

PROSTA REGRESJA LINIOWA

Termin "regresja" i metody badania związków pomiędzy dwoma zmiennymi pochodzą sprzed ponad 100 lat. Termin regresja po raz pierwszy wprowadzony został przez Francisa Galtona w roku 1908. Galton, znany brytyjski biologa, propagator metod ilościowych w nauce oraz kuzynem Karola Darwina był zaangażowany w badania nad dziedzicznością. Na podstawie danych ilościowych zaobserwował, że dzieci wysokich rodziców są wyższe niż przeciętne, ale nie tak wysokie jak ich rodzice. Termin "regres w kierunku średniej" dał asumpt to nazwania metody badania tego typu zależności regresją. Tak więc termin regresja opisują przede wszystkim relacje statystyczne między zmiennymi. W szczególności prosta regresja to metoda regresji służąca do opisanie zależności między jedną zalezną zmienną (y) i jedną zmienną niezależną (x). Następujące klasyczne dane zebrane przez Galtona zawierają informacje o wzroście rodzica i wzroście jego dziecka.

Parent	64.5	65.5	66.5	67.5	68.5	69.5	70.5	71.5	72.5
Children	65.8	66.7	67.2	67.6	68.2	68.9	69.5	69.9	72.2

Średnia wysokość wynosi 68,44 dla dzieci i 68,5 dla rodziców. Korzystając z metody regresji liniowej zależność wzrostu dziecka od wzrostu rodziców można opisać jako:

$$\text{wzrost dziecka} = 21,52 + 0,69 \text{ wzrost rodzica.}$$

Prosty model regresji liniowej jest zwykle zapisany w formie

$$y = \beta_0 + \beta_1 x + \epsilon,$$

gdzie y jest zmienną zależną, β_0 i β_1 są parametrami modelu, x jest zmienną niezależną natomiast ϵ jest błędem losowym. Zmienna zależna jest również nazywana zmienną odpowiedzi a zmienna niezależna nazywana jest zmienną objaśniającą lub predykcyjną. Model regresji można przedstawić w bardziej ogólnej postaci:

$$y = E(y) + \epsilon,$$

gdzie $E(y)$ jest wartością oczekiwaną zmiennej odpowiedzi ϵ jest błędem losowym. Gdy $E(y)$ jest liniową kombinacją zmiennych objaśniających $x_1, x_2, \dots, x_k$ to otrzymujemy model regresji liniowej. Jeśli $k = 1$, to otrzymujemy prostą regresję liniową. Jeśli $E(y)$ jest nieliniową funkcją $x_1, x_2, \dots, x_k$ to otrzymamy model regresji nieliniowej. Klasyczne założenia dotyczące błędu losowego to $E(y) = 0$ i stała wariancja $\text{Var}(\epsilon) = \sigma^2$. Typowy eksperyment dla prostej regresji liniowej polega na tym, że obserwujemy n par danych $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ na podstawie eksperymentu naukowego i modele otrzymane na podstawie tych danych można zapisać jako

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \text{dla } i = 1, 2, \dots, n,$$

gdzie $E(\epsilon_i) = 0$ i $\text{Var}(\epsilon) = \sigma^2$, a zmienne losowe ϵ_i są niezależne. Zauważmy, że wartość σ^2 jest zwykle nieznana. Zakładamy, że wartości x_i są mierzone "dokładnie" a więc bez błędu pomiarowego. Po określeniu postaci modelu i zebraniu danych, następnym krokiem jest estymacja parametrów β_0 i β_1 dla prostego modelu regresji liniowej, przy których model może najlepiej opisać dane pochodzące z eksperymentu naukowego. Postać estymatorów parametrów modelu wyprowadzimy i omówimy ich właściwości statystyczne w dalszej części wykładu.

1 Metoda najmniejszych kwadratów.

Zasada najmniejszych kwadratów dla prostego modelu regresji liniowej proponuje oszacować wartości parametrów b_0 i b_1 w taki sposób, aby suma kwadratu odległości przewidywanej przez model wartości $y_i = \beta_0 + \beta_1 x_i$ od rzeczywistej odpowiedzi y_i osiąga minimum spośród wszystkich możliwych wyborów współczynników regresji β_0 i β_1 .

$$(\beta_0, \beta_1) = \arg \min_{(\beta_0, \beta_1)} \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)]^2.$$

Korzystając z analizy matematycznej zagadnienie wyznaczenia metodą najmniejszych kwadratów wartości parametrów modelu prostej regresji liniowej można sprowadzić do rozwiązania następującego układu równań:

$$\frac{\partial}{\partial \beta_0} \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)]^2 = 0, \quad (1)$$

$$\frac{\partial}{\partial \beta_1} \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)]^2 = 0. \quad (2)$$

$$(3)$$

Zauważmy, że wygodniej jest rozwiązać zagadnienie estymacji parametrów regresji liniowej w przypadku scentrowanego modelu liniowego postaci:

$$y_i = \beta_0^* + \beta_1(x_i - \bar{x}) + \epsilon,$$

gdzie $\beta_0 = \beta_0^* - \beta_1 \bar{x}$. W przypadku tego scentrowanego modelu należy rozwiązać układ

$$\frac{\partial}{\partial \beta_0^*} \sum_{i=1}^n [y_i - (\beta_0^* + \beta_1(x_i - \bar{x}))]^2 = 0, \quad (4)$$

$$\frac{\partial}{\partial \beta_1} \sum_{i=1}^n [y_i - (\beta_0^* + \beta_1(x_i - \bar{x}))]^2 = 0. \quad (5)$$

Różniczkując względem β_0 i β_1 otrzymamy

$$\sum_{i=1}^n [y_i - (\beta_0^* + \beta_1(x_i - \bar{x}))] = 0, \quad (6)$$

$$\sum_{i=1}^n [y_i - (\beta_0^* + \beta_1(x_i - \bar{x}))](x_i - \bar{x}) = 0. \quad (7)$$

Zauważmy, że

$$\sum_{i=1}^n y_i = n\beta_0^* + \sum_{i=1}^n \beta_1(x_i - \bar{x}) = n\beta_0^* \quad (8)$$

Tak więc

$$\beta_0^* = \frac{1}{n} \sum_{i=1}^n y_i = \bar{y}.$$

Podstawiając $\bar{y}$ w (8) za β_0^* otrzymamy

$$\sum_{i=1}^n [y_i - (\bar{y} + \beta_1(x_i - \bar{x}))] (x_i - \bar{x}) = 0.$$

Oznaczając przez b_0 i b_1 rozwiązanie układu (1) i (2) mamy

$$b_1 = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \quad (9)$$

oars

$$b_0 = b_0^* - b_1 \bar{x} = \bar{y}. \quad (10)$$

Dopasowana na podstawie modelu prostej regresji liniowej wartość y_i będącej odpowiedzią na wartość zmiennej x_i jest definiowana jako $\hat{y}_i = b_0 + b_1 x_i$. Różnicę między y_i a dopasowaną na podstawie modelu wartością $\hat{y}_i$, $e_i = y_i - \hat{y}_i$, nazywamy reszduem regresji, wartością resztkową. Wartości resztowe regresji odgrywają ważną rolę w analizie jakości dopasowania oraz testowaniu modelu regresji. Zostanie to przedyskutowane w dalszej części wykładu. W odróżnieniu od błędów ϵ_i , które są nieobserwowalne, residua można obliczyć na podstawie zaobserwowanych odpowiedzi y_i i wyznaczonych wartości $\hat{y}_i$ a więc są obserwowalne.

2 Właściwości statystyczne oszacowania najmniejszych kwadratów.

W tej części wykładu omówimy właściwości statystyczne estymatorów modelu regresji liniowej otrzymanych metodą najmniejszych kwadratów. Najpierw omawiamy właściwości przy następujących założeniach na błąd. Przyjmijmy, że $E(\epsilon_i) = 0$ i $\text{Var}(\epsilon_i) = \sigma^2$, oraz ϵ_i dla $i = 1, 2, \dots, n$ są niezależnymi zmiennymi losowymi.

Twierdzenie 1. *Estymator najmniejszych kwadratów b_0 jest nieobciążonym estymatorem β_0 .*

Dowód.

$$\begin{aligned} Eb_0 &= E(\bar{y} - b_1 \bar{x}) = E\left(\frac{1}{n} \sum_{i=1}^n y_i\right) - Eb_1 \bar{x} = \frac{1}{n} \sum_{i=1}^n Ey_i - \bar{x} Eb_1 = \\ &= \frac{1}{n} \sum_{i=1}^n (\beta_0 + \beta_1 x_i) - \beta_1 \bar{x} = \frac{1}{n} \sum_{i=1}^n \beta_0 + \beta_1 \frac{1}{n} \sum_{i=1}^n x_i - \beta_1 \bar{x} = \beta_0. \end{aligned}$$

Twierdzenie 2. *Estymator najmniejszych kwadratów b_1 jest nieobciążonym estymatorem parametru β_1 .*

Dowód.

$$E(b_1) = E\left(\frac{S_{xy}}{S_{xx}}\right) \quad (11)$$

$$= \frac{1}{S_{xx}} E \sum_{i=1}^n (y_i - \hat{y})(x_i - \hat{x}) \quad (12)$$

$$= \frac{1}{S_{xx}} \frac{1}{n} \sum_{i=1}^n (x_i - \hat{x}) E y_i \quad (13)$$

$$= \frac{1}{S_{xx}} \frac{1}{n} \sum_{i=1}^n (x_i - \hat{x}) (\beta_0 + \beta_1 x_i) \quad (14)$$

$$= \frac{1}{S_{xx}} \frac{1}{n} \sum_{i=1}^n (x_i - \hat{x}) \beta_1 x_i \quad (15)$$

$$= \frac{1}{S_{xx}} \frac{1}{n} \sum_{i=1}^n (x_i - \hat{x}) \beta_1 (x_i - \hat{x}) \quad (16)$$

$$= \frac{1}{S_{xx}} \frac{1}{n} \sum_{i=1}^n (x_i - \hat{x})^2 \beta_1 = \frac{S_{xx}}{S_{xx}} \beta_1 = \beta_1. \quad (17)$$

■

Twierdzenie 3. *Wariancja estymatora b_1 parametru β_1 ma następującą postać*

$$\text{Var}(b_1) = \frac{\sigma^2}{nS_{xx}}.$$

Dowód.

$$\text{Var}(b_1) = \text{Var}\left(\frac{S_{xy}}{S_{xx}}\right) \quad (18)$$

$$= \frac{1}{S_{xx}^2} \text{Var}\left(\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})\right) \quad (19)$$

$$= \frac{1}{S_{xx}^2} \text{Var}\left(\frac{1}{n} \sum_{i=1}^n y_i (x_i - \bar{x})\right) \quad (20)$$

$$= \frac{1}{S_{xx}^2} \frac{1}{n^2} \sum_{i=1}^n (x_i - \bar{x})^2 \text{Var}(y_i) \quad (21)$$

$$= \frac{1}{S_{xx}^2} \frac{1}{n^2} \sum_{i=1}^n (x_i - \bar{x})^2 \sigma^2 = \frac{\sigma^2}{nS_{xx}} \quad (22)$$

Twierdzenie 4. *Estymator najmniejszych kwadratów b_1 oraz $\bar{y}$ są nieskorelowane. Przy dodatkowym założeniu, że y_i dla $i = 1, 2, \dots, n$ mają rozkład normalny, estymator najmniejszych kwadratów b_1 oraz $\bar{y}$ mają rozkłady normalne i są niezależne.*

Dowód.

$$\text{Cov}(b_1, \hat{y}) = \text{Cov}\left(\frac{S_{xy}}{S_{xx}}, \bar{y}\right) \quad (23)$$

$$= \frac{1}{S_{xx}} \text{Cov}(S_{xy}, \bar{y}) \quad (24)$$

$$= \frac{1}{nS_{xx}} \text{Cov}\left(\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}), \bar{y}\right) \quad (25)$$

$$= \frac{1}{nS_{xx}} \text{Cov} \left(\sum_{i=1}^n (x_i - \bar{x}) y_i, \bar{y} \right) \quad (26)$$

$$= \frac{1}{n^2 S_{xx}} \text{Cov} \left(\sum_{i=1}^n (x_i - \bar{x}) y_i, \sum_{i=1}^n y_i \right) \quad (27)$$

$$= \frac{1}{n^2 S_{xx}} \sum_{i=1}^n (x_i - \bar{x}) \text{Cov}(y_i, y_j). \quad (28)$$

Ponieważ $E\epsilon_i = 0$ oraz ϵ_i są niezależne otrzymujemy

$$\text{Cov}(y_i, y_j) = E[(y_i - Ey_i)(y_j - Ey_j)] = \text{Cov}(\epsilon_i, \epsilon_j) = \begin{cases} \sigma^2, & \text{gdy } i = j \\ 0, & \text{gdy } i \neq j. \end{cases}$$

Tak więc

$$\text{Cov}(b_1, \bar{y}) = \frac{1}{n^2 S_{xx}} \sum_{i=1}^n (x_i - \bar{x}) \sigma^2 = 0.$$

Niezależność b_1 i $\bar{y}$ wynika teraz z faktu, że nieskorelowane zmienne losowe normalne są niezależne.

Twierdzenie 5. *Wariancja estymatora b_0 parametru β_0 ma następującą postać*

$$\text{Var}(b_0) = \left(\frac{1}{n} + \frac{\bar{x}^2}{nS_{xx}} \right) \sigma^2.$$

Dowód.

$$\text{Var}(b_0) = \text{Var}(\bar{y} - b_1 \bar{x}) \quad (29)$$

$$= \text{Var}(\bar{y}) + (\bar{x})^2 \text{Var}(b_1) \quad (30)$$

$$= \frac{\sigma^2}{n} + (\bar{x})^2 \frac{\sigma^2}{nS_{xx}} \quad (31)$$

$$= \left(\frac{1}{n} + \frac{(\bar{x})^2}{\sigma^2 nS_{xx}} \right) \sigma^2. \quad (32)$$

Udowodnione właściwości 1-5 są istotne dla analizy modelu regresji liniowej. Zwłaszcza postać wariancje parametrów b_0 i b_1 jest istotna. Zauważmy, że wzór na wariancję estymatorów najmniejszych kwadratów b_0 i b_1 zawiera wariancję błędu. Wariancja błędu jest zwykle nieznana. Dlatego niezbędne jest oszacowanie wariancji błędu losowego w modelu regresji liniowej.

Niech y_i będzie obserwowaną zmienną odpowiedzi, a $\hat{y}_i = b_0 + b_1 x_i$, wartością odpowiedzi otrzymaną na podstawie dopasowanego metodą najmniejszych kwadratów. Zarówno y_i , jak i $\hat{y}_i$ są dla nas znane. Prawdziwa wartość σ^2 jest nieobserwowana, a więc wymagane jest dokonanie estymacji tego parametru modelu. Wielkość $y_i - \hat{y}_i$ możemy uznać za j empiryczną wartość błędu ϵ_i i za estymator wariancji błędu proponuje się przyjąć:

$$s^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

Zauważ, że w mianowniku jest $n-2$. Dzięki takiej normalizacji s^2 jest nieobciążonym estymatorem σ^2 , wariancji błędu.

3 Estymacja największej wiarygodności.

Estymacji metodą największej wiarygodności modelu regresji liniowej dokonamy przy dodatkowym założeniu, że zmienna zależna y_i ma rozkład normalny: $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$. Funkcja wiarygodności dla $(y_1, y_2, \dots, y_n)$ ma przy tym założeniu następującą postać:

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^n f(y_i) = \frac{1}{(2\pi)^{n/2} \sigma^n} \exp \left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \right).$$

Wartości parametrów β_0 i β_1 , które maksymalizują funkcję wiarygodności L , są takie same jak odpowiednie wartości, które minimalizują wykładniczą część funkcji wiarygodności, a więc są tymi samymi wartościami, które są rozwiązaniem zagadnienia najmniejszych kwadratów dla regresji liniowej. Zatem przy założeniu normalności estymatory wyznaczone metodą najmniejszych kwadratów i metodą największej wiarygodności są takie same. Po wyznaczeniu estymatorów b_1 i b_0 , możemy zapisać funkcję wiarygodności w terminach dopasowanych na podstawie modelu regresji liniowej wartości zmiennych odpowiedzi $\hat{y}_i$.

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^n f(y_i) = \frac{1}{(2\pi)^{n/2} \sigma^n} \exp \left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \right).$$

Biorąc pochodną cząstkową względem σ^2 logarytmu funkcji wiarygodności i przyrównując ją do zera otrzymujemy:

$$\frac{\partial \log(L)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = 0.$$

Estymator największej wiarygodności dla σ^2 jest więc postaci:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

Zauważ, że jest to estymator obciążony. Biorąc pod uwagę postać estymatora nieobciążonego

$$s^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2,$$

otrzymujemy, że $\frac{n}{n-2} \hat{\sigma}^2$ jest nieobciążonym estymatorem σ^2 . Dodatkowo zauważmy, że $\hat{\sigma}^2$ jest asymptotycznie nieobciążonym estymatorem, co jest zgodne z klasycznym rezultatem teorii estymatorów największej wiarygodności.

Krzysztof Topolski