Többszörös lineáris regresszió
Illeszkedés diagnosztika

Matematikai statisztika
2024. október 28.

Többváltozós lineáris regresszió - Modellminősítés és Illeszkedésdiagnosztika

Miért van szükségünk illeszkedés és modelldiagnosztikára?

A diagnosztikák célja, hogy megvizsgáljuk, mennyire megfelelő a regressziós modell az adatokra nézve. Az alábbi szempontok miatt elengedhetetlen a modell alapos ellenőrzése:

- **Modell feltevéseinek ellenőrzése**: A többváltozós lineáris regresszió bizonyos feltevéseken alapul, mint például a hibák normális eloszlása, homoszkedaszticitása és függetlensége. Ha ezek a feltevések sérülnek, a modell becslései és következtetései torzulhatnak.
- **A modell teljesítményének felmérése**: Az illeszkedésdiagnosztika megmutatja, hogy a modell mennyire jól magyarázza a függő változó változását, és mennyire pontosak az előrejelzések.
- **Potenciális problémák azonosítása**: Az illeszkedésdiagnosztika segít azonosítani olyan problémákat, mint az outlierek vagy a multikollinearitás, amelyek befolyásolhatják az eredményeket.

Többváltozós lineáris regresszió - Szórásfelbontás és Szórásnégyzetösszegek

Szórásfelbontás többváltozós esetben

A többváltozós lineáris regresszió modell illesztésének értékeléséhez használjuk a következő jelöléseket: $\hat{y} = \mathbf{X}\hat{\mathbf{b}}$, ahol:

- $\hat{y}$ az előrejelzett értékek vektora,
- $\mathbf{X}$ a magyarázó változók mátrixa,
- $\hat{\mathbf{b}}$ a becsült együtthatók vektora.

A modell illeszkedésének vizsgálata szórásnégyzetösszegeken alapul:

- **SST** (Teljes szórásnégyzetösszeg): Az adatok összes varianciája.
- **SSR** (Reprezentált szórásnégyzetösszeg): A modell által megmagyarázott variancia.
- **SSE** (Hiba szórásnégyzetösszeg): A modell által nem magyarázott variancia.

Szórásfelbontás elemei

A szórásnégyzetösszegek a modell különböző összetevőinek változékonyságát mérik:

- **SST** (Teljes szórásnégyzetösszeg): Méri a teljes varianciát az eredményváltozó (y) körül, definiálva mint:

$$\text{SST} = \sum_{i=1}^{n}(y_i - \overline{y})^2,$$

  ahol $\overline{y}$ az y változó átlaga.

- **SSR** (Regressziós szórásnégyzetösszeg): Az előrejelzett értékek ($\hat{y}$) varianciáját adja meg az eredményváltozó átlagához képest:

$$\text{SSR} = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 = \hat{y}^T\hat{y} - n\overline{y}^2.$$

- **SSE** (Hiba szórásnégyzetösszeg): Az eltérés a tényleges és az előrejelzett értékek között, amelyet a hibák négyzetösszegeként határozunk meg:

$$\text{SSE} = \vec{\varepsilon}^T\vec{\varepsilon},$$

  ahol $\vec{\varepsilon}$ a reziduális hibák vektora.

Többváltozós lineáris regresszió - Modellminősítés

SST felbontása

Hasonlóan az egyszerű lineáris regresszióhoz, a teljes szórásnégyzetösszeg felbontható a reprezentált szórásnégyzetösszeg és a hiba szórásnégyzetösszeg összegére:

$$\text{SST} = \text{SSR} + \text{SSE}.$$

Ez a felbontás segít megérteni, hogy a modell mennyire illeszkedik a megfigyelésekhez, és milyen arányban képes magyarázni a változók közötti kapcsolatot.

TLR - Determinációs együttható (R^2)

Determinációs együttható (R^2)

Az R^2 determinációs együttható megmutatja, hogy a modell mennyire képes magyarázni az eredményváltozó varianciáját, és a következőképpen definiálható:

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}.$$

Az R^2 értéke 0 és 1 között mozog, ahol a magasabb érték jobb illeszkedést jelez.

Determinációs együttható értelmezése

Az R^2 azt jelzi, hogy a modell által tartalmazott magyarázó változók mennyire járulnak hozzá az eredményváltozó előrejelzéséhez. Minél magasabb az R^2, annál több információt nyerünk a modellből a magyarázó változókról.

Többszörös korrelációs együttható

Az R^2 négyzetgyöke, $\sqrt{R^2}$, a többszörös korrelációs együttható, amely egyenlő $\text{corr}(y, \hat{y})$-vel. Ez a korrelációs együttható a magyarázó változók és az eredményváltozó közötti kapcsolat erősségét mutatja.

TLR - Túlillesztés elkerülése

Figyelmeztetés a túlillesztésre

Az R^2 értéke mindig növekszik, amikor újabb magyarázó változót adunk a modellhez, még akkor is, ha az új változó nem ad lényeges információt az eredményváltozóról. Ezért körültekintő változóválasztás szükséges, hogy csak valóban releváns tényezők kerüljenek be a modellbe.

Túl sok változó problémája

Ha túl sok irreleváns változót adunk a modellhez, az növeli a modell komplexitását, megnehezítheti az elemzést, és csökkentheti a modell általánosíthatóságát. Cél egy olyan modell kialakítása, amely a legkevesebb változóval magyarázza a lehető legtöbb információt.

Adjusztált R^2

Az adjusztált R^2 szükségessége

Az adjusztált R^2 egy módosított mérőszám, amely figyelembe veszi a modell változóinak számát. Ellentétben a sima R^2-tel, az adjusztált R^2 csökkenthető is, ha egy új prediktor nem járul érdemben hozzá a modellhez, így véd a túlillesztés ellen.

Adjusztált R^2 definíciója

Az adjusztált R^2 a következő képlettel számítható:

$$\overline{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1},$$

ahol:

- n az összes megfigyelés száma,
- k a magyarázó változók száma,
- R^2 a standard R^2 érték.

Adjusztált R^2 alternatív kifejezése

A négyzetösszegek alapján is kifejezhető az adjusztált R^2:

$$\overline{R}^2 = 1 - \frac{\text{SSE} \cdot (n-1)}{\text{SST} \cdot (n-k-1)}.$$

Az adjusztált R^2 tulajdonságai

Az adjusztált R^2 védi a túlillesztéstől

Az adjusztált R^2 csökkenhet, ha egy új változó nem javítja jelentősen a modell magyarázóerejét. A $\frac{n-1}{n-k-1}$ büntető kifejezés figyelembe veszi a modell komplexitását, így véd a túlillesztés ellen.

Adjusztált R^2 jellemzői

- **Összehasonlíthatóság**: Különböző számú prediktorral rendelkező modellek közötti jobb összehasonlítást tesz lehetővé.
- **Értéktartomány**: Általában kisebb vagy egyenlő, mint az R^2, és lehet negatív is, ha a modell gyenge.
- **Védelem**: Csak akkor növekszik, ha az új változók ténylegesen hozzájárulnak a modellhez.

Az adjusztált R^2 alkalmazása

Az adjusztált R^2 felhasználási esetei

- **Modellek összehasonlítása**: A különböző prediktorokkal rendelkező modellek közötti választáshoz.
- **Modellválasztás**: Az optimális prediktorok számának megállapítására, minimalizálva a túlillesztés kockázatát.

Gyakorlati példa

Például lakásárak előrejelzésekor az alapterület, szobák száma stb. alapján újabb változók hozzáadása növelheti az R^2-t, de az adjusztált R^2 csak akkor növekszik, ha ezek ténylegesen javítják az előrejelzést.

Modellépítési technikák - Problémák és megoldások

Változók kiválasztása R^2 és adjusztált R^2 alapján

Az R^2 és az adjusztált R^2 összevetése segít a változók kiválasztásában. Ha egy új változó valóban növeli a modell illeszkedését, az adjusztált R^2 is növekszik, de ha nem jelentős, az adjusztált R^2 csökken. Ezzel a módszerrel elkerülhető a túlillesztés, és biztosítható, hogy csak releváns változók kerüljenek a modellbe.

Modellépítési technikák - Parciális F-próba

A parciális F-próba célja

A parciális F-próbával meghatározható, hogy egy újonnan hozzáadott változó szignifikánsan javítja-e a modell magyarázóerejét. Tegyük fel, hogy egy p-edik változót kívánunk bevonni, és teszteljük, hogy érdemes-e a modellben tartani.

Hipotézisek megfogalmazása a parciális F-próbához

$$H_0: \quad R^2 = R_0^2 \quad \text{(az új változó nem növeli az } R^2\text{-t)}$$
$$H_1: \quad R^2 \neq R_0^2$$

Ha H_0 igaz, azaz az új változó nem jelentős, akkor a következő statisztika Fisher-eloszlást követ:

$$F = \frac{(R^2 - R_0^2)/(p - p_0)}{(1 - R^2)/(n - p - 1)} \sim F_{1,n-p-1}.$$

Ezzel a próbával adott szignifikanciaszinten eldönthetjük, hogy az új változó hozzájárul-e érdemben a modellhez.

Automatikus modellépítési technikák R-ben

Automatikus modellépítési módszerek

Az R program számos automatizált modellépítési módszert kínál, például az ENTER, FORWARD, BACKWARD és STEPWISE eljárásokat, amelyek a magyarázó változók kiválasztására szolgálnak.

ENTER módszer

Az ENTER módszer minden lehetséges változót egyszerre beilleszt a modellbe, nincs előzetes szűrés. Ez gyakran kiindulási alapot nyújt, de előfordulhat, hogy további változóellenőrzést igényel.

FORWARD módszer

A FORWARD módszer alulról építkező technika, amely minden lépésben hozzáadja azt a változót, amelynek F-próbája a legkisebb szignifikancia szinten van. Az eljárás addig folytatódik, amíg a bevonási szignifikancia szint egy előre meghatározott érték alatt marad.

További automatikus modellépítési technikák

BACKWARD módszer

A BACKWARD eljárás felülről lebontó technika: az összes változóval indul, és minden lépésben eltávolítja azt a változót, amely a legkevésbé szignifikáns a parciális F-próba alapján. A folyamat akkor ér véget, amikor az összes változó szignifikanciaszintje egy előre beállított határ alatt van.

STEPWISE módszer

A STEPWISE módszer a FORWARD eljárás módosítása. Minden új változó bevonásakor ellenőrzi a korábban bevont változók szignifikancia szintjét, és eltávolítja azokat, amelyek szignifikanciája meghalad egy küszöbértéket. Ez a ciklikus bevonási és eltávolítási folyamat a stabil modellhez vezethet.

Szokásos paraméterek R-ben

Az R programban általában a következő szignifikancia szinteket használjuk a modellépítéshez:

- PIN $= 0.05$ (bevonási szignifikancia szint),
- POUT $= 0.10$ (elhagyási szignifikancia szint).

Modellépítési technikák R-ben - Előrelépéses kiválasztás (Forward Selection)

Előrelépéses kiválasztás

A FORWARD egy iteratív eljárás, amely kezdetben egy üres modellből indul ki, majd fokozatosan hozzáadja a legjobban illeszkedő változókat. Az R-ben a `step()` függvény használatával végezhetjük ezt az eljárást az `direction = "forward"` paraméter megadásával.

```
# Adatok generálása
set.seed(123)
n <- 100
x1 <- rnorm(n)
x2 <- rnorm(n)
x3 <- rnorm(n)
y <- 2 + 1.5 * x1 + 0.5 * x2 + rnorm(n)

# Üres modell
null_model <- lm(y ~ 1, data = data.frame(y, x1, x2, x3))

# Teljes modell
full_model <- lm(y ~ x1 + x2 + x3, data = data.frame(y, x1, x2, x3))

# Előrelépéses modellépítés
forward_model <- step(null_model, scope = list(lower = null_model, upp
summary(forward_model)
```

Modellépítési technikák R-ben - Előrelépéses kiválasztás (Forward Selection)

R output - Előrelépéses kiválasztás

Az `summary(forward_model)` parancs kimenete megmutatja az előrelépéses eljárással kiválasztott legjobb modellt és az illesztett változókat.

Többi eljárást hasonlóan