



Analiza regresji

Statystyka i analiza danych 2019

Jurek Błaszczński,
na podstawie slajdów Wojtka Kotłowskiego
11 maja 2019

Rozkład zmienności Y

Na danych $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$ wyznaczono współczynniki regresji a, b metodą najmniejszych kwadratów.

Przypomnienie: $\hat{Y}_i = aX_i + b$.

Zachodzi:

$$\underbrace{\sum_{i=1}^n (Y_i - \bar{Y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}_{\text{SSR}} + \underbrace{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}_{\text{SSE}}$$

- **SST**: całkowita suma kwadratów odchyień – całkowita zmienność Y .
- **SSR**: regresyjna s.k.o. – część zmienności wyjaśniona przez model liniowy.
- **SSE**: resztowa s.k.o. – część zmienności nie wyjaśniona przez model liniowy.

Dowód rozkładu

$$\begin{aligned} \text{SST} &= \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 \\ &= \underbrace{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}_{\text{SSE}} + \underbrace{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}_{\text{SSR}} + 2 \sum_{i=1}^n (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) \end{aligned}$$

Wystarczy pokazać, że ostatni człon znika. Z poprzedniej prezentacji wynika, że $b = \bar{Y} - a\bar{X}$, a stąd:

$$\hat{Y}_i - \bar{Y} = aX_i + b - \bar{Y} = a(X_i - \bar{X})$$

$$Y_i - \hat{Y}_i = Y_i - aX_i - b = Y_i - \bar{Y} - a(X_i - \bar{X}).$$

Używając powyższego i definicji (z poprzedniej prezentacji) $a = \frac{s_{XY}}{s_X^2}$:

$$\begin{aligned} \sum_{i=1}^n (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) &= \sum_{i=1}^n a(Y_i - \bar{Y})(X_i - \bar{X}) - a^2 \sum_{i=1}^n (X_i - \bar{X})^2 \\ &= (n-1)a(s_{XY} - as_X^2) = (n-1)a\left(s_{XY} - \frac{s_{XY}}{s_X^2} s_X^2\right) = 0. \end{aligned}$$

Dlaczego SSR to część wyjaśniona przez model liniowy?

- Weźmy sytuację, w której wszystkie punkty leżą na prostej (idealna zależność liniowa). Wtedy $\hat{Y}_i = Y_i$ i

$$\text{SSE} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = 0,$$

a więc $\text{SST} = \text{SSR}$.

Dlaczego SSE to część niewyjaśniona przez model liniowy?

- Weźmy sytuację, w której brak jakiegokolwiek trendu liniowego ($a = 0$). Wtedy:

$$\text{SSR} = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = \sum_{i=1}^n (aX_i + b - \bar{Y})^2 = n(b - \bar{Y})^2.$$

Ponieważ $b = \bar{Y} - a\bar{X} = \bar{Y}$, mamy $\text{SSR} = 0$, a więc $\text{SST} = \text{SSE}$.

Współczynnik determinacji

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

Cześć zmienności Y wyjaśnionej przez model liniowy.

R^2 jest **kwadratem współczynnika korelacji**. Używając $a = r \frac{s_Y}{s_X}$ oraz $b = \bar{Y} - a\bar{X}$:

$$\begin{aligned} R^2 &= \frac{SSR}{SST} = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \frac{\sum_{i=1}^n (aX_i - b - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \\ &= \frac{\sum_{i=1}^n a^2 (X_i - \bar{X})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = a^2 \frac{s_X^2}{s_Y^2} = r^2 \frac{\cancel{s_Y^2} s_X^2}{\cancel{s_X^2} \cancel{s_Y^2}} = r^2. \end{aligned}$$

- Układ hipotez:

$$\underline{H_0 : a = 0}$$

$$\underline{H_1 : a \neq 0}$$

- Statystyka testowa:

$$F = \frac{SSR}{SSE}(n-2) \sim F(1, n-2),$$

gdzie $F(k, m)$ to rozkład F Snedecora o k i m stopniach swobody.

Istotność regresji vs. istotność korelacji

Istotność regresji

$$H_0 : a = 0$$

$$H_1 : a \neq 0$$

$$F = \frac{SSR}{SSE} (n - 2)$$

Istotność korelacji

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

$$T = \frac{r}{\sqrt{1 - r^2}} \sqrt{n - 2}$$

Ale $a = r \frac{s_Y}{s_X}$, więc $a = 0 \iff r = 0 \dots ?$

Istotność regresji vs. istotność korelacji

Istotność regresji

$$H_0 : a = 0$$

$$H_1 : a \neq 0$$

$$F = \frac{SSR}{SSE}(n-2)$$

Istotność korelacji

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

$$T = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$$

Ale $a = r \frac{s_Y}{s_X}$, więc $a = 0 \iff r = 0 \dots ?$

Jest to w zasadzie ten sam test:

$$T^2 = \frac{r^2}{1-r^2}(n-2) = \frac{\frac{SSR}{SST}}{\frac{SSE}{SST}}(n-2) = \frac{SSR}{SSE}(n-2) = F$$

Ta równość nie zachodzi dla **wielorakiej regresji**.

- Błąd standardowy oszacowania:

$$S = \sqrt{\frac{\text{SSE}}{n-2}}$$

- Błędy standardowe parametrów a i b :

$$s_a = \frac{S}{s_X} \sqrt{n-1}$$

$$s_b = S \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{s_X^2} (n-1)}$$

Globalny test na istotność regresji wielorakiej

Model liniowy z m zmiennymi objaśniającymi:

$$\hat{Y} = a_0 + \sum_{i=1}^m a_i X_i$$

- Układ hipotez:

$$H_0 : a_1 = a_2 = \dots = a_m = 0$$

$$H_1 : \text{Co najmniej jeden } a_i \neq 0$$

- Statystyka testowa:

$$F = \frac{SSR/m}{SSE/(n-m-1)} \sim F(m, n-m-1).$$

Uwaga: wyraz wolny nigdy nie wchodzi do układu hipotez!

Test pojedynczego parametru w regresji wielorakiej

- Układ hipotez:

$$\underline{H_0 : a_i = 0}$$

$$\underline{H_1 : a_i \neq 0}$$

- Statystyka testowa:

$$T = \frac{a_i}{S_{a_i}} \sim t(n - m - 1)$$

W przypadku prostej regresji liniowej ($m = 1$), jest to ten sam test, co na istotność współczynnika korelacji.