2025. április 23-i gyakorlat

Kétváltozós lineáris és lineárisra visszavezethető regresszió

Kétváltozós lineáris regressziós modell

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

$\beta_0,\ \beta_1$ becslése legkisebb négyzetek módszerével:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} d_{x_i} d_{y_i}}{\sum_{i=1}^{n} d_{x_i}^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - n \cdot \overline{x} \cdot \overline{y}}{\sum_{i=1}^{n} x_i^2 - n(\overline{x})^2}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

Varianciafelbontás a kétváltozós lineáris modellre:

$$SST = SSR + SSE$$

$SST = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum d_y^2$ teljes négyzetösszeg

$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$ belső négyzetösszeg, a hiba okozta (reziduális) négyzetösszeg

$SSR = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$ külső négyzetösszeg, regressziós vagy magyarázott négyzetösszeg

A *determinációs együttható* azt mutatja, hogy a regressziós modellel az y_i adatokban meglévő variancia hány százaléka szüntethető/magyarázható meg:

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

$$R^2_{\text{adjusted}} = 1 - \frac{n-1}{n-k-1}(1 - R^2)$$

$R^2 \approx 1$ - jó illeszkedés, nagy magyarázó erő
$R^2 \approx 0$ - gyenge modellteljesítmény.

A mintából számolt becsült *lineáris korrelációs együttható* a magyarázó- és eredményváltozó között:

$$r = \frac{\sum d_x d_y}{\sqrt{\sum d_x^2 \cdot \sum d_y^2}} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \cdot \sum_{i=1}^{n}(y_i - \overline{y})^2}} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right) \cdot \left(\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\right)}} = \hat{\beta}_1 \frac{s_x}{s_y}$$

CSAK kétváltozós lineáris esetben: $R^2 = r^2$.

Az *elaszticitás* (rugalmasság) azt méri, hogy az X változó 1%-os növekedése hány százalékos növekedést/csökkenést eredményez az Y változónál. Az elaszticitás kiszámítása a becsült eredményváltozóra:

$$El(\hat{y}, x) = \frac{\partial \hat{y}}{\partial x} \cdot \frac{x}{\hat{y}}$$

Kétváltozós lineáris esetben:

$$El(\hat{y}, x) = \hat{\beta}_1 \cdot \frac{x}{\hat{y}} = \frac{\hat{\beta}_1 x}{\hat{\beta}_0 + \hat{\beta}_1 x}$$

Intervallumbecslés a függvényértékekre:

- az átlagos értékre

$$Int_{1-\alpha}(E(Y_*)) = \hat{y}_* \pm t_{1-\frac{\alpha}{2}}(n-2)s_e\sqrt{\frac{1}{n} + \frac{(x_* - \overline{x})^2}{\sum d_x^2}},$$

- az egyedi értékre

$$Int_{1-\alpha}(Y_*) = \hat{y}_* \pm t_{1-\frac{\alpha}{2}}(n-2)s_e\sqrt{1 + \frac{1}{n} + \frac{(x_* - \overline{x})^2}{\sum d_x^2}},$$

ahol $s_e = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}}$ a *korrigált reziduális szórás.*

1. A hektáronkénti szőlőtermést befolyásolja az évenkénti permetezések száma. Az alábbi táblázat öt év adatait mutatja:

Permetezések száma (db)	Szőlőtermés (q)
5	10
6	10
7	12
7	10
5	8
$\sum$	$\sum$

   (a) Határozza meg és értelmezze a lineáris regresszió paramétereit!

	$d_x = x_i - \overline{x}$	$d_y = y_i - \overline{y}$	d_x^2	d_y^2	$d_x d_y$	$\hat{y}$	$e_i = y_i - \hat{y}$
1							
2							
3							
4							
5							
$\sum$							

   (b) Számítsa ki a két változó lineáris korrelációs együtthatóját!

   (c) Számítsa ki és értelmezze a lineáris regresszió determinációs együtthatóját!

   (d) Számítsa ki és értelmezze a korrigált reziduális szórást!

   (e) Számolja ki és értelmezze az elaszticitást az átlagos permetezésszámra!

   (f) Adjon becslést 6 permetezés esetén a szőlőtermés átlagos mennyiségére, majd szerkesszen konfidenciaintervallumot ugyanerre 95%-os megbízhatósági szinten!

   (g) Adjon 95%-os konfidenciaintervallumot egy 6-szor permetezett szőlőültetvény szőlőtermésére!

2. Egy bank 10 ügyfelét vizsgálva az életkor (X, év) és a havi jövedelem (Y, eFt-ban) kapcsolatát elemzi. Az alábbi részeredményeket kapták:

$$\sum_{i=1}^{n} x_i = 260, \qquad \sum_{i=1}^{n} y_i = 2\,040, \qquad \sum_{i=1}^{n} x_i y_i = 53\,754,$$

$$\sum_{i=1}^{n} x_i^2 = 6\,862, \qquad \sum_{i=1}^{n} y_i^2 = 421\,866, \qquad \sum_{i=1}^{n} e_i^2 = 708,$$

$$\sum_{i=1}^{n} \log x_i = 50, \qquad \sum_{i=1}^{n} \log y_i = 90, \qquad \sum_{i=1}^{n} x_i \log y_i = 675.$$

(a) Határozza meg és értelmezze a lineáris regresszió paramétereit!

(b) Számítsa ki a két változó lineáris korrelációs együtthatóját!

(c) Számítsa ki és értelmezze a lineáris regresszió determinációs együtthatóját!

(d) Adjon becslést a 30 évesek átlagos jövedelmére, majd szerkesszen konfidenciaintervallumot ugyanerre 98%-os megbízhatósági szinten!

(e) Adjon 98%-os konfidenciaintervallumot egy 30 éves ügyfél egyedi jövedelmére!

(f) Számolja ki és értelmezze az elaszticitást az átlagos életkorra!

3. Egy taxivállalat 15 véletlenszerűen kiválasztott fuvar alapján vizsgálja, hogy hogyan függ a menetidő a távolságtól (megtett km-től). A 15 fuvar esetén a távolság és a menetidő:

távolság (km)	menetidő (perc)	távolság (km)	menetidő (perc)
3	8	9	20
4	19	12	23
4	13	15	44
6	21	16	47
6	11	16	41
7	19	20	46
8	14	26	48
8	19		

$$\sum y = 393, \qquad \sum xy = 5\,433, \qquad \sum x \ln y = 545.8033,$$

$$\sum \ln y = 46.7381, \qquad \sum x = 160, \qquad \sum x^2 = 2\,328,$$

$$\sum \ln x \ln y = 106.6887, \qquad \sum (\ln x)^2 = 77.2063, \qquad \sum \ln x = 32.7487.$$

(a) Jellemezze a távolság és a menetidő közötti lineáris, exponenciális illetve hatvány kapcsolatot és értelmezze a paramétereket!

(b) Becsülje meg mindegyik modell alapján, hogy egy 15 km távolságú út hány percet vesz igénybe!

(c) Mekkora az elaszticitás az átlagos és a 15 km távolságú fuvar környezetében?

SPSS: Graphs ⇒ Scatter/Dot

Analyze ⇒ Regression ⇒ Curve Estimation: Linear, Compound, Power

Transform ⇒ Compute variable: becslés, hiba

Analyze ⇒ Regression ⇒ Linear