

1 O regresji liniowej i nie-tylko?

Powiedzmy, że mamy przeanalizować przyszły czas życia Kowalskiego (oznaczony przez T) studenta 3 roku lat 20 w dniu 20 stycznia 2010. Przypuśćmy, że wiemy, że dalszy czas życia Kowalskiego ma rozkład normalny $T \in N(\mu, \sigma)$. Zauważmy, że T możemy zapisać w postaci

$$T = \mu + \epsilon,$$

gdzie $\epsilon \in N(0, \sigma)$. Niestety nie wiemy jakie są parametry tego rozkładu. Co gorsza śmierć Kowalskiego nie ułatwi nam zadania. Co robimy? Zakładamy, że poznamy wyniki dalszego czasu życia t_1, \dots, t_n dla n studentów którzy dożyli do 3 roku (mieli 20 lat). Ponieważ jeszcze nie wybraliśmy w sposób niezależny tych studentów – co gorsza jeszcze niektórzy mogą żyć — zatem zakładamy, że dany jest ciąg niezależnych zmiennych losowych T_1, \dots, T_n o rozkładzie T , czyli $N(\mu, \sigma)$. Zauważmy, że model ten możemy zapisać w postaci

$$T_i = \mu + \epsilon_i,$$

gdzie ϵ_i $i = 1, \dots, n$ tworzą ciąg niezależnych zmiennych losowych o rozkładzie $N(0, \sigma)$. Jak estymować teraz parametry np. μ ? Otóż estymator to

$$\hat{\mu} = \frac{1}{n} \sum_{j=1}^n T_j.$$

Czasem zapisujemy to w postaci

$$\hat{\mu} = \frac{1}{n} \sum_{j=1}^n t_j,$$

gdzie rozumiemy, że realizacje t_1, \dots, t_n dopiero poznamy. Oczywiście możemy mieć wyłącznie dane historyczne.

Okazało się, powyższy model jest mało dokładny gdyż jeśli znamy x wzrost studenta Kowalskiego, to oczekiwany czas życia Kowalskiego

$$\mu(x) = \beta x + \alpha.$$

gdzie β i α to pewne parametry. Zauważmy, że teraz model jest postaci

$$T = \mu(x) + \epsilon,$$

$$T = \beta x + \alpha + \epsilon,$$

gdzie $\epsilon \in N(0, \sigma)$. Problem jest ten sam. Tym razem zakładamy, że poznamy wyniki dalszego czasu życia t_1, \dots, t_n dla n studentów którzy dożyli do 3 roku (mieli 20 lat) oraz ich wzrost x_1, \dots, x_n . Zakładam, że wyniki t_1, \dots, t_n są realizacjami niezależnych zmiennych losowych T_1, \dots, T_n o rozkładzie odpowiednio $N(\mu(x_i), \sigma)$. Zauważmy, że model ten możemy zapisać w postaci

$$T_i = \mu(x_i) + \epsilon_i,$$

$$T_i = \beta x_i + \alpha + \epsilon_i,$$

gdzie ϵ_i $i = 1, \dots, n$ tworzą ciąg niezależnych zmiennych losowych o rozkładzie $N(0, \sigma)$. Powiększyła nam się liczba parametrów. Teraz estymować trzeba β, α, σ . Zwykle wzory pisze się dla realizacji (t_i, x_i) , $i = 1, 2, \dots, n$.

Ponadto są dwie możliwości badań. Badania prospektywne – wybieramy studentów o ustalonym wzroście i czekamy. Badania retrospektywne mamy dane historyczne i osoby już nie żyją tutaj wzrost jest dany z wyborem grupy historycznej.

2 Korelacja

Miarą relacji liniowej wektora losowego (X, Y) o skończonych i niezerowych wariancjach jest współczynnik korelacji

$$\rho = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}},$$

gdzie

$$Cov(X, Y) = E(X - EX)(Y - EY),$$

$$0 < \sigma_X = \sqrt{Var(X)} < \infty, \quad 0 < \sigma_Y = \sqrt{Var(Y)} < \infty.$$

Przy powyższych założeniach $Cov(X, Y)$ jest dobrze określona gdyż z nierówności Schwartza

$$|Cov(X, Y)| \leq \sigma_X \sigma_Y.$$

Wprowadzamy oznaczenia

$$\mu_X = EX, \quad \mu_Y = EY.$$

Estymatorem parametru ρ z próby losowej (X_j, Y_j) , $j = 1, 2, \dots, n$ jest

$$r_n = \frac{\frac{1}{n} \sum_{j=1}^n (X_j - \bar{X})(Y_j - \bar{Y})}{S_x S_y},$$

gdzie

$$S_x = \sqrt{\frac{1}{n} \sum_{j=1}^n (X_j - \bar{X})^2}, \quad S_y = \sqrt{\frac{1}{n} \sum_{j=1}^n (Y_j - \bar{Y})^2},$$

$$\bar{X} = \frac{1}{n} \sum_{j=1}^n X_j, \quad \bar{Y} = \frac{1}{n} \sum_{j=1}^n Y_j$$

Definicja Wektor losowy (X, Y) ma niezdegenerowany dwuwymiarowy rozkład normalny z parametrami $(\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho)$ jeśli jego gęstość ma postać

$$f(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \times \\ \text{Exp} \left(-\frac{1}{2(1-\rho^2)} \left(\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - 2\rho \frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} \right) \right).$$

Test o nieskorelowaniu Niech (X, Y) ma niezdegenerowany dwuwymiarowy rozkład normalny. Wysuwamy hipotezę zerową $H : \rho = 0$ wobec hipotezy alternatywnej $H_a : \rho \neq 0$. Przy założeniu hipotezy zerowej, statystyka testowa

$$t = \frac{r_n}{\sqrt{1-r_n^2}} \sqrt{n-2}$$

ma rozkład t-Studenta z $n-2$ stopniami swobody.

UWAGA We wzorze na r_n możemy pisać (x_j, y_j) , $j = 1, 2, \dots, n$ jako możliwa realizacja z próby losowej (X_j, Y_j) , $j = 1, 2, \dots, n$.