Görbeillesztés

Lineáris regresszió, legkisebb négyzetek módszere

Illés Gergő és Sarkadi Balázs

PTE TTK Fizikai Intézet

2023. március 3.

Tartalom

Lineáris regresszió – Bevezetés

- A lineáris regresszió a leggyakrabban használt görbeillesztési módszer.
- Lineáris regresszió használatakor lineáris kapcsolatot feltételezünk a független- és függő változó között.
- Nemlineáris kapcsolatokra is alkalmazható a függő- és független változók közti kapcsolat linearizálásával.

Lineáris regresszió – Általános eset

Tételezzük fel, hogy rendelkezünk egy adathalmazzal amely n darab statisztikai egységet tartalmaz. Ezt mátrix formájában a következőképp írhatjuk le.

$$\begin{bmatrix} \{y_1, [x_{11}, \ldots, x_{1p}]\} \\ \vdots \\ \{y_n, [x_{n1}, \ldots, x_{np}]\} \end{bmatrix}$$

Ebben az írásmódban y a függő változó, $\vec{x}$ p hosszúságú vektor pedig az úgynevezett regresszor ami a független változókat tartalmazza.

Lineáris regresszió – Általános eset

A becslésünk jóságára vezessünk be egy hibaváltozót, ez legyen ϵ. Lineáris függést, valamint ϵ hibát feltételezve az egyes y_i-k a következőképp írhatók fel:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i = \vec{x_i}\vec{\beta}^T + \epsilon_i, \qquad i = 1, 2, \ldots, n$$

Az egyenletből látszik, hogy itt feltételezünk egy 0-ik x_{i0} elemet, ami minden $\vec{x}$ esetén 1-nek adódik.

Lineáris regresszió – Általános eset

A lineáris egyenletrendszerek témakörben szerzett tudásunk alapján beláthatjuk, hogy amennyiben $\vec{x}_i$-kből mátrixot képzünk, valamint ϵ_i-kből vektort képzünk, egy lineáris egyenletrendszert írhatunk fel mátrixműveletek formájában a következő módon:

$$\vec{y} = \mathbf{X}\vec{\beta} + \vec{\epsilon}.$$

Célunk innentől az ϵ tag „minimalizálása".

Lineáris regresszió – Egy darab független változó esete

- Az eddig tárgyalt általános esetben $\vec{x_i}$ egy $1 \times (p+1)$ méretű vektor volt, azonban 1 darab független változóval dolgozunk az esetek többségében.
- A továbbiakban 1 darab független változóval dolgozunk.
- Ebből adódik, hogy $\mathbf{X}$ egy $n \times 2$ méretű mátrix lesz aminek a második oszlopban lévő elemeit x_i-vel jelöljük.
- A $\vec{\beta}$ együttható vektor pedig 1×2-es vektor lesz, ennek elemit β_i-vel jelöljük.

Lineáris regresszió – Egy darab független változó esete

Definiáljuk $Q(\vec{\beta})$ függvényt az egyes x_i-khez tartozó hibák négyzetösszegeként.

$$
\begin{aligned}
Q(\vec{\beta}) &= \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \\
&= \sum_{i=1}^{n} (y_i^2 + \beta_0^2 + \beta_1^2 x_i^2 - 2y_i\beta_0 - 2y_i\beta_1 x_1 + 2\beta_0\beta_1 x_i)
\end{aligned}
$$

Célunk azon β_0 és β_1 paraméterek megkeresésre amelyre $Q(\vec{\beta})$ függvény minimális értéket vesz fel.

Lineáris regresszió – Egy darab független változó esete

Minimalizáljuk $Q(\vec{\beta})$-t β_0 szerint:

$$\begin{aligned}
\frac{\partial Q(\vec{\beta})}{\partial \beta_0} &= \sum_{i=1}^{n} (2\beta_0 - 2y_i + 2\beta_1 x_i) \\
&= 2n\beta_0 - 2\sum_{i=1}^{n}(y_i) + 2\beta_1 \sum_{i=1}^{n}(x_i) \\
&= \beta_0 - \overline{y} + \beta_1 \overline{x} = 0
\end{aligned}$$

Lineáris regresszió – Egy darab független változó esete

Most tegyük ugyanezt β_1 szerint:

$$\begin{aligned}
\frac{\partial Q(\vec{\beta})}{\partial \beta_1} &= \sum_{i=1}^{n}(2\beta_1 x_i^2 - 2y_i x_i + 2\beta_0 x_i) \\
&= 2\beta_1 \sum_{i=1}^{n}(x_i^2) - 2\sum_{i=1}^{n}(x_i y_i) + 2\beta_0 \sum_{i=1}^{n}(x_i) \\
&= \beta_1 \overline{x^2} - \overline{xy} + \beta_0 \overline{x} = 0
\end{aligned}$$

Lineáris regresszió – Egy darab független változó esete

A minimalizációval kapott lineáris egyenletrendszert megoldva β_0 és β_1 a következőknek adódik:

$$\beta_0 = \frac{\overline{x}\,\overline{xy} - \overline{x^2}\,\overline{y}}{\overline{x}^2 - \overline{x^2}}$$
$$\beta_1 = \frac{\overline{x}\,\overline{y} - \overline{xy}}{\overline{x}^2 - \overline{x^2}}$$

β_0 és β_1 paraméterek segítségével már meghatározhatjuk az adathalmazra négyzetesen legjobban illeszkedő egyenes paramétereit.

Legkisebb négyzetek módszere – Bevezetés

- A legkisebb négyzetek módszerének, lényege az, hogy a modell (magyarázó függvény) paramétereit úgy hangoljuk, hogy a görbe a lehető legjobban illeszkedjen az adathalmazra.
- Az előző fejezetben az egyenes paramétereinek kiszámításakor ugyan így a legkisebb négyzetek módszerét alkalmaztuk.
- Két legkisebb négyzetes módszert különböztetünk meg:
  - Lineáris/közönséges négyzetek: a maradékok a paraméterektől lineárisan függ
  - Nemlineáris négyzetek: a maradékok nemlineárisan függnek a paraméterektől

Legkisebb négyzetek módszere – Általános megoldás

Tegyük fel, hogy mérési eredményeként kaptunk egy x_i és y_i értékekből álló adathalmazt utóbbi tartalmaz némi zajt. Feltételezzük továbbá, hogy a valódi y értékek előállnak egy ismert függvény értékeként, ez legyen $f(x;\mathbf{a})$. A mérési eredményeket ekkor a következőképp fejezhetjük ki:

$$y_i = f(x;\mathbf{a}) + n_i$$

Legkisebb négyzetek módszere – Általános megoldás

Tegyük fel, hogy a hibák függetlenek és normál eloszlásúak $\left(N(0, \sigma_i)\right)$ ekkor bevezethetünk egy mennyiséget:

$$\chi^2(\mathbf{a}) = \sum_{i=1}^{N} \left[\frac{y_i - f(x_i; \mathbf{a})}{\sigma_i} \right]^2$$

Legkisebb négyzetek módszere – Általános megoldás

$\chi^2(\mathbf{a})$ minimalizálásával megkaphatjuk a legjobban illeszkedő görbét. Ezt az a_i paraméterek szerinti deriválással tehetjük meg a következőképp:

$$\left.\frac{\partial \chi^2(\mathbf{a})}{\partial a_i}\right|_{\mathbf{a}=\mathbf{a}_{LS}} = 0 \qquad\qquad i = 1, 2, \ldots, N$$

$\chi^2(\mathbf{a})$-be behelyettesítve a megoldandó egyenlet a következő:

$$\sum_{i=1}^{N} \left.\left((f(x_i;\mathbf{a}) - y_i) \frac{\partial f(x_i;\mathbf{a})}{\partial a_i} \right)\right|_{\mathbf{a}=\mathbf{a}_{LS}} = 0$$