Analiza regresji

Statystyka i analiza danych 2017/2018

Jurek Błaszczyński,
na podstawie slajdów Wojtka Kotłowskiego

20 maja 2018

Rozkład zmienności Y

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

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

Zachodzi:

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

- **SST**: całkowita suma kwadartów odchyleń – 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

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

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

$$\begin{aligned}
\widehat{Y}_i - \overline{Y} &= aX_i + b - \overline{Y} = a(X_i - \overline{X}) \\
Y_i - \widehat{Y}_i &= Y_i - aX_i - b = Y_i - \overline{Y} - a(X_i - \overline{X}).
\end{aligned}$$

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

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

SSR i SSE

Dlaczego SSR to część wyjaśniona przez model liniowy?

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

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

  a więc $\mathrm{SST} = \mathrm{SSR}$.

Dlaczego SSE to część niewyjaśniona przez model liniowy?

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

$$\mathrm{SSR} = \sum_{i=1}^{n} (\widehat{Y}_i - \overline{Y})^2 = \sum_{i=1}^{n} (aX_i + b - \overline{Y})^2 = n(b - \overline{Y})^2.$$

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

Współczynnik determinacji

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

Cześć zmienności Y wyjaśnionej przez model liniowy.

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

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = \frac{\sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} = \frac{\sum_{i=1}^{n}(aX_i - b - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}$$

$$= \frac{\sum_{i=1}^{n} a^2(X_i - \overline{X}_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} = a^2\frac{s_X^2}{s_Y^2} = r^2\frac{\cancel{s_Y^2}}{\cancel{s_X^2}}\frac{\cancel{s_X^2}}{\cancel{s_Y^2}} = r^2.$$

Test na istotność regresji

- **Układ hipotez**:

$$H_0: \quad a = 0$$
$$H_1: \quad a \neq 0$$

- **Statystyka testowa**:

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

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

Istotność regresji vs. istotność korelacji

Istotność regresji

$$H_0: \quad a = 0$$
$$H_1: \quad a \neq 0$$

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

Istotność korelacji

$$H_0: \quad \rho = 0$$
$$H_1: \quad \rho \neq 0$$

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

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

Istotność regresji vs. istotność korelacji

Istotność regresji

$$\begin{aligned} H_0 &: \quad a = 0 \\ H_1 &: \quad a \neq 0 \end{aligned}$$

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

Istotność korelacji

$$\begin{aligned} H_0 &: \quad \rho = 0 \\ H_1 &: \quad \rho \neq 0 \end{aligned}$$

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

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

Jest to w zasadzie ten sam test:

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

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

Pozostałe współczynniki

- Błąd standardowy oszacowania:

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

- Błędy standardowe parametrów a i b:

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

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

Globalny test na istotność regresji wielorakiej

Model liniowy z m zmiennymi objaśniającymi:

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

- **Układ hipotez**:

H_0 :	$a_1 = a_2 = \ldots = a_m = 0$
H_1 :	Co najmniej jeden $a_i \neq 0$

- **Statystyka testowa**:

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

Uwaga: wyraz wolny nigdy nie wchodzi do układu hipotez!

Test pojedynczego parametru w regresji wielorakiej

- **Układ hipotez**:

$$\begin{array}{ll} \hline H_0 : & a_i = 0 \\ H_1 : & a_i \neq 0 \\ \hline \end{array}$$

- **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.