Többszörös lineáris regresszió
Modell diagnosztika

Matematikai statisztika
2024. október 28.

Többváltozós lineáris regresszió - Modell tesztelése

A modell tesztelésének célja

A modell tesztelése elengedhetetlen ahhoz, hogy megbizonyosodjunk a becsült paraméterek és az egyes magyarázó változók valódi hozzájárulásáról. A modell tesztelésére nem csak az együtthatókra, hanem az egész modellre vonatkozóan is szükség van. Mielőtt az egyes együtthatók jelentőségét vizsgálnánk, ellenőrizzük, hogy a modell összességében releváns-e.

Miért szükséges az egész modell tesztelése?

Az egész modell tesztelése segít meghatározni, hogy az összes magyarázó változó együttesen mennyire járul hozzá a függő változó varianciájának magyarázatához. Ha a modell egészében nem releváns, akkor felesleges az egyes együtthatók további vizsgálata.

A hibatag eloszlása és következményei

Hibatag normális eloszlásának feltétele

Feltételezzük, hogy a hibatagok normális eloszlásúak, azaz:

$$\varepsilon_i \sim N(0, \sigma_0^2).$$

Ez az eloszlás feltétel lehetővé teszi a modell becsült paramétereinek (azaz $\hat{\mathbf{b}}$) normális eloszlással való közelítését.

Következmény - Együtthatók eloszlása

A hibatag normális eloszlása alapján az együtthatók becslései is normális eloszlást követnek:

$$\hat{\mathbf{b}} \sim N\left(\mathbf{b}, \sigma_0^2(\mathbf{X}^T\mathbf{X})^{-1}\right),$$

ahol $\sigma_0^2(\mathbf{X}^T\mathbf{X})^{-1}$ a kovarianciamátrix. Ez alapvető az intervallumbecslés és a hipotézisvizsgálatok elvégzéséhez.

A modell egészének relevanciájának vizsgálata

A teljes modell nullhipotézise

A modell teljes magyarázóerejét úgy teszteljük, hogy vizsgáljuk: az összes magyarázó változó együttvéve szignifikáns hatást gyakorol-e a függő változóra. A nullhipotézis:

$$H_0 : b_0 = b_1 = \ldots = b_k = 0 \quad \text{(nincs magyarázóerő)}.$$

Ha H_0 igaz, akkor egyik változó sem járul hozzá jelentősen a függő változó előrejelzéséhez.

Alternatív hipotézis

Az alternatív hipotézis szerint legalább egy változó szignifikáns:

$$H_1 : \exists i : b_i \neq 0.$$

Ez a teszt megmutatja, hogy az egész modell hasznos-e, mielőtt az egyes változók szignifikanciáját külön megvizsgálnánk.

F-próba a modell relevanciájának vizsgálatához

F-próba statisztika

A teljes modell relevanciáját F-próbával (ANOVA) teszteljük. Az F-statisztika így számítható:

$$F = \frac{\text{SSR}/k}{\text{SSE}/(n-k-1)},$$

ahol:

- SSR: a magyarázott szórásnégyzetösszeg,
- SSE: a hiba szórásnégyzetösszeg,
- k: a magyarázó változók száma,
- n: a megfigyelések száma.

F-eloszlás H_0 esetén

Ha a nullhipotézis igaz (H_0), az F-statisztika $F_{k,n-k-1}$ eloszlást követ. Ha az F-statisztika túl nagy, elutasíthatjuk H_0-t, ezzel igazolva, hogy a modell szignifikáns magyarázóerővel rendelkezik.

Többváltozós lineáris regresszió - Együtthatók tesztelése

A változók relevanciájának tesztelése

Ha a teljes modell szignifikáns, azaz nem egy nullmodell, akkor az egyes magyarázó változók relevanciája is megvizsgálható. Az adott változó akkor **releváns**, ha a hozzá tartozó regressziós együttható (b_i) szignifikánsan eltér nullától. Ez azt jelzi, hogy az adott változó statisztikailag szignifikáns hatást gyakorol a függő változóra.

A relevancia fontossága

Az egyes változók külön vizsgálata segít feltárni, hogy mely tényezők járulnak hozzá a modell pontosságához, és melyek csak zajt visznek be. Ezzel biztosíthatjuk, hogy csak a releváns változók maradjanak a modellben.

Hipotézisek és tesztelési módszer

Nullhipotézis és alternatív hipotézis

Minden egyes változó esetében a nullhipotézis (H_0) azt feltételezi, hogy az adott változónak nincs hatása:

$$H_0 : b_i = 0 \quad \text{(nincs hatása)}.$$

Az alternatív hipotézis (H_1) szerint az együttható különbözik nullától, azaz a változó hatással bír:

$$H_1 : b_i \neq 0 \quad \text{(jelentős hatással bír)}.$$

Tesztelési módszer - t-próba

Az egyes együtthatók nullától való eltérésének vizsgálatához a t-próbát alkalmazzuk, mivel a következő arány t-eloszlást követ:

$$t = \frac{\hat{b}_i}{s_{\hat{b}_i}} \sim t_{n-k-1},$$

ahol $s_{\hat{b}_i}$ az $\hat{b}_i$ becslésének standard hibája.

Próbastatisztika és döntési kritérium

Próbastatisztika kiszámítása

A t-próbában használt próbastatisztika az egyes változók tesztelésére így számítható:

$$t = \frac{\hat{b}_i - 0}{s_{\hat{b}_i}}.$$

Ez azt jelenti, hogy a becsült együtthatót ($\hat{b}_i$) elosztjuk a becslésének standard hibájával ($s_{\hat{b}_i}$).

Döntési kritérium

Ha a t-érték nagyobb a kritikus szintnél a választott szignifikanciaszinten, akkor elutasítjuk H_0-t, és arra következtetünk, hogy az adott együttható statisztikailag jelentős. Ezáltal a változó relevánsnak tekinthető a modellben.

Intervallumbecslés az együtthatókra

Intervallumbecslés kiszámítása

A t-eloszlás segítségével egy adott megbízhatósági szinten konfidenciaintervallumot határozhatunk meg az együtthatókra. Az intervallumbecslés képlete:

$$b_i \pm t_{\alpha/2,n-k-1} \cdot s_{\hat{b}_i},$$

ahol:

- b_i: a valódi együttható,
- $t_{\alpha/2,n-k-1}$: a választott szignifikanciaszinthez tartozó kritikus érték,
- $s_{\hat{b}_i}$: az $\hat{b}_i$ becslésének standard hibája.

Az intervallumbecslés jelentése

Az intervallumbecslés segítségével megbecsülhetjük, hogy az egyes együtthatók milyen tartományban mozoghatnak a választott szignifikanciaszinten belül, és ez által megbízhatósági becslést adhatunk az együtthatókra.

A modell hibájának becslése és szórásnégyzet

A modell hibatagjainak varianciája

A modell hibatagjainak varianciáját becsléssel számíthatjuk ki:

$$\widehat{\sigma^2} = \frac{\varepsilon^T \varepsilon}{n - (k+1)},$$

ahol ε a reziduális hibatagok vektora, n a megfigyelések száma, és k a magyarázó változók száma. Ez a becslés alapvető a modell pontosságának értékeléséhez.

A szórásnégyzet szerepe az együtthatók megbízhatóságában

Az együtthatók becslésének szórásnégyzete a hibatagok varianciájának ismeretében adható meg:

$$\widehat{\sigma^2_{\hat{b}}} = \widehat{\sigma^2} \cdot (\mathbf{X}^T \mathbf{X})^{-1}.$$

Ezáltal pontosan meghatározhatjuk az együtthatók becslésének pontosságát, amely elengedhetetlen az intervallumbecslések és a hipotézisvizsgálatok számára.

Lehetséges hibák a modellekben

Alapfeltevések a modellről

A modell helyessége érdekében több alapfeltevést teszünk, amelyeknek teljesülniük kell. Ezek közé tartozik, hogy:

- **Lineáris kapcsolat** van a magyarázó változók és az eredményváltozó között:

$$Y_i = b_0 + b_1 X_{i,1} + b_2 X_{i,2} + \ldots + b_k X_{i,k} + \varepsilon_i.$$

- **Nincs egzakt multikollinearitás**: A magyarázó változók között nincs tökéletes lineáris kapcsolat.
- **Erős exogenitás** áll fenn:

$$\mathbb{E}(\varepsilon_i | \mathbf{X}) = 0.$$

További feltételek a hibatagokra

További feltételezések

A modell pontosságának biztosítása érdekében további feltevéseket teszünk a hibatagokra:

- **Homoszkedaszticitás**: A hibatagok varianciája konstans, azaz

$$\sigma^2(\varepsilon_i|\mathbf{X}) = \sigma_0^2.$$

- **Nincs autokorreláció**: A különböző megfigyelések hibái függetlenek, vagyis

$$\mathrm{Cov}(\varepsilon_i, \varepsilon_j|\mathbf{X}) = 0, \quad \text{ha} \quad i \neq j.$$

- **Hibatagok normalitása**: Feltételezzük, hogy a hibatagok normális eloszlásúak.

Multikollinearitás fogalma

Multikollinearitás meghatározása

A multikollinearitás a magyarázó változók közötti erős lineáris kapcsolatot jelenti. Ez problémát jelenthet a többszörös lineáris regresszióban, mivel az egyes együtthatók becslései instabillá válhatnak, ha a magyarázó változók szorosan korrelálnak egymással.

Az egzakt multikollinearitás kizárása

Feltételezzük, hogy nincs tökéletes lineáris kapcsolat a magyarázó változók között, azaz:

$$\mathbb{P}(\mathbf{X} \text{ rangja} = k + 1) = 1.$$

Ezzel biztosítjuk, hogy a modellben szereplő változók teljes rangúak legyenek, így egyértelműen becsülhetők a regressziós együtthatók.

Multikollinearitás hatása az együtthatók becslésére

Multikollinearitás és az együtthatók szórásnégyzete

A multikollinearitás jelenléte növeli az együtthatók becslésének varianciáját. Az egyes becslések varianciája a következőképpen alakul:

$$\sigma^2(\hat{b}_i|\mathbf{X}) = \frac{\sigma_0^2}{SST_i(1-R_i^2)},$$

ahol SST_i a változó teljes szórásnégyzetösszege, és R_i^2 a magyarázó változó más változókkal vett determinációs együtthatója.

Instabilitás a multikollinearitás miatt

Ha az R_i^2 érték közel van az 1-hez, akkor az együttható becslésének varianciája nagymértékben megnő, így a modell instabillá válik. Ezért fontos a multikollinearitás mértékének értékelése.

Multikollinearitás mérőszámai

A multikollinearitás jelenlétének és mértékének azonosításához több mérőszámot is használhatunk:

- **Tolerancia**: Az i-edik változó többi változóból történő lineáris magyarázottságát méri.
- **Variancia infláló faktor** (VIF): A becslési varianciát mutatja, ha a magyarázó változók korreláltak.
- **Kondíciós index** (CI): Az $\mathbf{X}^T\mathbf{X}$ mátrix sajátértékeiből számított mutató.

Tolerancia és Variancia infláló faktor (VIF)

Tolerancia kiszámítása

A tolerancia értéke $1 - R_i^2$, ahol R_i^2 az i-edik változónak a többi magyarázó változóval vett determinációs együtthatója. Ha a tolerancia alacsony (nullához közeli), akkor az i-edik változó nagyfokú lineáris kapcsolatban áll a többivel.

VIF - Variancia infláló faktor

A VIF a tolerancia reciproka:

$$\text{VIF} = \frac{1}{1 - R_i^2}.$$

Ha a VIF értéke 10 fölött van, akkor jelentős multikollinearitás áll fenn. Ha a magyarázó változók korrelálatlanok, a VIF értéke 1.

Kondíciós index (CI)

Kondíciós index (CI) kiszámítása

A kondíciós index az $\mathbf{X}^T\mathbf{X}$ mátrix legnagyobb és legkisebb sajátértékének négyzetgyökös hányadosa:

$$\text{CI} = \sqrt{\frac{\lambda_{\text{max}}}{\lambda_{\text{min}}}}.$$

Ha a CI értéke 15-nél nagyobb, akkor erős multikollinearitás jelenlétére utal, ami a modell stabilitásának szempontjából kedvezőtlen.

CI értékelése

A CI magas értéke azt jelzi, hogy a magyarázó változók között erős lineáris kapcsolat van, ami csökkenti a modell megbízhatóságát. Jó értéknek számít, ha a CI kisebb, mint 15.

Multikollinearitás kezelése

Lehetséges megoldások a multikollinearitás kezelésére

Ha multikollinearitás lép fel, az alábbi lépésekkel csökkenthetjük annak hatását:

- **Nagyobb mintaméret**: Ha több adatot gyűjtünk, az együtthatók becslésének varianciája csökkenhet.
- **Főkomponens-analízis (PCA)**: Az eredeti magyarázó változókból olyan új, korrelálatlan változókat képezhetünk, amelyekkel csökkenthető a multikollinearitás.

PCA alkalmazása multikollinearitás kezelésére

A PCA eltávolítja a magyarázó változók közötti korrelációkat, így új, korrelálatlan főkomponensekkel dolgozhatunk. Ez egy hatékony módszer a multikollinearitás kezelésére, bár a kapott komponensek gyakran kevésbé intuitívak.

A modell problémáinak tesztelése - Többszörös lineáris regresszió

A többszörös és egyszerű lineáris regresszió tesztelési hasonlóságai

A többszörös lineáris regresszió esetében a fő problémák, mint a normalitás, homoszkedaszticitás és autokorreláció tesztelésére ugyanazokat a módszereket használhatjuk, mint az egyszerű lineáris regressziónál. Az alábbi tesztek alkalmazhatók:

- **Normalitás ellenőrzése**: A hibatagok normális eloszlását Q-Q plotokkal vagy Shapiro-Wilk teszttel vizsgálhatjuk.
- **Homoszkedaszticitás vizsgálata**: A Breusch-Pagan vagy White-tesztet alkalmazhatjuk a hibatagok konstans varianciájának ellenőrzésére.
- **Autokorreláció vizsgálata**: A Durbin-Watson teszt segítségével ellenőrizhetjük, hogy van-e korreláció a hibatagok között, különösen időbeli adatok esetén.

Example

Miért fontos a problémák tesztelése? Ezen feltételek megsértése torzítja az eredményeket, növeli a becslések varianciáját, és csökkenti a hipotézisvizsgálatok érvényességét. A tesztek biztosítják, hogy a modell megbízható eredményeket szolgáltasson.