_j) = \frac{\sigma^2}{RSS_j} \; \forall j = 1..q$$

  - RSS_j: a reziduális négyzetösszeg abban az esetben ha x_j a függőváltozó, a többi a magyarázóváltozó
- A σ^2 becsülhető

$$\sigma^2 = \frac{RSS}{n - q - 1}$$

- A standard hiba $SE(b_j)$ a variancia négyzetgyöke

Együtthatók

- Az együtthatók táblázata az együtthatókon valamint a standard hibákon felül a t-próba adatait tartalmazza
- Az együtthatók t-eloszlást követnek $DF = n - q - 1$ szabadságfokkal
- A próbastatisztika

$$t_j = \frac{b_j}{SE(b_j)} \quad \forall j = 1..\mathrm{q}$$

Gauss-Markov feltételek ellenőrzése

- A becslésünk akkor BLUE (Best Linear Unbiased Estimate – Legjobb Lineáris Torzítatlan Becslés), ha a Gauss-Markov feltételek teljesülnek
- További elemzések:
  - Hibatagok normalitásának vizsgálata
  - Multikollinearitás
  - Heteroszkedaszticitás
  - Autokorreláció (Erről a jövő órán lesz szó)

Hibatagok normalitásának ellenőrzése

- A hibatagok normalitása nem tartozik a Gauss-Markov feltételek közé, csak a klasszikus lineáris modell feltételeibe
- Azonban, ha a hibatagok normális eloszlást követnek
  - $E(u_i) = \bar{u}$
  - $E(u_i) = 0$ feltétel egy egyszerű t-próbával ellenőrizhető
- Hibatagok normalitásának ellenőrzése
  - Kolmogorov-Smirnov próba (2000-es mintanagyság felett)
  - Shapiro-Wilk próba (2000-es mintanagyság alatt)
  - χ^2 illeszkedésvizsgálat
- Amennyiben a hibatagok nem normális eloszlást követnek
  - y logaritmizálása
  - y^{λ} becslése, alkalmas λ kitevő megválasztásával (-2, -1, -0,5, 0,5, 2)

Kolmogorov-Smirnov próba

- ! $X(X_1, X_2, \dots, X_n)$ egy statisztikai minta
- ! $F(x)$ a minta eloszlásfüggvénye
- ! $F_0(x)$ egy elméleti eloszlás eloszlásfüggvénye
- $H_0: F(x) = F_0(x)\ \forall x \in \mathbb{R}$
- $H_1: \exists x: F(x) \neq F_0(x)$
- ! $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i<x\}}$ a minta empirikus eloszlásfüggvénye
- ! $D_n = \sqrt{n}\max_{x\in\mathbb{R}}|F_n(x) - F_0(x)|$
- Ha $D_n < K_\varepsilon \Rightarrow H_0$ elfogadható

ε	K_ε
0,9	1,23
0,95	1,36
0,99	1,63
0,999	1,96

χ^2 próba illeszkedésvizsgálathoz

- $\chi^2 = \frac{(O-E)^2}{E}$
  - O: megfigyelt értékek (observed)
  - E: várt értékek (expected)
- A χ^2 próba elengedhetetlen része az osztályközökbe való sorolás
  - Osztályközök száma: $k = 2\sqrt[5]{N^2}$
  - Az osztályközök elemeinek a száma (f_i) legyen egyenlő
- $\chi^2 = \frac{(O-E)^2}{E} = \sum_{i=1}^{k} \frac{(f_i - Np_i)^2}{Np_i}$
  - p_i: adott átlag és szórás értékekre meghatározott normális eloszlás esetén mekkora a valószínűsége az osztályközbe való esésnek
- $DF = k - 3$

Multikollinearitás

- A multikollinearitás azonosítása, veszélyességének megállapítása és elhárítása igen komoly és összetett probléma
- Az előzőekben megadott ökölszabály csupán vészjelzőnek alkalmas
- Szakirodalomban gyakorta alkalmazott a VIF mutató, amely azonban szintén vészjelzőnek alkalmas

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

  - Ahol R_i^2 az i. magyarázóváltozó és a többi közti Pearson-féle korrelációs együttható
- Értékelése
  - $1 \leq VIF_j \leq 2$ – gyenge multikollinearitás
  - $2 \leq VIF_j \leq 5$ – zavaró multikollinearitás
  - $5 \leq VIF_j$ – nagyon erős, káros multikollinearitás

Heteroszkedaszticitás

- Heteroszkedaszticitásról akkor beszélhetünk, ha a hibatagok szórása nem állandó
- Heteroszkedaszticitás esetén a becslés torzítatlan, azonban nem hatásos
- Egyik lehetséges megoldás a változók logaritmizálása ($y = a + bx$ helyett $\ln y = a + b \ln x$ alkalmazása)
- Heteroszkedaszticitás azonosítása
  - White teszt
  - Breusch-Pagen teszt

White teszt

- Alapfelvetés: $var(u_i) = \sigma^2 f(z_i)$
- Lépései (két változó esetén)
  1. $y_i = a + b_1 x_{1i} + b_2 x_{2i} + u_i$
  2. A hibatagra a következő modell felállítása
     $u_i^2 = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i}^2 + \beta_4 x_{2i}^2 + \beta_5 x_{1i} x_{2i}$
  3. A 2. lépés modelljéből az R^2 érték meghatározása
  4. A következő állítás igaz
     $$nR^2 \sim \chi^2_{DF}$$
     ahol DF a 2. lépésben felállított modell magyarázóváltozóinak száma (q)
  5. Amennyiben a próbastatisztika értéke nagyobb, mint a kritikus érték adott DF és szignifikancia szint mellett a homoszkedaszticitási nullhipotézist el kell utasítani

Tartalom


Empirikusan tesztelhető modell
Adatgyűjtés
Modell becslése
Előrejelzés, döntéselőkészítés
Órai példa bemutatása

Modell értékelése

- R^2
  - Minél nagyobb annál jobban magyarázzák a függő változót a függetlenek
  - Százalékos értékként értelmezhető
- F-érték
  - Amennyiben az F-próba nem szignifikáns, akkor nem zárható ki, hogy a modell csak a véletlentől függ
- t-érték
  - Amennyiben a t-próba nem szignifikáns, akkor nem zárható ki, hogy az adott változó nem befolyásolja a modellt

Előrejelzés, döntéselőkészítés

- Amennyiben az R^2 alacsony, a modell előrejelzésre nem alkalmas
  - Ettől még döntéselőkészítési célokat szolgálhat
  - Példa: a magyarázóváltozók szignifikáns hatása
- Döntéstámogatás két esetben értelmezhető
  - Adott modell alkalmas előrejelzésre – R^2 érték magas
  - Adott magyarázóváltozó szignifikáns hatást mutat – a t-próbához tartozó szignifikanciaszint alacsony

BETA mutató

- A modellben szereplő együtthatók nem egyenlő mértékben befolyásolják a modellt
- Annak eldöntésére, hogy melyik változó mennyire fontos a BETA mutató alkalmazható

$$BETA_j = b_j \frac{s_j}{s_y}$$

- Ahol:
  - b_j: a j. magyarázóváltozó együtthatója
  - s_y: a függő változó szórása
  - s_j: a j. magyarázóváltozó szórása
- A BETA mutató minél nagyobb abszolút értékben, annál fontosabb szerepet tölt be az adott változó a modellben
- Beépítve csak az SPSS számolja

Tartalom

- Empirikusan tesztelhető modell
- Adatgyűjtés
- Modell becslése
- Előrejelzés, döntéselőkészítés
- Órai példa bemutatása

Órai példa

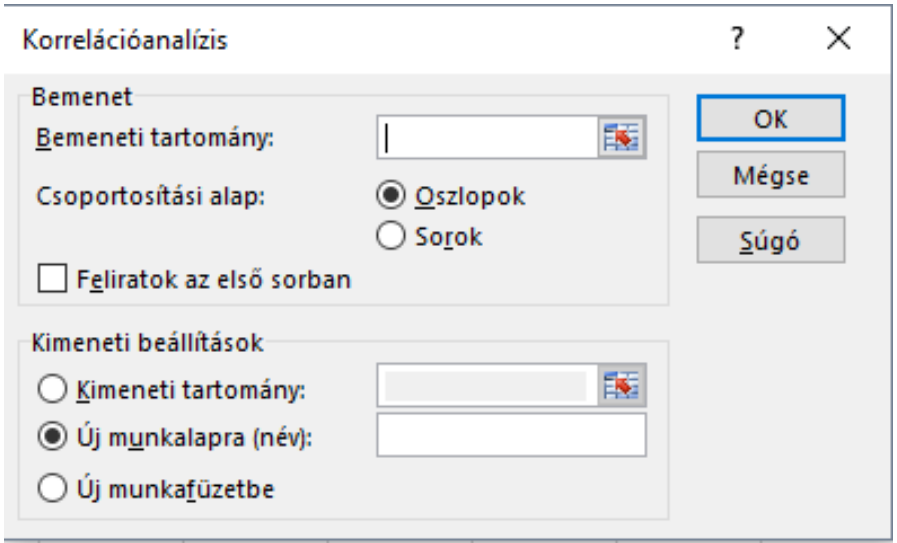
- Településeken a helyi közösségi közlekedési szolgáltatás vizsgálata
- Lehetséges függőváltozók
  - Járatszám: hány helyi járat működik az adott településen
  - Indulásszám: egy átlagos hétköznapon hány járatpár közlekedik

Lehetséges magyarázóváltozók

- Lak(2018): az adott település lakosságszáma 2018-ban
- Terület: az adott település területe hektárban
- Népsűrűség: a fenti két érték hányadosa
- Rang: az adott település milyen NUTS terület központja
- Pest megye: az adott település Pest megyében van-e
- Vasút: van-e a településen vasútállomás
- Állomás táv: ha van, akkor milyen messze
- Mot.fok: motorizációs fok
- Átlag járásjövedelem: átlagos személyijövedelemadó-alapot képző jövedelem egy állandó lakosra
- Bp táv: Budapesttől való távolság

Korrelációszámítás – Excel

- Adatok->Adatelemzés->Korrelációanalízis
- Bemeneti tartomány: Itt lehet megadni azokat az oszlopokat, vagy sorokat amiket figyelembe vesz (csak összefüggő tartomány)
- Feliratok az első sorban: ha a táblázatban vannak feliratok, és ezeket is kijelöljük, az eredménytáblázatban megjelennek a címkék

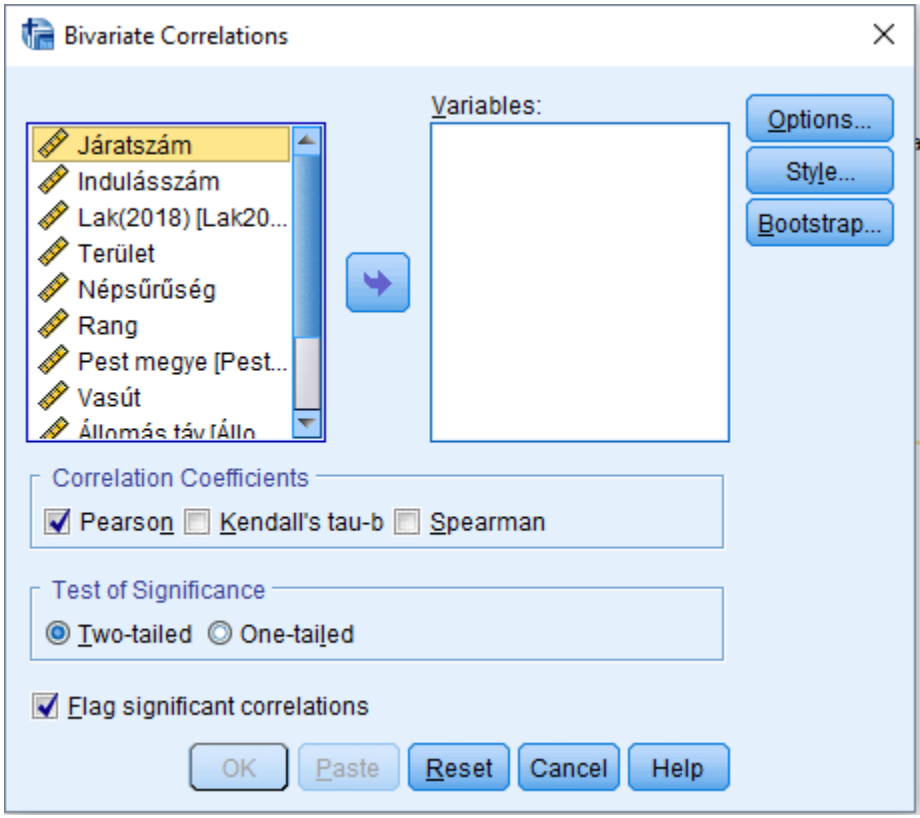


Korrelációszámítás – Excel

	Járatszám	*Indulásszám*	*Lak(2018)*	*Terület*	*Népsűrűség*	*Rang*	*Pest megye*	*Vasút*	*Állomás táv*	*Mot.fok*	*Átlag járásjövedelem*	*Bp táv*
Járatszám	1											
Indulásszám	0,86580307	1										
Lak(2018)	0,90675341	0,89295612	1									
Terület	0,42941387	0,37802621	0,58940388	1								
Népsűrűség	0,45888414	0,36724062	0,52508103	0,12674066	1							
Rang	-0,70740803	-0,60386412	-0,77856919	-0,5638988	-0,41587977	1						
Pest megye	0,04213186	0,03018776	0,1184878	0,03413045	0,39839114	-0,03012181	1					
Vasút	0,00133955	-0,00044479	0,00171349	0,04277775	-0,00775839	0,00694198	-0,00774637	1				
Állomás táv	0,00107034	-0,00053436	0,00119229	0,04265471	-0,00802463	0,00676254	-0,00754615	0,97415273	1			
Mot.fok	0,0122775	0,00755339	0,00798843	-0,04633788	0,0413416	-0,0065465	0,12049811	-0,00362942	-0,00376384	1		
Átlag járásjövedelem	0,10127733	0,08538141	0,12323146	-0,02427648	0,27292697	-0,02971444	0,26183326	0,01065994	0,00043059	0,20977233	1	
Bp táv	-0,07673078	-0,04962365	-0,13810545	-0,13355269	-0,31553526	0,06661873	-0,47131182	-0,00906215	-0,00382893	-0,11565056	-0,49926345	1

Korrelációszámítás – SPSS

- Analyze->Correlate->Bivariate
- Bal oldali oszlop: lehetséges változók
- Jobb oldali oszlop (Variables): kiválasztott változók
- Korreláció típusai
  - Pearson
  - Kendall's tau-b: előjelkorreláció
  - Spearman: rangkorreláció
- Flag significant correlations: szignifikanciaszint számítása



Correlations

		Járatszám	Indulásszám	Lak(2018)	Terület	Népsűrűség	Rang	Pest megye	Vasút	Állomás táv	Mot.fok	Átlag járásjövedelem	Bp táv
Járatszám	Pearson Correlation	1	,866**	,907**	,429**	,459**	-,707**	,042*	0,001	0,001	0,012	,101**	-,077**
	Sig. (2-tailed)		0,000	0,000	0,000	0,000	0,000	0,018	0,940	0,952	0,491	0,000	0,000
	N	3154	3154	3154	3154	3154	3154	3154	3154	3154	3154	3154	3154
Indulásszám	Pearson Correlation	,866**	1	,893**	,378**	,367**	-,604**	0,030	0,000	-0,001	0,008	,085**	-,050**
	Sig. (2-tailed)	0,000		0,000	0,000	0,000	0,000	0,090	0,980	0,976	0,672	0,000	0,005
	N	3154	3154	3154	3154	3154	3154	3154	3154	3154	3154	3154	3154
Lak(2018)	Pearson Correlation	,907**	,893**	1	,589**	,525**	-,779**	,118**	0,002	0,001	0,008	,123**	-,138**
	Sig. (2-tailed)	0,000	0,000		0,000	0,000	0,000	0,000	0,923	0,947	0,654	0,000	0,000
	N	3154	3154	3154	3154	3154	3154	3154	3154	3154	3154	3154	3154
Terület	Pearson Correlation	,429**	,378**	,589**	1	,127**	-,564**	0,034	,043*	,043*	-,046**	-0,024	-,134**
	Sig. (2-tailed)	0,000	0,000	0,000		0,000	0,000	0,055	0,016	0,017	0,009	0,173	0,000
	N	3154	3154	3154	3154	3154	3154	3154	3154	3154	3154	3154	3154
Népsűrűség	Pearson Correlation	,459**	,367**	,525**	,127**	1	-,416**	,398**	-0,008	-0,008	,041*	,273**	-,316**
	Sig. (2-tailed)	0,000	0,000	0,000	0,000		0,000	0,000	0,663	0,652	0,020	0,000	0,000
	N	3154	3154	3154	3154	3154	3154	3154	3154	3154	3154	3154	3154
Rang	Pearson Correlation	-,707**	-,604**	-,779**	-,564**	-,416**	1	-0,030	0,007	0,007	-0,007	-0,030	,067**
	Sig. (2-tailed)	0,000	0,000	0,000	0,000	0,000		0,091	0,697	0,704	0,713	0,095	0,000
	N	3154	3154	3154	3154	3154	3154	3154	3154	3154	3154	3154	3154
Pest megye	Pearson Correlation	,042*	0,030	,118**	0,034	,398**	-0,030	1	-0,008	-0,008	,120**	,262**	-,471**
	Sig. (2-tailed)	0,018	0,090	0,000	0,055	0,000	0,091		0,664	0,672	0,000	0,000	0,000
	N	3154	3154	3154	3154	3154	3154	3154	3154	3154	3154	3154	3154
Vasút	Pearson Correlation	0,001	0,000	0,002	,043*	-0,008	0,007	-0,008	1	,974**	-0,004	0,011	-0,009
	Sig. (2-tailed)	0,940	0,980	0,923	0,016	0,663	0,697	0,664		0,000	0,839	0,550	0,611
	N	3154	3154	3154	3154	3154	3154	3154	3154	3154	3154	3154	3154
Állomás táv	Pearson Correlation	0,001	-0,001	0,001	,043*	-0,008	0,007	-0,008	,974**	1	-0,004	0,000	-0,004
	Sig. (2-tailed)	0,952	0,976	0,947	0,017	0,652	0,704	0,672	0,000		0,833	0,981	0,830
	N	3154	3154	3154	3154	3154	3154	3154	3154	3154	3154	3154	3154
Mot.fok	Pearson Correlation	0,012	0,008	0,008	-,046**	,041*	-0,007	,120**	-0,004	-0,004	1	,210**	-,116**
	Sig. (2-tailed)	0,491	0,672	0,654	0,009	0,020	0,713	0,000	0,839	0,833		0,000	0,000
	N	3154	3154	3154	3154	3154	3154	3154	3154	3154	3154	3154	3154
Átlag járásjövedelem	Pearson Correlation	,101**	,085**	,123**	-0,024	,273**	-0,030	,262**	0,011	0,000	,210**	1	-,499**
	Sig. (2-tailed)	0,000	0,000	0,000	0,173	0,000	0,095	0,000	0,550	0,981	0,000		0,000
	N	3154	3154	3154	3154	3154	3154	3154	3154	3154	3154	3154	3154
Bp táv	Pearson Correlation	-,077**	-,050**	-,138**	-,134**	-,316**	,067**	-,471**	-0,009	-0,004	-,116**	-,499**	1
	Sig. (2-tailed)	0,000	0,005	0,000	0,000	0,000	0,000	0,000	0,611	0,830	0,000	0,000	
	N	3154	3154	3154	3154	3154	3154	3154	3154	3154	3154	3154	3154

**. Correlation is significant at the 0.01 level (2-tailed).

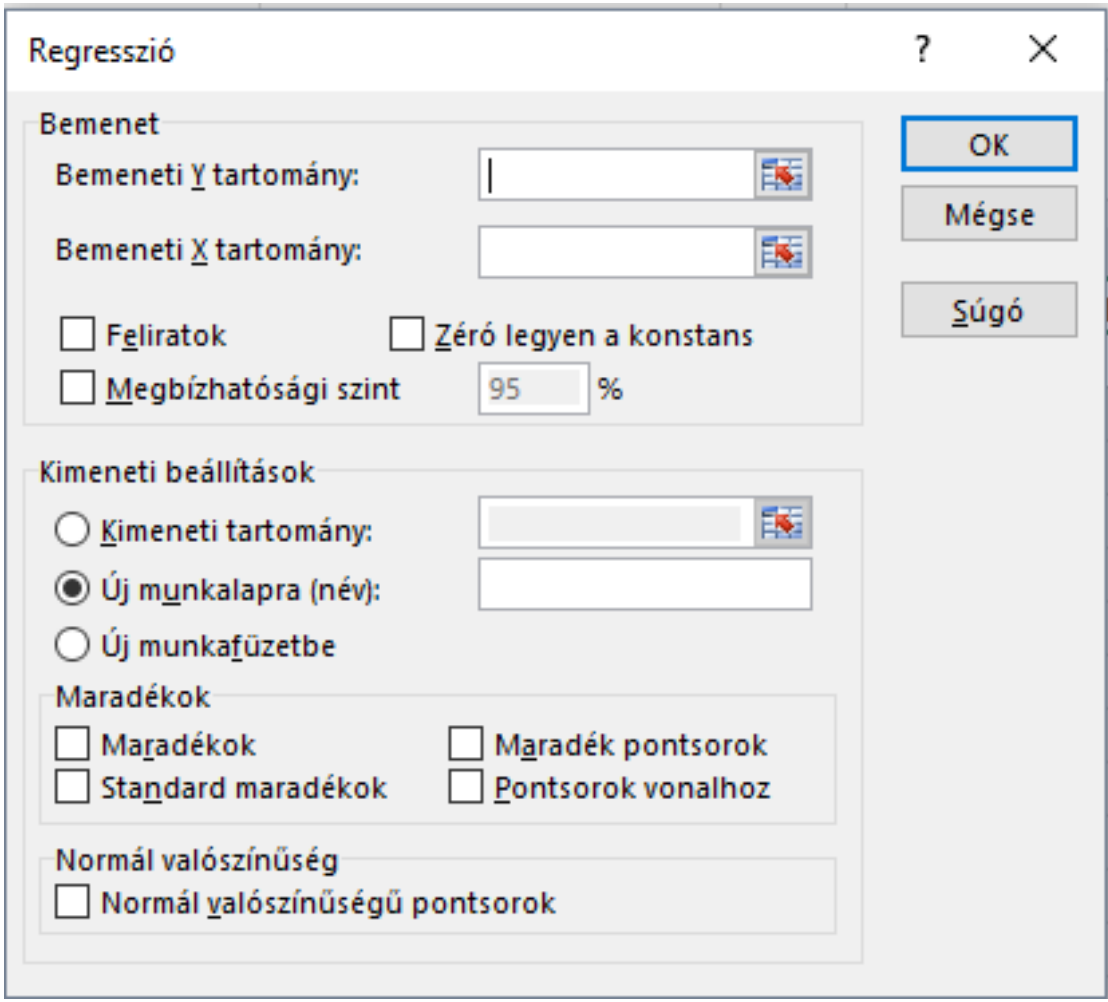
*. Correlation is significant at the 0.05 level (2-tailed).

Modellállítás

- Három magyarázóváltozós modell
  - Lakosságszám
  - Motorizációs fok
  - Rang
  - Függő változó: indulásszám természetes alapú logaritmusa
- Ok
  - Az Excel 16 magyarázóváltozót tud maximum kezelni
  - Így a White-teszt is bemutatható

Regressziószámítás – Excel

- Adatok->Adatelemzés ->Regresszió
- Bemeneti Y tartomány: függő változó
- Bemeneti X tartomány: független változók
- Maradékok
  - Maradékok: maradéktagok
  - Standard maradékok: maradéktagok standardizálva



Regressziószámítás – Excel

- Látható, hogy az R^2 érték alapján a magyarázóváltozók a függőváltozók ~53% magyarázzák

Regressziós statisztika	
r értéke	0,731586006
r-négyzet	0,535218084
Korrigált r-négyzet	0,527840593
Standard hiba	1,284919656
Megfigyelések	193

Regressziószámítás – Excel

- Az F-próba szignifikanciaszintje alapján
  - Elutasítható a H_0 hipotézis, hogy a modell csak a véletlentől függne

	df	*SS*	*MS*	*F*	*F szignifikanciája*
Regresszió	3	359,3315133	119,7771711	72,54744237	2,90514E-31
Maradék	189	312,0425007	1,651018522		
Összesen	192	671,374014			

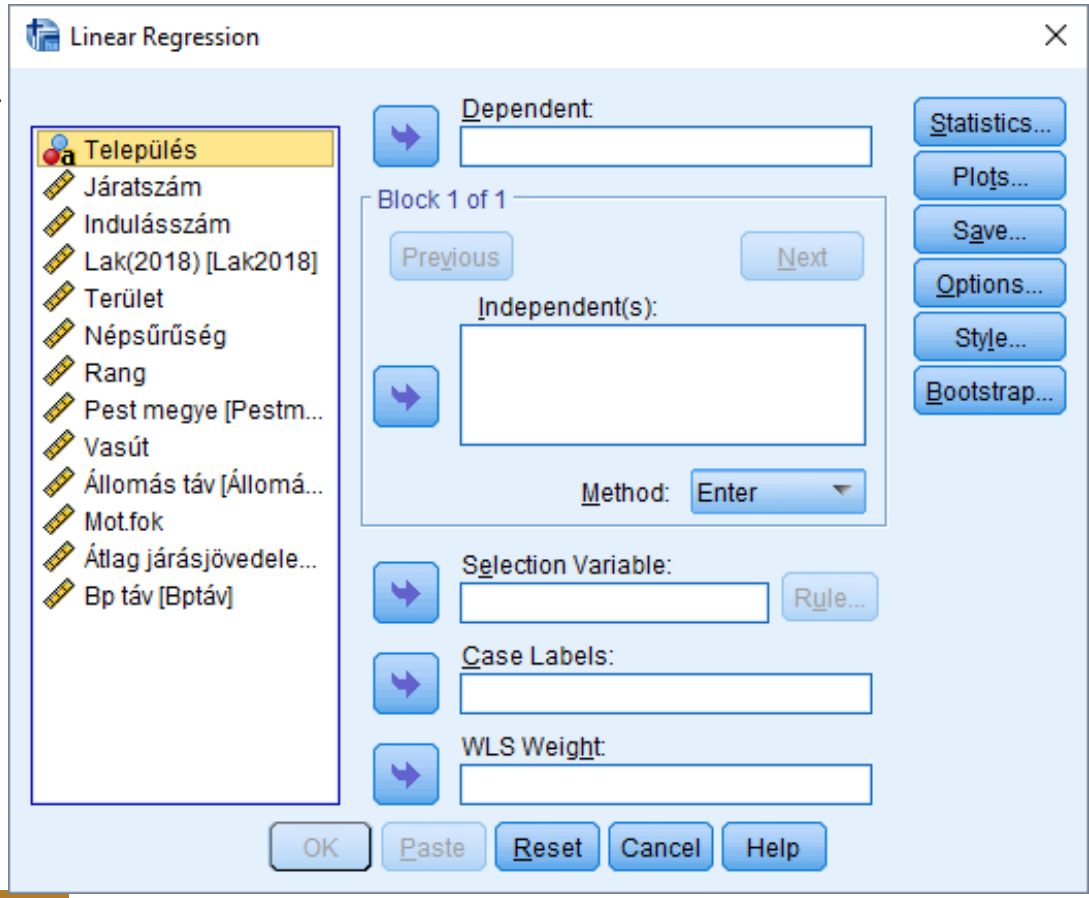
Regressziószámítás – Excel

- A t-próba szignifikanciaszintjei alapján a népsűrűség és a rang esetében elvethető a H_0 hipotézis, miszerint az együttható 0 lenne
- Ezt azonban a Motorizációs fokról nem lehet elmondani
- Az együtthatók előjele megfelelőnek tekinthető

	Koefficiensek	*Standard hiba*	*t érték*	*p-érték*	*Alsó 95%*	*Felső 95%*
Tengelymetszet	2,915803846	1,147702835	2,540556454	0,011871619	0,651850881	5,179756811
Lakosság	3,40568E-05	4,99861E-06	6,813265187	1,23945E-10	2,41966E-05	4,39171E-05
Mot.fok	0,004456264	0,001641178	2,715284353	0,007235517	0,001218885	0,007693643
Rang	-0,548067462	0,205535945	-2,666528534	0,008328203	-0,953506656	-0,142628268

Regressziószámítás – SPSS

- Analyze->Regression->Linear…
- Bal oldalt találhatók a lehetséges változók
- Jobb oldalt felül lehet megadni a függő változót
- Ez alatt pedig a függetleneket



White-teszt – Excel

- A jobb oldali táblázatból kiolvasható mind az n, mind pedig az R^2
- $nR^2 = 39{,}421$
- $\chi^2_{krit_{DF=9,p=95\%}} = 16{,}92$
- $nR^2 > \chi^2_{krit_{DF=9,p=95\%}}$
- Heteroszkedaszticitás áll fenn

Regressziós statisztika	
r értéke	0,451947843
r-négyzet	0,204256853
Korrigált r-négyzet	0,165121944
Standard hiba	2,108672383
Megfigyelések	193

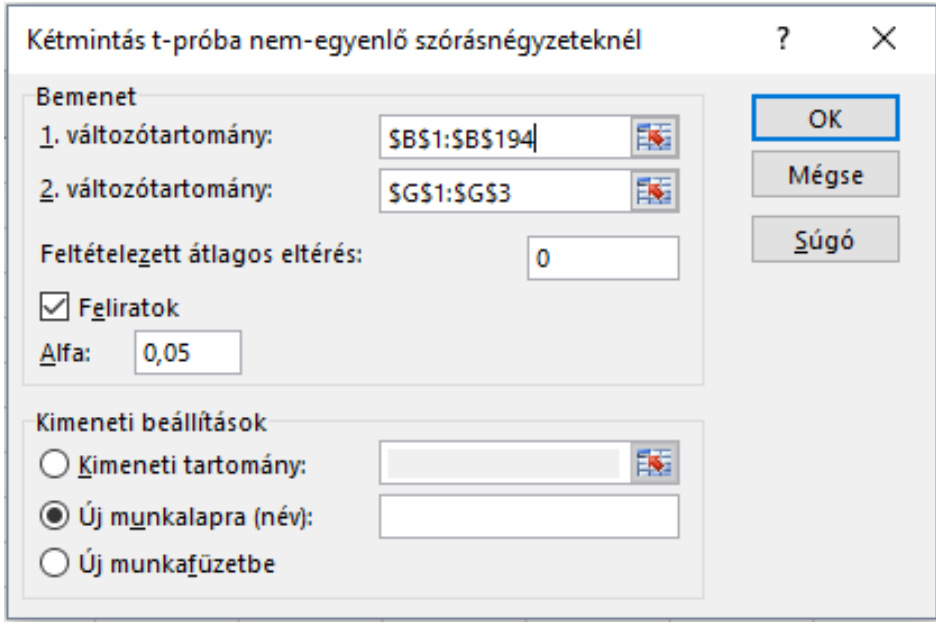
Normalitásvizsgálat

- A oszlop: sorszám
- B: hibatagok növekvő sorrendben
- C: minta eloszlásfüggvénye $(^{k}/_{n})$
- D: standard hibatagok $(^{(u-\bar{u})}/_{\sqrt{\sigma(u)}})$
- E: D alapján számított standard normális eloszlás (elméleti)
- F: különbség abszolútértéke
- $D_n = \sqrt{n} \max_{x \in \mathbb{R}} |F_n(x) - F_0(x)| = 0{,}9959$
- $D_n < K_\varepsilon$

	A	B	C	D	E	F
1			Fn(x)	Z-score	F0(x)	|Fn(x)-F0(x)|
2	1	-3,814861547	0,005181	-3,0002	0,001349	0,003832346
3	2	-3,568647747	0,010363	-2,80657	0,002504	0,007859072
4	3	-3,31027602	0,015544	-2,60337	0,004616	0,010928435
5	4	-3,027286122	0,020725	-2,38081	0,008637	0,012088145
6	5	-2,735675102	0,025907	-2,15147	0,015719	0,010187371
7	6	-2,637105966	0,031088	-2,07396	0,019042	0,012046345
8	7	-2,608584915	0,036269	-2,05152	0,020108	0,016161493
9	8	-2,45389187	0,041451	-1,92987	0,026812	0,014639068
10	9	-2,072382805	0,046632	-1,62983	0,051569	0,004936809
11	10	-2,049661835	0,051813	-1,61196	0,053485	0,001671958
12	11	-1,994686205	0,056995	-1,56872	0,058356	0,001361388
13	12	-1,983389856	0,062176	-1,55984	0,059399	0,002777239
14	13	-1,944857192	0,067358	-1,52954	0,063066	0,004291612
15	14	-1,901712854	0,072539	-1,4956	0,067378	0,005160467
16	15	-1,840632338	0,07772	-1,44757	0,073869	0,003851186
17	16	-1,807473719	0,082902	-1,42149	0,077587	0,005314357
18	17	-1,737193911	0,088083	-1,36622	0,085935	0,002147664
19	18	-1,720073128	0,093264	-1,35275	0,088067	0,005197065
20	19	-1,684049628	0,098446	-1,32442	0,092681	0,005764271
21	20	-1,664528304	0,103627	-1,30907	0,095255	0,008371671
22	21	-1,631385506	0,108808	-1,28301	0,099745	0,009063144

t-próba: $E(u) = 0$ tesztelésére

- Ha a hibatag eloszlása normális, akkor az $E(u) = 0$ ellenőrizhető egyszerű t-próbával
- Sajnos az Excelben nincsen beépített egymintás t-próba
- Felveszünk egy új oszlopot (legyen a neve Dummy) amiben két darab 0 szerepel
- Erre az oszlopra, illetve a hibatagra futtatunk egy kétmintás t-próbát, nem megegyező szórásnégyzettel



t-próba: $E(u) = 0$ tesztelésére

- Kétszélű próba
- $t < t_{krit}$
- $H_0: E(u) = 0$ elfogadható

Kétmintás t-próba nem-egyenlő szórásnégyzeteknél

	u	*Dummy*
Várható érték	3,41E-16	0
Variancia	1,625221	0
Megfigyelések	193	2
Feltételezett átlagos eltérés	0	
df	192	
t érték	3,71E-15	
P(T<=t) egyszélű	0,5	
t kritikus egyszélű	1,652829	
P(T<=t) kétszélű	1	
t kritikus kétszélű	1,972396	

Ellenőrző kérdések

- Mondja ki a Gauss-Markov tételt!
- Miért fontos az ökonometriában a korrelációanalízis?
- Miért jó mutatószám a korrigált R^2 érték?
- Miért alkalmazható a t-próba a paraméterbecslés jóságának ellenőrzésére?

KÖSZÖNÖM A MEGTISZTELŐ FIGYELMÜKET!

Dr. Sipos Tibor
Dr. Török Ádám
Szabó Zsombor

KUKG

BUDAPESTI MŰSZAKI ÉS GAZDASÁGTUDOMÁNYI EGYETEM
Közlekedésmérnöki és Járműmérnöki Kar