Lineáris regresszió

Statisztika I., 4. alkalom

Lineáris regresszió

Ha két folytonos változó lineáris kapcsolatban van egymással, akkor az egyik segítségével előre jelezhetjük a másik értékét. Szükségünk van a függő és független változó kiválasztására, de ez *nem jelent oksági kapcsolatot!* Azt sem jelenti, hogy megértettük volna a kapcsolatot, de az összefüggés segítheti a megértését a kapcsolatnak és legfőképp releváns előrejelzéseink lehetnek.

Példák:

Évszakok váltakozása és az ókori görögök.

- Képességteszt és adott pozícióban való beválás.
- Felvételi vizsgapontszám és egyetemi előmenetel.
- Adott árucikkel szembeni attitűd és vásárlási hajlandóság.
- Kapcsolat szubjektív erőssége és interakciók heti gyakorisága.

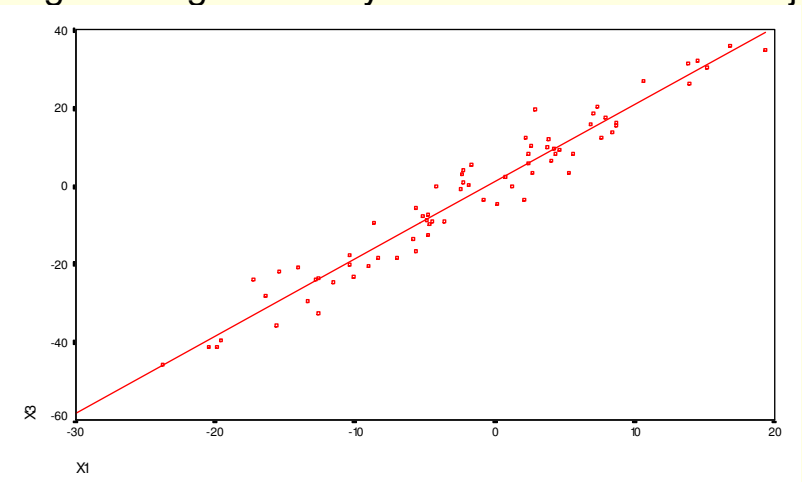
Ha az előrejelzés egy változó segítségével történik, akkor egyváltozós lineáris regresszió számításnak nevezzük az eljárást.

Lineáris regresszió

Minél szorosabb két változó kapcsolata, annál kisebb lesz az előrejelzés hibája.

Pl. Egy varrónővel szembeni elégedettséget szeretnénk előre jelezni egy varrási gyorsasági teszt alapján.

Számos már dolgozó varrónővel elvégezzük a tesztet, és informálódunk főnökük munkájukkal való elégedettségéről. Pl. korrelációval megvizsgáljuk, hogy a számszerűsített két változó összefügg-e egymással. Amennyiben a teszt használhatónak tűnik, regressziós egyenest illesztünk az adatokra, hogy a teszt alapján a főnök elégedettségét bármely teszt érték esetén előre jelezhessük.



Ábra a Máth jegyzetből.

Lineáris regresszió

A lineáris kapcsolat természetesen egy egyenessel ragadható meg a legjobban. Ezt regressziós egyenesnek nevezzük.

Az általános képlete egy egyenesnek::

$$Y = \beta_0 + \beta_1 X$$

β_0 konstans, az a pont ahol az egyenes metszi az y tengelyt, az az érték, ami a legjobb becslés x=0 esetén

β_1 a változó súlya, azt fejezi ki, hogy x egységnyi változása mekkora növekedést idéz elő y becslésében

A becslés csak tökéletes kapcsolat esetén lenne hibamentes (r=1 vagy r= -1).

Az eljárás elnevezésének háttere:

Sir Francis Galton a 19. században kutatta gyermekek genetikus meghatározottságát. Úgy fogalmazta meg eredményeit, hogy a gyermekek magassága a szülők magasságához képest regrediál az átlagosság irányába.

A jelenség generalizálható teszt-reteszt szituációkra, ez is mutatja, hogy a regressziós hatás egy természetes jelenség.

Lineáris regresszió

A becslés csak tökéletes kapcsolat esetén lenne hibamentes (r=1 vagy r= -1).

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

A lehető legkisebb hibájú becslés a cél. A hibáról feltételezzük, hogy független X-től és átlaga nulla.

A négyzetes hiba minimalizálására épülő “legkisebb négyzetek” segítségével számolhatjuk becslését. A becslések normális eloszlásúak, így tesztelhető, hogy nullával egyenlőek-e a populáció szintjén.

$$\beta_1 = \frac{\sum_i (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_i (X_i - \overline{X})^2}$$

$$SE(\beta_1) = \sigma \sqrt{\frac{1}{(N-1)s_x^2}}$$

$$\hat{\sigma} = \frac{\sum_i Y_i - \hat{Y}_i}{N-2}$$

$$\beta_0 = \overline{Y} - \beta_1 \overline{X}$$

$$SE(\beta_0) = \sigma \sqrt{\frac{1}{N} + \frac{\overline{X}^2}{(N-1)s_x^2}}$$

Két változó kapcsolata

Ha két változó normális eloszlású, akkor csak *lineáris kapcsolat* képzelhető el közöttük, azaz, ha nincs közöttük lineáris kapcsolat, akkor függetlenek egymástól.

Ha két változó normális eloszlású és *korrelációjuk* nulla, akkor függetlenek egymástól, ha korrelációjuk nullától különbözik, akkor lineáris kapcsolatban vannak, és ez a kapcsolat egy egyenessel megragadható. Fontos a korreláció mértéke is (r=0.01)

A *regressziós egyenes* segítségével egyik változó értékének ismeretében a másik változó értékét előre jelezhetjük. Meg kell határoznunk a függő és független változót, ki kell számítanunk a *regressziós együtthatókat* (β_0 , és β_1).

Ha a független változó értékei köz nem szerepel egy érték, de a megfigyelt tartományban van (megfigyelt min. és max. érték között), akkor következtethetünk a függő változó értékére (*interpoláció*), ha a megfigyelt tartományon kívül van, akkor ezt nem tehetjük meg (*extrapoláció)*.

Ha a *független változó súlya* (β_1) a populáció szintjén különbözik nullától, akkor a független változó hatása szignifikáns.

Lineáris regresszió

A lineáris regresszió terminológiája megtévesztő:

-függő változó
-független változó
-változó hatása

Csak akkor beszélhetünk oksági kapcsolatról, ha random kísérletből származó adatokkal dolgozunk és minden más, a vizsgált kapcsolat szempontjából releváns, tényezőt kontrollálunk. (A független változót mi manipuláljuk és a személyeket random módon soroltuk a függő változó szerinti csoportokba).

Ha megfigyelésről van szó, számos külső tényező befolyással lehet mind a függő, mind a független változóra, oksági kapcsolatról megfigyelés esetén nem beszélhetünk.

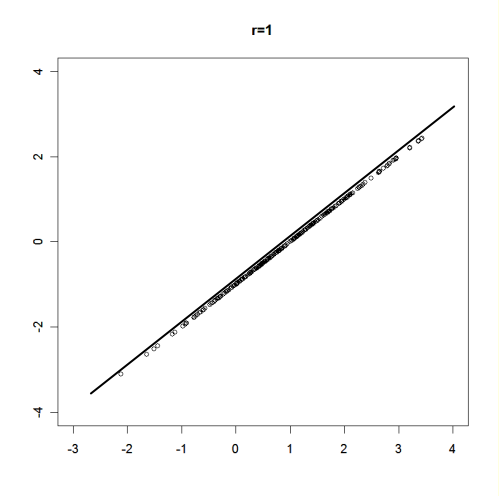
Lineáris regresszió

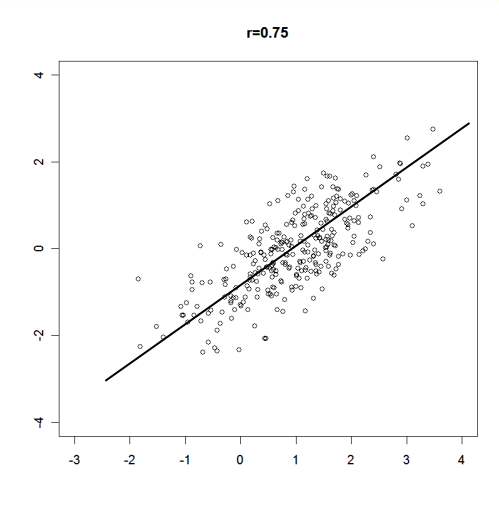
Ha a regresszió tökéletes előrejelzésre ad lehetőséget, azaz a megfigyelt értékek , pontdiagrammon ábrázolva tökéletesen illeszkednek egy egyenesre, akkor szokás *függvénykapcsolatról* beszélni.

Pl. Eladott termék száma, eladásból származó bevétel.

Az esetek döntő többségében azonban csak úgynevezett *statisztikai kapcsolatról* van szó, ahol az előrejelzés nem tökéletes, az előrejelzés hibája vizuálisan a pontok távolsága az illesztett egyenestől.

Pl. az anya intelligenciájával próbáljuk bejósolni a gyermek intelligenciáját.





Lineáris regresszió

Az általános képlete egy egyenesnek:

$$Y_i = \beta_0 + \beta_1 X_i$$

β_0 az a pont ahol az egyenes metszi az y tengelyt

β_1 azt fejezi ki, hogy x egységnyi változása mekkora növekedést idéz elő y-ban

$$\hat{Y}_i = \overline{Y_i}\big|X_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \qquad Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \varepsilon_i$$

$$\varepsilon_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i) = Y_i - \hat{Y}_i$$

A becslés csak tökéletes kapcsolat esetén lenne hibamentes (r=1 vagy r= -1).

ε_i a becslés hibája.

A lehető legkisebb hibájú becslés a cél. A hibáról feltételezzük, hogy független X-től és átlaga nulla. A négyzetes hiba minimalizálására épülő “legkisebb négyzetek”eljárás segítségével számolhatjuk β_0 és β_1 becslését.

Lineáris regresszió

A becslés hibája:

$$\varepsilon_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i) = Y_i - \hat{Y}_i$$

A regressziós egyenes hibája, az ún. reziduális hiba vagy hibavariancia:

$$\text{Res} = E[(Y_i - \hat{Y}_i)^2] = E[\varepsilon_i^2] = E[(\varepsilon_i - 0)^2] = \sigma_{\varepsilon_i}^2$$

$$\sigma_{Y_i}^2 = \sigma_{\hat{Y}_i}^2 + \sigma_{\varepsilon_i}^2$$

$$ss_{Y_i}^2 = ss_{\hat{Y}_i}^2 + ss_{\varepsilon_i}^2$$

$$R^2 = \frac{ss_{\hat{Y}}}{ss_Y}$$

Az R négyzet érték, a determinációs együttható, azt mutatja meg, hogy az Y változó varianciájának mekkora részét tudjuk megragadni az y becsült értékével.

Ez pontosan a korreláció négyzete lesz.