Analiza regresji

Na podstawie materiałów WK

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 kwadratów (s.k.o.) 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.

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.

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 ($a = 0$). Wtedy:

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

Ponieważ $b = \overline{Y} - a\overline{X} = \overline{Y}$, 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** (dla modelu z jedną zmienną niezależną).
Używając $a = r\frac{s_Y}{s_X}$ oraz $b = \overline{Y} - a\overline{X}$:

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

Test na istotność regresji

- **Zmiana oznaczeń**: $a \to \beta_1/b_1$, $b \to \beta_0/b_0$
- **Układ hipotez**:

$$\begin{array}{ll} H_0: & \beta_1 = 0 \\ H_1: & \beta_1 \neq 0 \end{array}$$

- **Statystyka testowa**:

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

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

Istotność regresji vs. istotność korelacji

Istotność regresji	**Istotność korelacji**
$H_0: \quad \beta_1 = 0$	$H_0: \quad \rho = 0$
$H_1: \quad \beta_1 \neq 0$	$H_1: \quad \rho \neq 0$
$F = \dfrac{\mathsf{SSR}}{\mathsf{SSE}}(n-2)$	$T = \dfrac{r}{\sqrt{1-r^2}}\sqrt{n-2}$

Ale $b_1 = r\frac{s_Y}{s_X}$, więc $b_1 = 0 \iff r = 0 \ldots ?$

Istotność regresji vs. istotność korelacji

Istotność regresji	**Istotność korelacji**
$H_0: \quad \beta_1 = 0$	$H_0: \quad \rho = 0$
$H_1: \quad \beta_1 \neq 0$	$H_1: \quad \rho \neq 0$
$F = \frac{\mathsf{SSR}}{\mathsf{SSE}}(n-2)$	$T = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$

Ale $b_1 = r\frac{s_Y}{s_X}$, więc $b_1 = 0 \iff r = 0 \ldots ?$

Jest to w zasadzie ten sam test (jeśli chodzi o wynik wnioskowania - test F jest prawostronny!):

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

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

Pozostałe współczynniki

- Błąd standardowy oszacowania:

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

- Błędy standardowe parametrów b_1 i b_0:

$$S_{b_1} = \frac{S}{\sqrt{SS_X}}$$

$$S_{b_0} = S\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{SS_X}}$$

Globalny test na istotność regresji wielorakiej

Model liniowy z m zmiennymi objaśniającymi:

$$\widehat{Y} = b_0 + \sum_{i=1}^{m} b_i X_i$$

- **Układ hipotez**:

$$\begin{array}{ll} H_0: & \beta_1 = \beta_2 = \ldots = \beta_m = 0 \\ H_1: & \text{Co najmniej jeden } \beta_i \neq 0 \end{array}$$

- **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**:

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

- **Statystyka testowa**:

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

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

Skorygowany współczynnik determinacji

Skorygowany (*adjusted*) współczynnik R^2 uwzględnia liczbę predyktorów (wykorzystanych zmiennych niezależnych) i ich wypływ na "jakość" predykcji (w stosunku do poprawy przypadkowej).
Jest on zdefiniowany wzorem:

$$R^2_{adj} = 1 - \left[\frac{\left(1 - R^2\right)(n-1)}{n-k-1} \right],$$

gdzie n to liczba punktów, a k to liczba predyktorów.