Analiza regresji

Wojciech Kotłowski

Statystyka i analiza danych 2019/2020

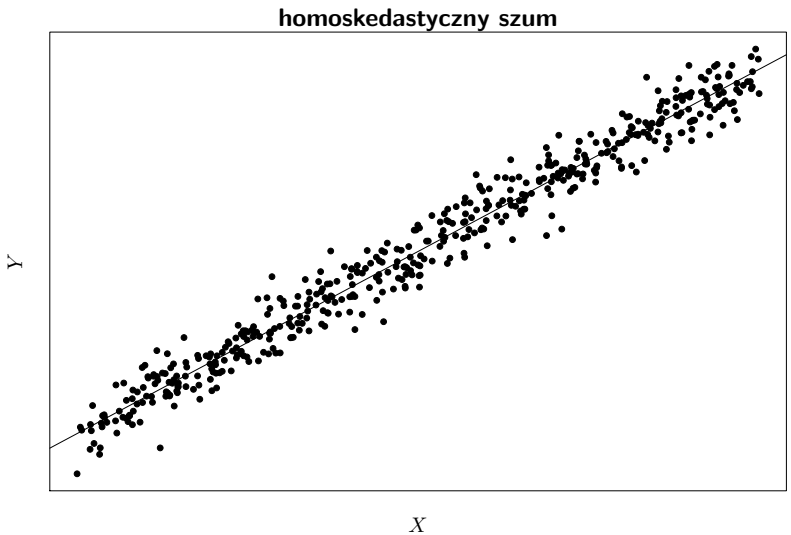
12.05.2020

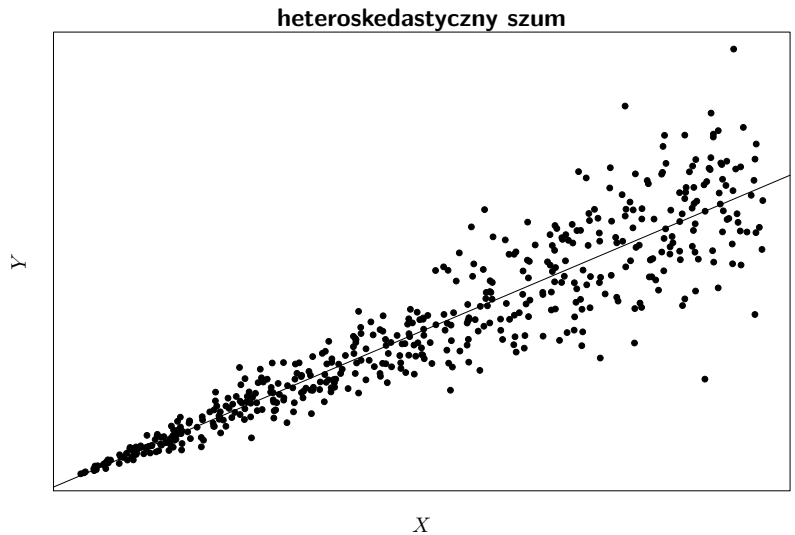
Regresja liniowa: założenia

Zmienna Y jest funkcją liniową zmiennej X wraz z dodatkowym składnikiem losowym („szumem") ϵ:

$$Y = aX + b + \epsilon, \qquad \epsilon \sim N(0, \sigma^2)$$

- Współczynniki a, b możemy traktować jako nieznane parametry populacji
- **Homoskedastyczność** szumu: jego wariancja σ^2 jest stała i nie zależy od X





Rozkład zmienności Y

Na danych $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ wyznaczono współczynniki regresji $\hat{a}, \hat{b}$ metodą najmniejszych kwadratów.
Możemy je traktować jako estymatory nieznanych współczynników a, b populacji

Oznaczmy $\widehat{Y}_i = \hat{a}X_i + \hat{b}$. 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

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

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

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

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

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

$$\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)(\widehat{Y}_i - \overline{Y}) = \sum_{i=1}^{n}\hat{a}(Y_i - \overline{Y})(X_i - \overline{X}) - \hat{a}^2\sum_{i=1}^{n}(X_i - \overline{X})^2$$

$$= (n-1)\hat{a}(s_{XY} - \hat{a}s_X^2) = (n-1)\hat{a}\left(s_{XY} - \frac{s_{XY}}{\cancel{s_X^2}}\cancel{s_X^2}\right) = 0.$$

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

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

a więc SST = SSR.

Dlaczego SSE to część niewyjaśniona przez model liniowy?

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

$$\text{SSR} = \sum_{i=1}^{n} (\widehat{Y}_i - \overline{Y})^2 = \sum_{i=1}^{n} (\hat{a}X_i + \hat{b} - \overline{Y})^2 = n(\hat{b} - \overline{Y})^2.$$

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

Współczynnik determinacji

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

Cześć zmienności Y wyjaśnionej przez model liniowy.

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

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

Test na istotność regresji

- **Układ hipotez**:

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

- **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	**Istotność korelacji**
$H_0: \quad a = 0$	$H_0: \quad \rho = 0$
$H_1: \quad a \neq 0$	$H_1: \quad \rho \neq 0$
$F = \dfrac{\mathsf{SSR}}{\mathsf{SSE}}(n-2)$	$T = \dfrac{r}{\sqrt{1-r^2}}\sqrt{n-2}$

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

Istotność regresji vs. istotność korelacji

Istotność regresji	**Istotność korelacji**
$H_0: \quad a = 0$	$H_0: \quad \rho = 0$
$H_1: \quad a \neq 0$	$H_1: \quad \rho \neq 0$
$F = \dfrac{\mathsf{SSR}}{\mathsf{SSE}}(n-2)$	$T = \dfrac{r}{\sqrt{1-r^2}}\sqrt{n-2}$

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

Jest to w zasadzie ten sam test:

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

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

Pozostałe współczynniki

- Błąd standardowy oszacowania (estymator rozrzutu szumu σ):

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

- Błędy standardowe parametrów $\hat{a}$ i $\hat{b}$:

$$s_{\hat{a}} = \frac{S}{s_X}\sqrt{n-1}$$

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

Globalny test na istotność regresji wielorakiej

Model liniowy z m zmiennymi objaśniającymi:

$$\widehat{Y} = \beta_0 + \sum_{i=1}^{m} \beta_i X_i$$

- **Układ hipotez**:

$H_0:$	$\beta_1 = \beta_2 = \ldots = \beta_m = 0$
$H_1:$	Co najmniej jeden $\beta_i \neq 0$

- **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**:

$H_0:$	$\beta_i = 0$
$H_1:$	$\beta_i \neq 0$

- **Statystyka testowa**:

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

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