

# Hipotézisvizsgálat és intervallumbecslés lineáris modellben

Ferenci Tamás  
tamas.ferenci@medstat.hu

Utoljára frissítve: 2023. május 9.

## Tartalom

## Tartalomjegyzék

1	Alkalmazási feltételek	1
2	Egy paraméter	1
3	Modell egésze	2
4	Tetszőleges számú paraméter	3
5	Lineáris megkötés(ek)	5

## 1. Alkalmazási feltételek

### Emlékeztetőül

- A most következő eredmények csak akkor egzaktak, ha a hibanormalitás is fennáll
- Ám aszimptotikusak, így közelítőleg akkor is fennállnak, ha elég nagy a mintanagyság (minél nagyobb, annál inkább)

Ez a centrális határeloszlás-tétel miatt van így.

## 2. Egy paraméter

### Becsült regressziós koefficiensek mintavételi eloszlása

- A  $\hat{\beta}_i$  becsült regressziós koefficiens mintavételi ingadozását tehát a következő összefüggés írja le:

$$\frac{\hat{\beta}_i - \beta_i}{\text{se}(\hat{\beta}_i)} \sim \mathcal{N}(0, 1),$$

$$\text{ahol } \text{se}(\hat{\beta}_i) = \sqrt{\sigma^2 \left[ \left( \underline{\underline{X}}^T \underline{\underline{X}} \right)^{-1} \right]_{kk}}$$

- Sajnos ezzel a gyakorlatban nem sokra megyünk, mert  $\sigma^2$ -et általában nem ismerjük
- Helyettesítsük a jó tulajdonságú becslőjével,  $\hat{\sigma}^2$ -tel!
- Így persze már más lesz az eloszlás, de szerencsére meghatározható, hogy mi, és nem bonyolult:  $n - (k + 1)$  szabadságfokú  $t$ -eloszlás

### Változó relevanciája

Egy változót relevánsnak nevezünk, ha a sokasági paramétere nem nulla:  $\beta_i \neq 0$ .

### Hipotézisvizsgálat változó relevanciájára

Ez alapján már konstruálhatunk próbát változó relevanciájának vizsgálatára:

1.  $H_0 : \beta_i = 0$
2. Ekkor (azaz ha ez fennáll!) a  $t_{\text{emp},i} = \frac{\hat{\beta}_i}{\text{se}(\hat{\beta}_i)}$  kifejezés  $n - (k + 1)$  szabadságfokú  $t$ -eloszlást követ (nulleloszlás)
3. Számítsuk ki a konkrét  $t_{\text{emp},i}$ -t a mintánkból és döntsük el, hogy hihető-e, hogy  $t_{n-(k+1)}$ -ből származik

### Hipotézisvizsgálat változó relevanciájára

A hipotézisvizsgálat elvégzéséhez szükséges minden tudnivalót – a nullhipotézisen kívül – összefoglal tehát a következő kifejezés (a későbbiekben is ezt a sémát fogjuk használni hipotézisvizsgálatok megadására):

$$t_{\text{emp},i} = \frac{\hat{\beta}_i}{\text{se}(\hat{\beta}_i)} \stackrel{H_0}{\sim} t_{n-(k+1)}.$$

E próba precíz neve: változó relevanciájára irányuló (parciális)  $t$ -próba

### 3. Modell egésze

#### Modell egészének relevanciája

- A korábban látott  $t$ -próba azért volt „parciális”, mert egy változó irrelevanciáját vizsgálta
- Felmerül a kérdés, hogy definiálható-e a modell *egészének* irrelevanciája
- Igen, mégpedig úgy, hogy *valamennyi* magyarázó változó paramétere *együttesen is* irreleváns:

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$$

- (Természetesen a  $\beta_0$  nincs felsorolva!)
- Rövid jelölés arra, hogy  $\beta_1 = 0$  és  $\beta_2 = 0$  stb. és  $\beta_k = 0$  (semmilyen más eset jelölésére *ne* használjuk az egyenlőségláncot!)
- Figyelem: az „egyszerre nulla mindegyik” *több* mint, hogy „külön-külön nulla mindegyik”!

#### Modell egészének relevanciája

- A modell egészének irrelevanciájára magyarul azt jelenti, hogy a modell nem tér el lényegesen a nullmodelltől
- Implikálja, hogy minden magyarázó változó külön-külön is irreleváns (tartalmazza ezeket a hipotéziseket)  $\rightarrow$  előbb teszteljük a modell egészének irrelevanciáját, és csak ennek elvetése után teszteljük a változókat parciálisan
- A próba konkrét alakja:

$$F_{\text{emp}} = \frac{RSS/k}{ESS/[n - (k + 1)]} \stackrel{H_0}{\sim} \mathcal{F}_{k, n-(k+1)}$$

#### Modell egészének relevanciája

- A tesztstatisztika átírható mint

$$\frac{RSS/k}{ESS/[n - (k + 1)]} = \frac{R^2/k}{(1 - R^2) / [n - (k + 1)]}$$

- Persze: a „nem tér el lényegesen a nullmodelltől” úgy is megfogalmazható, hogy az „ $R^2$  nem tér el lényegesen a nullától” ( $H_0 : R^2 = 0$  is mondható lett volna)

### Modell egészének relevanciája

- A próba neve: a modell egészének relevanciájára irányuló (globális)  $F$ -próba
- Szokás ANOVA-próbának is nevezni (a  $TSS = ESS + RSS$  variancia-felbontáson alapszik; számlálóban és nevezőben a fokszámmal normált szórásnégyzetek vannak)
- Tipikus eredményközlés az ún. ANOVA-táblában

## 4. Tetszőleges számú paraméter

### Felvezető gondolatok

- Valamennyi eddigi próba felírható úgy, hogy van egy modellünk, a nullhipotézis pedig egy megkötést jelent arra a modellre
- Azaz lényegében két modellünk van, egy megkötés nélküli és egy megkötött
- Mellesleg a megkötött modell szükségképp rosszabb, de legalábbis nem jobb (szűkebb tartományon vett optimum nem lehet jobb, mint egy bővebben vett), emiatt úgy is megfogalmazható a kérdés, hogy a különbség lényeges-e
- Az ilyen helyzetre – mint bármilyen helyzetre – többféle elven lehet tesztet konstruálni
- Wald-elv, LM-elv, LR-elv
- Az eddigi két próba Wald-elven is kihozható

### Tetszőleges számú paraméter tesztelése Wald-elven

- Most felírjuk a két modellt explicite is, mert a nullhipotézis alakja szebb lesz (ez pusztán formai kérdés):
- Az egyik modell a bővebb (U – unrestricted), a másik a szűkebb (R – restricted):

$$\begin{aligned} U : Y &= \beta_0 + \beta_1 X_1 + \dots + \beta_{q-1} X_{q-1} + \beta_q X_q + \beta_{q+1} X_{q+1} + \dots + \beta_{q+m} X_{q+m} + \varepsilon_U \\ R : Y &= \beta_0 + \beta_1 X_1 + \dots + \beta_{q-1} X_{q-1} + \beta_q X_q + \varepsilon_R \end{aligned}$$

- $H_0 : \beta_{q+1} = \beta_{q+2} = \dots = \beta_{q+m} = 0$ , tehát megadott  $m$  darab változó még összességében sem bír lényeges magyarázó erővel

## Tetszőleges számú paraméter tesztelése Wald-elven

A próba:

$$\begin{aligned} F_{\text{emp}} &= \frac{(ESS_R - ESS_U) / m}{ESS_U / (n - q - m)} = \\ &= \frac{(R_U^2 - R_R^2) / m}{(1 - R_U^2) / (n - q - m)} \stackrel{H_0}{\sim} \mathcal{F}_{m, n-q-m}. \end{aligned}$$

Ebből a felírásból látszik jól, hogy ez a teszt úgy is felfogható, mint ami a többszörös determinációs együtthatók különbségét ítéli meg.

### Speciális esetek

- Vegyük észre, hogy ez az általános megközelítés a két, eddig látott tesztet is tartalmazza speciális esetként!
- Ha  $m = 1$ , akkor  $F = t_j^2$ : visszakaptuk a  $t$ -tesztet
  - Ám figyelem: a Wald-teszt *nem* ekvivalens a  $t$ -próba  $m$ -szeri elvégzésével (külön-külön az egyes változókra)!
- Ha  $m = k$ , akkor  $F_{\text{Wald}} = F_{\text{ANOVA}}$ : visszakaptuk a függetlenségvizsgálatot
- Logikusak, hiszen a nullhipotézisek is azonos alakúak lettek

Az első állításhoz hozzá kell még tenni, hogy az  $(1, n - k)$  paraméterű  $F$ -eloszlás épp az  $n - k$  szabadságfokú  $t$ -eloszlás négyzetével esik egybe.

### Kitérő: a Lagrange Multiplikátor (LM)-elv

- Az LM (Lagrange Multiplikátor) próba hipotézispárja *teljesen* azonos alakú a Wald- $F$ -teszttel:

$$\begin{aligned} U : Y &= \beta_0 + \beta_1 X_1 + \dots + \beta_{q-1} X_{q-1} + \beta_q X_q + \beta_{q+1} X_{q+1} + \dots + \beta_{q+m} X_{q+m} + \varepsilon_U \\ R : Y &= \beta_0 + \beta_1 X_1 + \dots + \beta_{q-1} X_{q-1} + \beta_q X_q + \varepsilon_R \end{aligned}$$

$$\text{és } H_0 : \beta_{q+1} = \beta_{q+2} = \dots = \beta_{q+m} = 0$$

- A különbség a modellezés filozófiájában van (ld. később), a teszt tulajdonságai, alkalmazhatósága is eltérő
- Alapötlet: becsüljük meg a szűkebb modellt, és számítsuk ki ez alapján a becsült reziduumokat. Ha fennáll  $H_0$ , akkor ezek a reziduumok nem magyarázhatóak lényegesen sem a szűkebb modell változóival (OLS következménye), sem a vizsgált változókkal ( $H_0$  következménye). Azaz: ha a becsült reziduumokat kiregresszáljuk az összes változóval, akkor sem tudjuk azt lényegesen magyarázni, ha fennáll a  $H_0$ .

### Az LM-próba próbafüggvénye

- Ezen intuitív indoklás után a próbafüggvény:

$$n \cdot R_{\hat{e}_R|X_1, X_2, \dots, X_{q+m}}^2 \stackrel{H_0}{\sim} \chi_m^2$$

- Itt  $\hat{e}_R$  jelölés arra utal, hogy a szűkebb ( $R$ ) modellből kapott reziduumokról van szó

## 5. Lineáris megkötés(ek)

### Lineáris kombináció tesztelése

- A séma:

$$r_1\beta_1 + r_2\beta_2 + \dots + r_k\beta_k = r$$

- Avagy röviden:  $\mathbf{r}^T \boldsymbol{\beta} = r$
- Több koefficiens is érinthet, de csak egy egyenletet tartalmazhat
- Például:
  - Két koefficiens egyezik,  $\beta_l = \beta_m$  (ekkor  $r_l = +1$ ,  $r_m = -1$ , a többi  $r_i$  nulla és  $r = 0$ )
  - Egyik koefficiens  $c$ -szerese a másiknak,  $\beta_l = c\beta_m$  (ekkor  $r_l = +1$ ,  $r_m = -c$ , a többi  $r_i$  nulla és  $r = 0$ )
  - Az összes koefficiens összege épp nulla (ekkor mindegyik  $r_i$  1 és  $r = 0$ )

### Lineáris kombináció tesztelése

- A normális lineáris modellben erre teszt szerkeszthető
- Megvalósítás: egyik lehetőség, hogy a  $t$ -próbához hasonló alakra vezetjük vissza
- Legyen  $r_1\hat{\beta}_1 + r_2\hat{\beta}_2 + \dots + r_k\hat{\beta}_k = \hat{r}$ , ekkor

$$\frac{\hat{r} - r}{\text{se}(\hat{r})} \stackrel{H_0}{\sim} t_{n-(k+1)}$$

- Ez az ún. *közvetlen  $t$ -próba*
- Vizsgálható Wald-jellegű próbával is

### Speciális esetek

- Ez tartalmazza speciális esetként a parciális  $t$ -próbát
- De más nem: kettő vagy több paraméter *egyidejű* nulla mivolta több megkötést jelent
- Szerencsére az előbbi kiterjeszthető több megkötés tesztelésére is:

$$\begin{aligned} \mathbf{r}_1^T \boldsymbol{\beta} &= r_1 \\ \mathbf{r}_2^T \boldsymbol{\beta} &= r_2 \\ &\vdots \\ \mathbf{r}_m^T \boldsymbol{\beta} &= r_m \end{aligned}$$

- Az  $\mathbf{r}_i^T$  sorvektorokat rakjuk össze egy  $\mathbf{R}$  mátrixba, az  $r_i$  skalárokat egy  $r$  oszlopvektorba

### Több megkötés egyidejű tesztelése

- Célszerű felírás:

$$H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r},$$

ahol  $\mathbf{R}$   $m \times k$  típusú (tehát  $m$  a megszorítások száma)

- Az erre adható teszt:

$$F_{\text{emp}} = \frac{(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})^T \left[ \mathbf{R} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{R}^T \right]^{-1} (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}) / m}{\text{ESS} / [n - (k + 1)]} \stackrel{H_0}{\sim} \mathcal{F}[m, n - (k + 1)]$$

Feltétel még, hogy  $\mathbf{R}$  teljes sorrangú legyen ( $\text{rank } \mathbf{R} = m$ ), ami azt a kézenfekvő követelményt fogalmazza meg, hogy a megszorítások ne legyenek (lineáris értelemben) redundánsak.

### Konkrét példák a fenti sémára

- Ellenőrizhető, hogy ha például...
  - $\dots \mathbf{R} = \begin{pmatrix} 0 & 0 & \dots & 0 & 1 & 0 \dots 0 \end{pmatrix}$  és  $r = 0$ , akkor a  $t$ -tesztet ...

$$\begin{aligned}
- \text{...}\mathbf{R} &= \begin{pmatrix} 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix} \text{ és } \mathbf{r} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} \text{ akkor az ANOVA-t...} \\
- \text{...}\mathbf{R} &= \begin{pmatrix} \lambda_{\beta_1} & \lambda_{\beta_2} & \dots & \lambda_{\beta_k} \end{pmatrix} \text{ és } r = \Lambda, \text{ akkor a lineáris kombináció tesztelése...}
\end{aligned}$$

- ...kapjuk vissza.

### **Speciális esetek**

- Ez a képlet viszont *minden* eddig látott dolgot tartalmaz speciális esetként!
- Wald-elven