Gauss-Markov feltevések

Ahhoz, hogy az OLS elven becsült lineáris regressziónak fennálljanak bizonyos előnyös tulajdonságai, meghatározott feltevéseknek teljesülniük kell:

1. **Modell linearitás**: Feltesszük, hogy az Y és az X közötti kapcsolat lineárisan leírható
2. Az adatok a sokaság **véletlen mintái**. A legtöbbször használt i.i.d. ennél szigorúbb, tehát teljesíti a véletlen mintás feltevést.
3. X mátrix **teljes oszlop rangú**. Ezen feltevés alpján tehát nincs egzakt multikolinearitás. Ha ezt nem teszük fel, akkor $(X^T X)$ nem lenne invertálható tetszőleges X esetén.
4. $\mathbb{E}[\epsilon|X] = 0$. Ez a feltevés (0 feltételes átlag) alapján a hibatagok átlaga 0. Ebből következik, hogy $\mathbb{E}(Y) = X\beta$.
5. $\mathbb{E}[\epsilon\epsilon^T|X] = \sigma^2 I$. Ezen feltevés szerint a hibatagok homoszkedasztikusak és autokorrelálatlanok.

Z-érték

A hipotézis, amit a becsült modell paraméterek egyikére (k) határozunk meg:

$$H_0 : \hat{\beta}_k = b_k$$

Ebben az esetben b_k előre ismert érték, amihez mért távolságot fogjuk mérni a **null hipotézis** keretében. A null hipotézist az **alternatív hipotéssiel** szemben vizsgáljuk ($H_1 : \hat{\beta}_k \neq b_k$), α szignifikancia szint mellett.

Z-érték

Amennyiben **feltesszük**, hogy a Gauss-Markov feltevések teljesültek, és a hibatag normális eloszlású, akkor

$$(\hat{\beta}_k - b_k)|X \sim N(0, \sigma^2((X^T X)^{-1})_{kk}),$$

ahol $((X^T X)^{-1})_{kk})$ a $(X^T X)^{-1}$ mátrix főátlójának k-adik eleme.
Ekkor:

$$z_k \equiv \frac{\hat{\beta}_k - b_k}{\sqrt{\sigma^2((X^T X)^{-1})_{kk}}} \sim N(0,1).$$

Z-érték tulajdonságai

A Z értéknek nagyon jó statisztikai tulajdonságai vannak, **feltéve ha** ismerjük a varianciát. Ekkor z_k

1. értéke a mintából kiszámolható (értsd, a nevező kiszámolható becslés nélkül),
2. eloszlása nem függ az X eloszlásától,
3. eloszlása előzetesen ismert, tehát nem függ ismeretlen paraméterektől.

Nuisance parameters: Azok az ismeretlen paraméterek, amelyktől a test statisztika eloszlása függ.

Mivel a gyakorlatban a σ^2 értékét nem ismerjük, ezért a kézenfekvő megoldás, hogy az ismeretlen variancia helyére az OLS becslésből számított minta varianciát (s^2) illesztjük.

$$s^2 = \frac{\epsilon^T \epsilon}{n-K},$$

ahol n a minta elemszáma, K pedig a becsült paraméterek száma.

T-érték

A minta variancia behelyettesítése után a teszt statisztika értékét t-értéknek nevezzük. Ebben az esetben a nevezőbe a β_k paraméter OLS elvű becsült értékének standard hiábja kerül, amit $SE(\hat{\beta}_k)$-val jelölünk.

$$SE(\hat{\beta}_k) \equiv \sqrt{s^2((X^T X)^{-1})_{kk})}$$

Mivel s^2 egy valószínűségi változó (hiszen a minta függvényében változik az értéke), ez a behelyettesítés megváltoztatja a teszt statisztika eloszlását. Szerencsére az új eloszlás szintén nem függ függ ismeretlen paraméterektől, sem X-től.

T-érték

Amennyiben **feltesszük**, hogy a Gauss-Markov feltevések teljesültek, és a hibatag normális eloszlású, akkor

$$t_k \equiv \frac{\hat{\beta}_k - b_k}{SE(\hat{\beta}_k)} \equiv \frac{\hat{\beta}_k - b_k}{\sqrt{s^2((X^T X)^{-1})_{kk})}} \sim t(n - K). \tag{3}$$

T-teszt

A t-érték alapú null hipozézis tesztelést t-teszt-nek nevezzük. T-tesztet a következő módon végzünk:

1. Számold ki a t-értéket a (3) egyenlet alapján.
2. Számold ki a kritikus értéket: $(n-K)$ szabadságfokú t eloszlás inverz kumulatív eloszlás függvényébe (CDF^{-1} v. PPF) helyettesítsd be az $\alpha/2$ értékét (regressziós paramétereket jellemzően 2 oldalon tesztelünk). A kapott értéket $t_{\alpha/2}$-vel jelöljük.
3. Amennyiben az 1. lépésben számolt t-érték a 2. lépésben számolt $[-t_{\alpha/2}; t_{\alpha/2}]$ intervallumba esik, akkor a H_0 null hipotézist nem tudjuk elvetni. Ha az intervallumon kívül, akkor 1-α konfidencia szint mellett H_0 null hipotézist elvetjük.

Amennyiben a becsült paraméterre szeretnénk **konfidencia intervallumot** (CI) meghatározni, akkor a t-érték definíciójából kiindulva, és azt átrendezve megkapjuk az $[\hat{\beta}_k - SE(\hat{\beta}_k)t_{\alpha/2}; \hat{\beta}_k + SE(\hat{\beta}_k)t_{\alpha/2}]$ intervallumot, amely már $\hat{\beta}_k$ mértékegységében értelmezhető.

P-érték alapú döntéshozatal t-teszt esetén

A t-érték alapú null hipozézis tesztelést t-teszt-nek nevezzük. T-tesztet a következő módon végzünk:

1. Számold ki a t-értéket a (3) egyenlet alapján.
2. $(n-K)$ szabadságfokú t eloszlás kumulatív eloszlás függvényébe (CDF), helyettesítsd be az előző lépésben kapott t-értéket.
3. A 2. lépésben kapott értéket vond ki 1-ből, a kapott értéket szorozd meg 2-vel.

$$\text{p-value} = [1 - CDF_t(t)] * 2$$

4. Ha p $> \alpha$, akkor H_0-t nem tudjuk elvetni, ellenkező esetben, H_0-t elvetjük.

Lineáris hipotézisek

A null hipotézis nem feltetélenül 1 paraméterre fogalmazhatók meg, számos esetben a model több paraméterét egyszerre, azok lineáris kombinációját kívánjuk vizsgálni. Ebben az esetben egyenletrendszer formájában tesztelünk a következő módon:

$$H_0 : R\hat{\beta} = r, \tag{4}$$

ahol, R restrikciós mátrix $(m \times K)$ és r pedig a restrikciós egyenletrendszer skalárjainak vektora $(m \times 1)$.

Példa

Szeretnénk megvizsgálni egy olyan regressziót, amelyben 4 jobboldali változónk van (az első változó a konstans). Amennyiben azt vizsgáljuk, hogy $\beta_2 = \beta_3$ és $\beta_4 = 0$, akkor az a (4) egyenlet formájában a következő módon írható fel:

$$R = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, r = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Wald statisztika

A wald statisztika az $R\hat{\beta}$ vektor és az r vektor közötti távolságot vizsgálja. Amennyiben a null hipotézis igaz, akkor az $R\hat{\beta} - r \approx 0$.

Wald statisztika

Kis mintás környezetbe $R\hat{\beta} - r$ eloszlása:

$$R\hat{\beta} - r \sim N(0, \sigma^2 R(X^T X)^{-1} R^T)$$

Ez alapján a nullhipotézis H_0: $R\hat{\beta} - r = 0$, és az alternatív H_1: $R\hat{\beta} - r \neq 0$, a teszt statisztika értéke pedig:

$$W_{Infeasible} = \frac{(R\hat{\beta} - r)^T [R(X^T X)^{-1} R^T]^{-1} (R\hat{\beta} - r)}{\sigma^2} \sim \chi^2_m$$

Ez a statisztika közvetlen formában nem használható, mert a sokasági variancia ismeretlen, ezért ismét behelyettesítéssel kell élnünk.

Kis mintás Wald teszt

Ahhoz, hogy a Wald statisztika értéke kiszámolható legyen, be kell helyettesítenünk a minta varinaciát (s^2) a sokasági variancia helyére.

Wald teszt

Amennyiben **feltesszük**, hogy a Gauss-Markov feltevések teljesültek, és a hibatag normális eloszlású, akkor

$$W = \frac{(R\hat{\beta} - r)^T [R(X^T X)^{-1} R^T]^{-1} (R\hat{\beta} - r)/m}{s^2} \sim F_{m,n-K}.$$

A variancia behelyettesítés hatására megváltozott a Wald statisztika eloszlása, amelyre a hipotézisvizsgálat során oda kell figyelnünk. A hipotézis vizsgálat, a t-teszttel analóg módon végezhető el.

Illeszkedés

A becsült paraméterekre vonatkozó hipotézisvizsgálatok mellett azt is meg kell állapítanunk, hogy a modell milyen jól illeszkedik az adatokra. Erre lineáris regresszió során többek között a (Centrált) R^2 és a korrigált R^2 mutatók szolgálnak. Ezen mutatók azt regadják meg, hogy a **a minta variancia mekkora részét magyarázza a modell**.

Centrált R^2 és korrigált R^2

$$R^2 = 1 - \frac{\epsilon^T \epsilon}{\tilde{Y}^T \tilde{Y}},$$

ahol $\tilde{Y} = Y - \overline{Y}$, tehát a célváltozó átlagtól szűrt értékei. Mivel az R^2 értéke minden hozzáadott változóval nő, ezért ezt az értéket korrigálni érdemes, hogy a modell bővítés mellett megmaradjon az összehasonlíthatóság.

$$Adj.R^2 = 1 - \frac{n-1}{n-K-1}(1 - R^2),$$