

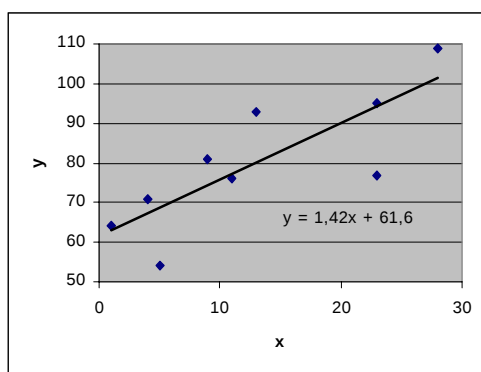
1 Wykład 12

Przykład 1.1 (powtórzenie z wykładu 3) $Y^T = [y_1, y_2, \dots, y_n]$ jest wektorem obserwacji, którego i -ta współrzędna y_i jest ilością fosforu w glebie po 38 dniach wsiania na i -tym poletku nawozu, zawierającego x_i jednostek fosforu. Wyniki zawarte są w tabeli:

x_i	1	4	5	9	11	13	23	23	28
x_i^*	-12	-9	-8	-4	-2	0	10	10	15
y_i	64	71	54	81	76	93	77	95	109

Prosta regresji będzie miała parametry

$$\|x^*\| = 27.09, \hat{\beta} = \frac{\langle x^*, y \rangle}{\|x^*\|^2} = 1.42, \hat{\alpha}^* = \bar{y} = 80, \hat{\alpha} = \alpha^* - \beta^* \bar{x} = 61.6$$



■

Spróbujemy powiedzieć coś o jakości tego oszacowania.

Twierdzenie 1.1 Niech $R = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ gdzie $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$, $\bar{x} = 0$, $s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$, $s_x^2 = \frac{1}{n-1} \sum_{i=1}^n x_i^2$. Wtedy:

1. $\hat{\alpha}$ ma rozkład $\mathcal{N}(\alpha, \sigma/\sqrt{n})$,
2. $\hat{\beta}$ ma rozkład $\mathcal{N}(\beta, \sigma/\|x\|)$,
3. $R = (n-1)(s_y^2 - s_x^2 \hat{\beta}^2)$ ma rozkład $\sigma^2 \chi_{n-2}^2$
4. $\hat{\alpha}, \hat{\beta}, R$ są niezależne
5. $\hat{\sigma}^2 \stackrel{\text{def}}{=} R/(n-2)$ jest nieobciążonym estymatorem σ^2 .

Estymatory α i β są nieobciążone, estymator α w miarę wzrostu liczby obserwacji jego wariancja zbiega do zera. Estymator parametru β ma wariancję, która jest tym mniejsza im większy jest rozrzut wartości x . Przy planowaniu eksperymentu, w którym chcemy oszacować równanie regresji, należy wybrać zestaw argumentów x_i tak aby estymatory były jak najbliższe rzeczywistej wartości. Należy więc wybrać ich dużo, co zwiększa dokładność oszacowania α i muszą one być jak najbardziej rozrzucone, co zwiększa dokładność oszacowania β .

Przykład 1.2 (c.d.) W naszym przykładzie:

$$\begin{aligned} s_y^2 &= 284.25, s_y = 16.86, s_x^2 = 91.75, s_x = 9.58, \\ R &= 8(284.25 - 91.75 \cdot 1.42^2) = 793.96, \\ \widehat{\sigma^2} &= \frac{793.96}{7} = 113.42, \widehat{\sigma} = 10.65 \end{aligned}$$

Parametr $\widehat{\sigma}$ opisuje wpływ nieobserwowanych czynników, modyfikujących postulowaną zależność liniową. Gdy liczba ta jest duża, należy się zastanowić, czy postulowany model jest adekwatny. ■

Wniosek 1.1 Jeżeli prawdziwą wartością β jest β_0 to zmienna losowa

$$\frac{\widehat{\beta} - \beta_0}{\widehat{\sigma}/\|x\|}$$

ma rozkład Studenta t_{n-2} z $n-2$ stopniami swobody.

Dowód. Wynika z faktu, że $\widehat{\beta}, R$ są niezależne oraz z rozkładu normalnego $\widehat{\beta}$ i rozkładu $\sigma^2 \chi_{n-2}^2$ dla R . ■

Wniosek 1.2 Zbiór krytyczny dla testowania

$$\begin{aligned} H_0 &: \beta = \beta_0, \\ H_1 &: \beta \neq \beta_0 \end{aligned}$$

ma postać

$$C = \left\{ y : \frac{|\widehat{\beta} - \beta_0|}{\widehat{\sigma}/\|x\|} > k \right\}$$

Przedział ufności dla β na poziomie ufności $1 - \gamma$ ma postać

$$\left(\widehat{\beta} - d(y, \gamma), \widehat{\beta} + d(y, \gamma) \right)$$

i dokładność $d(y, \alpha)$ wyraża się wzorem:

$$d(y, \gamma) = t_{n-2, \gamma} \frac{\widehat{\sigma}}{\|x\|}$$

gdzie $P(|t_{n-2}| > t_{n-2, \gamma}) = \gamma$

Dowód. Postać zbioru krytycznego wynika z tezy poprzedniego wniosku. Konstrukcja przedziału ufności wynika z faktu, że dopełnienie zbioru krytycznego¹ na poziomie istotności γ jest przedziałem ufności na poziomie ufności $1 - \gamma$.² ■

¹zwane zbiorem akceptacji

²wynika to z dualności testowania hipotez i przedziałów ufności (twierdzenie 7.1)

Przykład 1.3 (c.d.) Aby obliczyć przedział ufności dla β na poziomie ufności 0.95 należy najpierw obliczyć jego długość. Z tablic rozkładu Studenta z 7 stopniami swobody obliczamy wartość krytyczną $t_{7,0.05} = 2.365$. Dokładność przedziału wynosi

$$d(y, \gamma) = t_{n-2, \gamma} \frac{\hat{\sigma}}{\|x^*\|} = 2.365 \frac{10.65}{27.09} = 0.93$$

Przedział ufności dla β na poziomie ufności 0.95 ma więc postać

$$(1.42 - 0.93, 1.42 + 0.93) = (0.49, 2.35)$$

Z tego wynika, że żadna z wartości między 0.49 a 2.35 nie będzie odrzucana przez test na poziomie 0.05. Co więcej każda wartość spoza tego przedziału będzie odrzucana na poziomie 0.05. Oznacza to, między innymi, że odrzucona będzie hipoteza $\beta = 0$ czyli hipoteza o nieistotności modelu regresji ■

Praktycznym zastosowaniem regresji jest możliwość prognozy wartości y w wybranym przez nas punkcie x_0 . Zazwyczaj prognozowaną wartość odczytuje się z prostej regresji $\hat{y}(x_0) = \hat{\alpha} + \hat{\beta}x_0$. Jednak ta wartość jest obarczona nie tylko błędem modelu ε ale również błędami estymacji parametrów regresji. Dlatego też warto znaleźć przedział ufności dla prognozy w zadanym punkcie.

Propozycja 1.1 Przedział ufności na poziomie ufności $1 - \gamma$ dla prognozy w punkcie x_0 ma postać

$$(\hat{y}(x_0) - d(y, \gamma, x_0), \hat{y}(x_0) + d(y, \gamma, x_0))$$

gdzie

$$d(y, \gamma, x_0) = t_{n-2, \gamma} \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{x_0^2}{\|x\|^2}},$$

$$P(|t_{n-2}| > t_{n-2, \gamma}) = \gamma$$

Dowód.

$$y(x_0) - \hat{y}(x_0) = \alpha + \beta x_0 + \varepsilon - \hat{\alpha} - \hat{\beta} x_0 = (\alpha - \hat{\alpha}) + (\beta - \hat{\beta}) x_0 + \varepsilon$$

Wszystkie trzy składniki sumy są niezależne, mają wartość oczekiwaną 0 i wariancję, równą sumie ich wariancji:

$$\frac{\sigma^2}{n} + \frac{\sigma^2 x_0^2}{\|x\|^2} + \sigma^2 = \sigma^2 \left(1 + \frac{1}{n} + \frac{x_0^2}{\|x\|^2} \right)$$

Stąd wynika, że

$$\frac{y(x_0) - \hat{y}(x_0)}{\hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{x_0^2}{\|x\|^2}}}$$

ma rozkład Studenta t_{n-2} z $n - 2$ stopniami swobody. Stąd

$$P \left(\left| \frac{y(x_0) - \hat{y}(x_0)}{\hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{x_0^2}{\|x\|^2}}} \right| > t_{n-2, \gamma} \right) = \gamma$$

Po rozwiązaniu tej nierówności względem $y(x_0)$ otrzymamy szukany przedział ufności. ■

Prognostyczny przedział ufności nie ma stałej szerokości. Jest on największy w punkcie $x_0 = 0$ a więc w średniej arytmetycznej wartości x_i . Najmniej mylimy się w pobliżu wartości typowych. Im bardziej oddalamy się od tych wartości tym większy popełniamy błąd prognostyczny.

Przykład 1.4 (c.d.) Chcemy podać przedział ufności dla prognozy w punkcie $x_0 = 10$. Ponieważ średnia wartość argumentów, użytych do konstrukcji przedziału ufności nie jest równa 0 musimy skorygować nasze argumenty: $x_0^* = 10 - 13 = -3$.

$$\begin{aligned}\hat{y}(-3) &= 80 + 1.42(-3) = 75.74, \\ \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{x_0^2}{\|x\|^2}} &= 10.65 \sqrt{1 + \frac{1}{9} + \frac{(-3)^2}{793.96}} = 11.283, \\ d(y, 0.05, -3) &= 2.365 \cdot 11.283 = 26.69\end{aligned}$$

Przedział ufności dla wartości w punkcie $x_0 = 10$ będzie miał postać

$$(75.74 - 26.69, 75.74 + 26.69) = (49.05, 102.43)$$

Patrząc na wykres danych zauważmy, jak szeroki jest ten przedział ufności. ■

Ważną rolę w modelu regresji liniowej pełni współczynnik korelacji

1.1 Współczynnik korelacji

Dane są dwa wektory $a^T = (a_1, a_2, \dots, a_n)$ i $b^T = (b_1, b_2, \dots, b_n)$ takie, że $\bar{a} = \bar{b} = 0$.

Definicja 1.1 Współczynnikiem korelacji wektorów a i b nazywamy wyrażenie

$$r(a, b) \stackrel{\text{def}}{=} \cos(\angle(a, b)) = \frac{\langle a, b \rangle}{\|a\| \|b\|}$$

Dla dowolnych wektorów definiujemy współczynnik korelacji

$$r(a, b) \stackrel{\text{def}}{=} r(a^*, b^*)$$

gdzie

$$\begin{aligned}a^{*T} &= (a_1 - \bar{a}, a_2 - \bar{a}, \dots, a_n - \bar{a}), \\ b^{*T} &= (b_1 - \bar{b}, b_2 - \bar{b}, \dots, b_n - \bar{b})\end{aligned}$$

Wniosek 1.3 Współczynnik korelacji ma następujące własności:

1. $|r(a, b)| \leq 1$,
2. $|r(a, b)| = 1 \iff \exists (\alpha, \beta) \ b_i = \alpha + \beta a_i$

Dowód. Pierwsza własność wynika z tego, że współczynnik korelacji jest cosinusem kąta. Druga z faktu, że jeśli cosinus kąta jest co do modułu równy 1 to wektory są do siebie równoległe, czyli dla pewnego β , $b_i - \bar{b} = \beta(a_i - \bar{a})$ ■

Wniosek 1.4 Dla regresji liniowej niech $r \stackrel{\text{def}}{=} r(y, x)$. Wtedy:

$$\hat{\beta} = r \frac{s_y}{s_x}, \quad (1)$$

$$R = (n-1) s_y^2 (1-r^2), \quad (2)$$

$$\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - r^2, \quad (3)$$

$$\hat{\sigma}^2 = s_y^2 (1-r^2) \frac{n-1}{n-2} \quad (4)$$

Dowód. Wszystkie własności wymienione we wniosku wynikają z podstawienia wzoru (1). Aby ten wzór udowodnić wystarczy zauważyć, że

$$\begin{aligned} \hat{\beta} &= \frac{\langle y, x \rangle}{\|x\|^2} = \frac{\langle y^*, x \rangle}{\|x\|^2} + \frac{\langle y - y^*, x \rangle}{\|x\|^2} = \frac{\langle y^*, x \rangle}{\|x\| \|y\| \|x\|} + \frac{\sum_i \bar{y} x_i}{\|x\|^2} = \\ &= r \frac{\|y\| / \sqrt{n-1}}{\|x\| / \sqrt{n-1}} = r \frac{s_y}{s_x} \end{aligned}$$

■

Własność (3) opisuje, na ile regresja wyjaśnia zmienność obserwacji y . W liczniku mamy kwadraty odchyłeń obserwacji i prognoz na podstawie regresji (błąd regresji), w mianowniku - całkowitą zmienność. Stosunek tych wartości wyraża, jaka część obserwacji *nie* jest wyjaśniona przez regresję. Zależy ona od kwadratu współczynnika korelacji. W praktyce uważa się, że korelacja jest "mocna", jeśli regresja wyjaśnia co najmniej połowę zmienności. To znaczy, korelacja jest "mocna", gdy $r^2 > 0.5$, czyli gdy $r > 0.7$.

Wniosek 1.5 Dla testowania hipotezy istotności regresji

$$\begin{aligned} H_0 &: \beta = 0, \\ H_1 &: \beta \neq 0 \end{aligned}$$

zbiór krytyczny ma postać

$$C = \{y : |T(y)| > k\}$$

gdzie

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

Gdy H_0 jest prawdziwa to $T(y)$ ma rozkład Studenta t_{n-2} z $n-2$ stopniami swobody.

Dowód. Korzystając z wniosku 1.2 zbiór krytyczny będzie miał postać:

$$\begin{aligned} C &= \left\{ y : \frac{|\hat{\beta}|}{\hat{\sigma}/\|x\|} > k \right\} = \left\{ y : \frac{\left| r \frac{s_y}{s_x} \right| \|x\|}{s_y \sqrt{1-r^2} \sqrt{\frac{n-1}{n-2}}} > k \right\} = \\ &= \left\{ y : \frac{|r| \frac{s_y}{s_x} (n-1) s_x}{s_y \sqrt{1-r^2} \sqrt{\frac{n-1}{n-2}}} > k \right\} = \left\{ y : \left| \frac{r}{\sqrt{1-r^2}} \sqrt{n-2} \right| > k \right\} \end{aligned}$$

co kończy dowód. ■

Przykład 1.5 (c.d.) Wykonamy obliczenia, związane z własnościami współczynnika korelacji i testowaniem istotności regresji.

$$r = 1.42 \frac{9.58}{16.86} = 0.81$$

Część zmienności nie wyjaśniona przez regresję wynosi $1 - r^2 = 1 - 0.81^2 = 0.34$ (34%).

Poziom krytyczny testu na istotność regresji wyraża prawdopodobieństwo

$$P\left(|T| > \frac{0.81}{\sqrt{1 - 0.81^2}} \sqrt{7}\right) = P(|T| > 3.654) < P(|T| > 3.499) = 0.01$$

Możemy więc, z ryzykiem popełnienia błędu I rodzaju mniejszym od 0.01, odrzucić hipotezę o nieistotności regresji. ■