Regressziószámítás

Statisztikai elemzéseknél gyakran vetődik fel az a kérdés, hogy sztochasztikus kapcsolat esetén az egyik ismérv (vagy több ismérv) által hordozott információt hogyan tudnánk felhasználni a másik ismérv értékeinek meghatározására. Az összefüggéseket ok(x)-okozati(y) kapcsolattal leíró egyik ilyen módszert **regressziószámítá**snak nevezzük. Ekkor egy egyenlettel adott kapcsolatot létesítünk a változók között, melynek segítségével a magyarázóváltozók (x) alapján becslést adhatunk az eredményváltozóra (y). Például, ha jégkrémet árulunk, akkor egy adott napra az eladott mennyiséget előre becsülhetjük pusztán hasra ütésszerűen is, de akár azt is mondhatjuk, hogy az értékesítést befolyásolja a külső hőmérséklet, és ez alapján becsüljük meg az értékesítést.
A regressziós modellben alkalmazott függvény típusának fontos szerepe van; ez egyszerűbb esetekben lineáris, de az empirikus elemzéseknél gyakran nemlineáris (például, exponenciális, vagy hatványkitevős). Néhány példa függvényalakokra.

Típus	Kétdimenziós modell	Háromdimenziós modell
Lineáris	$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \ x$	$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \ x_1 + \hat{\beta}_2 \ x_2$
Exponenciális	$\hat{y} = \hat{\beta}_0 \hat{\beta}_1^x$	$\hat{y} = \hat{\beta}_0 \hat{\beta}_1^{x_1} \hat{\beta}_2^{x_2}$
Hatványkitevős	$\hat{y} = \hat{\beta}_0 x^{\hat{\beta}_1}$	$\hat{y} = \hat{\beta}_0 x_1^{\hat{\beta}_1} x_2^{\hat{\beta}_2}$

A regressziószámítás során a számítógépes szoftverek egy olyan eljárásra építenek, amelynek célja a regressziós függvény által létrehozott becslések hibatagjának minmalizálása; ezt az eljárást a legkisebb négyzetek módszerének (LNM) nevezzük (angolul OLS – ordinary least squares). A regressziószámítás outputján látható, hogy milyen változók és milyen szerepkörben szerepelnek a modellben.

Regressziószámítás során feltétlenül meg kell vizsgálnunk

- a **többszörös determinációs együttható**t (*R*-square), ami a modell magyarázó erejét adja meg, azaz azt, hogy a magyarázóváltozók együttesen milyen mértékben magyarázzák az eredményváltozó értékeinek különbözőségét. Egy adott modell magyarázóereje általában 80 százalék felett tekinthető elfogadhatónak. Ha egy adott modellbe újabb változókat csatolunk, akkor a modell magyarázóereje biztosan nem csökken. Ez által a magyarázó erő akár 100% közelébe is felhúzható, azonban ettől mindenkit óva intenénk, ugyanis egyrészt túl bonyolulttá teszi a modellt, másrészt sok negatív következménnyel jár. Kétdimenziós modell esetén a mutató az alábbi összefüggéssel határozható meg:

$$r_{y.x}^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST} = r_{yx}^2$$

amely alapja a variancafelbontás, ahol

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

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

$$SST = SSR + SSE$$

- a **többszörös korrelációs együttható**t (*R*), ami a többszörös determinációs együttható gyöke. Megmutatja, hogy a magyarázóváltozók együttese (mint változók halmaza) milyen szoros kapcsolatban áll (mennyire mozog együtt) az eredményváltozóval (y). Kétdimenziós modell esetén a mutató megegyezik a lineáris korrelációs együttható abszolút értékével:

$$r^2 = r_{y.x}^2 = r_{yx}^2$$

$$\sqrt{r_{y.x}^2} = \sqrt{r_{yx}^2} = |r_{yx}|$$

- a **reziduális szórás**t (standard error of the estimates), ami az eredményváltozó tényleges és a modell alapján becsült értékeinek az átlagos eltérését adja meg. Kétdimenziós modell esetén a mutató az alábbi összefüggéssel határozható meg:

$$s_e = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}}$$

- a **regressziós modell illeszkedésének jóságát**. Ez a vizsgálat egy variancianalízis végrehajtását jelenti. **A próba nullhipotézise szerint a modell illeszkedése nem megfelelő, azaz azt hogy a többszörös korrelációs együttható értéke szignifikánsan nem különbözik nullától.** A vizsgálat eredményét egy ANOVA táblázatban kapjuk meg. Fontos megjegyeznünk, hogy a nullhipotézis elvetése csak azt jelenti, hogy a modell alkalmazható a probléma vizsgálatára. (Megjegyzendő, hogy kétdimenziós modell esetén a modell illeszkedésének tesztelése és a korrelációs együttható tesztelése ugyanarra a próbafüggvényre vezethető vissza – a próbák gyakorlatilag ugyanazt a kérdést vizsgálják). A tesztek megkülönböztetésének többdimenziós modellek esetén van relevanciája, azonban a tesztek alkalmazási feltételei könnyen sérülhetnek. Az alkalmazási feltételek kezelése a mesterképzés statisztika kurzusainak a részét képezik.)

- Végre kell hajtanunk a **paraméterek tesztelését**. Ekkor mindegyik magyarázóváltozó fontosságát, magyarázó erejét fogjuk tesztelni külön-külön. **A próba nullhipotézise szerint a vizsgált magyarázóváltozó szignifikánsan nem befolyásolja az eredményváltozót.** Azokat a változókat, amik nincsenek szignifikáns hatással az eredményváltozóra nem érdemes szerepeltetnünk a modellben. Azonban a változók „kidobálásával" óvatosan kell bánni. Ugyanis, ha egy a vizsgálat szempontjából releváns változót nem léptetünk be a modellbe azzal az okkal, hogy ennek hatása statisztikailag nem szignifikáns, akkor torzított modellhez jutunk. Az is – különösen szignifikáns multikollinearitás (magyarázóváltozók nem függetlenek) esetén – lehetséges, hogy egy magyarázóváltozó ereje önmagában nem csekély, de – a modellben levő többi magyarázóváltozó által hordozott megmagyarázott hányadon felül – további többletinformációval nem rendelkezik. Fontos megjegyezni, hogy szignifikáns multikollinearitás esetén a teszt eredményei nem értelmezhetőek, hiszen ekkor nem lehet parciális hatásokról beszélni. Ha csak előrejelzés a célunk, akkor a multikollinearitás nem jelent problémát. (Megjegyzendő a paraméterek tesztelésével kapcsolatban is, hogy kétdimenziós modell esetén a paraméterek tesztelése, a modell illeszkedésének tesztelése és a korrelációs együttható tesztelése ugyanarra a próbafüggvényre vezethető vissza – a próbák gyakorlatilag ugyanazt a kérdést vizsgálják). A tesztek megkülönböztetésének többdimenziós modellek esetén van relevanciája, azonban a tesztek alkalmazási feltételei sérülhetnek. Az alkalmazási feltételek kezelése a mesterképzés statisztika kurzusainak a részét képezik.)
- Miután megállapítottuk, hogy megfelelő a modell magyarázó ereje, a modellben csak megfelelő változók szerepelnek, felírhatjuk a **regressziós modell egyenletét, majd értelmezzük ennek paramétereit**. A regressziós paraméterek megadják, hogy az adott magyarázóváltozó változására – ceteris paribus – átlagosan hogyan változik az eredményváltozó értéke. A pontos értelmezés a modell típusától függ.
    - **Lineáris modell** esetében: az adott magyarázóváltozó (x) értékének 1 egységnyi növekedése esetén az eredményváltozó (y) értéke átlagosan és megközelítőleg $\hat{\beta}_1$ egységgel változik, minden más változatlansága mellett.
    - **Exponenciális modell** esetében: az adott magyarázóváltozó (x) értékének 1 egységnyi növekedése esetén az eredményváltozó (y) értéke átlagosan és megközelítőleg $\hat{\beta}_1$ -szeresére változik, minden más változatlansága mellett
    - **Hatványkitevős modell** esetében: az adott magyarázóváltozó (x) értékének bármilyen szintről történő 1 százalékos növekedése esetén az eredményváltozó (y) értéke átlagosan és megközelítőleg $\hat{\beta}_1$ százalékkal változik, minden más változatlansága mellett.

A fentieken kívül még számos modell diagnosztikai kérdést ajánlatos megvizsgálni (**autokorreláció, heteroszkedaszticitás, multikollinearitás, outlierek**), ezek nem képezik a tananyag részét.

SZEGEDI TUDOMÁNYEGYETEM
GAZDASÁGTUDOMÁNYI KAR
KÖZGAZDÁSZ KÉPZÉS
TÁVOKTATÁSI TAGOZAT
LECKESOROZAT

JELEN TANANYAG
A SZEGEDI TUDOMÁNYEGYETEMEN KÉSZÜLT
AZ EURÓPAI UNIÓ TÁMOGATÁSÁVAL.
PROJEKT AZONOSÍTÓ: EFOP-3.4.3-16-2016-00014

SZÉCHENYI 2020

MAGYARORSZÁG KORMÁNYA

Európai Unió
Európai Szociális Alap

BEFEKTETÉS A JÖVŐBE