Hipotézisvizsgálat és intervallumbecslés lineáris modellben

Ferenci Tamás
tamas.ferenci@medstat.hu

Utoljára frissítve: 2023. május 9.

Tartalom

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

Emlékeztetőül

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

Becsült regressziós koefficiensek mintavételi eloszlása

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

$$\frac{\widehat{\beta}_i - \beta_i}{\mathrm{se}\left(\widehat{\beta}_i\right)} \sim \mathcal{N}\left(0,1\right),$$

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

- Sajnos ezzel a gyakorlatban nem sokra megyünk, mert σ^2-et általában nem ismerjük
- Helyettesítsük a jó tulajdonságú becslőjével, $\widehat{\sigma^2}$-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 $t_{\mathrm{emp},i} = \frac{\widehat{\beta}_i}{\mathrm{se}\left(\widehat{\beta}_i\right)}$ kifejezés $n-(k+1)$ szabadságfokú t-eloszlást követ (nulleloszlás)
3. Számítsuk ki a konkrét $t_{\mathrm{emp},i}$-t a mintánkból és döntsük el, hogy hihető-e, hogy $t_{n-(k+1)}$-ból 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):

$$t_{\mathrm{emp},i} = \frac{\widehat{\beta}_i}{\mathrm{se}\left(\widehat{\beta}_i\right)} \overset{H_0}{\sim} t_{n-(k+1)}.$$

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

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:
$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$
- (Természetesen a β_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 utána teszteljük a változókat parciálisan
- A próba konkrét alakja:

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

Modell egészének relevanciája

- A tesztstatisztika átírható mint

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

- 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

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):

$$U : Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_{q-1} X_{q-1} + \beta_q X_q + \beta_{q+1} X_{q+1} + \ldots + \beta_{q+m} X_{q+m} + \varepsilon_U$$
$$R : Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_{q-1} X_{q-1} + \beta_q X_q + \varepsilon_R$$

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

Tetszőleges számú paraméter tesztelése Wald-elven

A próba:

$$F_{\mathrm{emp}} = \frac{\left(ESS_R - ESS_U\right)/m}{ESS_U/\left(n-q-m\right)} = \\ = \frac{\left(R_U^2 - R_R^2\right)/m}{\left(1-R_U^2\right)/\left(n-q-m\right)} \overset{H_0}{\sim} \mathcal{F}_{m,n-q-m}.$$

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 $F_{\mathrm{Wald}} = F_{\mathrm{ANOVA}}$: visszakaptuk a függetlenségvizsgálatot
- Logikusak, hiszen a nullhipotézisek is azonos alakúak lettek

Kitérő: a Lagrange Multiplikátor (LM)-elv

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

$$U: Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_{q-1} X_{q-1} + \beta_q X_q + \beta_{q+1} X_{q+1} + \ldots + \beta_{q+m} X_{q+m} + \varepsilon_U$$
$$R: Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_{q-1} X_{q-1} + \beta_q X_q + \varepsilon_R$$

  és $H_0 : \beta_{q+1} = \beta_{q+2} = \ldots = \beta_{q+m} = 0$
- 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:

$$n \cdot R^2_{\widehat{e}_R|X_1,X_2,\ldots,X_{q+m}} \overset{H_0}{\sim} \chi^2_m$$

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

Lineáris kombináció tesztelése

- A séma:

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

- Avagy röviden: $\mathbf{r}^T\boldsymbol{\beta} = r$
- Több koefficienst 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 $r_1\widehat{\beta}_1 + r_2\widehat{\beta}_2 + \ldots + r_k\widehat{\beta}_k = \widehat{r}$, ekkor

$$\frac{\widehat{r} - r}{\mathrm{se}\left(\widehat{r}\right)} \overset{H_0}{\sim} t_{n-(k+1)}$$

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

Speciális esetek

- Ez tartalmazza speciális esetként a parciális t-próbát
- De mást 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:

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

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

- Az erre adható teszt:

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

Konkrét példák a fenti sémára

- Ellenőrizhető, hogy ha például…
  - …$\mathbf{R} = \begin{pmatrix} 0 & 0 & \ldots & 0 & 1 & 0\ldots 0 \end{pmatrix}$ és $r = 0$, akkor a t-tesztet …
  - …$\mathbf{R} = \begin{pmatrix} 0 & 1 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 1 & 0 \\ 0 & 0 & 0 & \ldots & 0 & 1 \end{pmatrix}$ és $\mathbf{r} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}$ akkor az ANOVA-t…
  - …$\mathbf{R} = \begin{pmatrix} \lambda_{\beta_1} & \lambda_{\beta_2} & \ldots & \lambda_{\beta_k} \end{pmatrix}$ és $r = \Lambda$, akkor a lineáris kombináció tesztelését…
- …kapjuk vissza.

Speciális esetek

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