Metody nieparametryczne w statystyce

Tomasz Wójtowicz

Wydział Zarządzania
AGH Akademia Górniczo-Hutnicza im. S. Staszica w Krakowie

Testowanie

Niech $X_1, \dots, X_n$ będzie próbą losową prostą z pewnego rozkładu F.
Interesuje nas wartość parametru θ rozkładu F.
Niech $\hat{\theta} = T(X_1, .., X_n)$ będzie pewną statystyką.

Próbę losową prostą $\boldsymbol{X}^* = (X_1^*, \dots, X_n^*)$ z rozkładu $\hat{F}$ dla ustalonej realizacji $x_1, \dots, x_n$ nazywamy **próbą typu bootstrap** (**próbą bootstrap**).

W celu otrzymania realizacji próby bootstrap dokonuje się n-krotnego losowania **ze zwracaniem** spośród elementów oryginalnej próby $x_1, \dots, x_n$.

Rozkład statystyki $T(\boldsymbol{X}^*) - \hat{\theta}$ dla próby bootstrap przy ustalonych wartościach realizacji $x_1, \dots, x_n$ jest dla regularnych statystyk T bliski rozkładowi $T(\boldsymbol{X}) - \theta$.

Testowanie

Kolejne kroki:

- na podstawie realizacji $x_1, \dots, x_n$ obliczyć wartość $\hat{\theta}$,
- na podstawie realizacji $x_1, \dots, x_n$ wylosować niezależne próby bootstrap $\boldsymbol{X_1^*, \dots, X_k^*}$,
- obliczyć wartości $\hat{\theta}_1^* = T(\boldsymbol{X_1^*}) - \hat{\theta}$, $\hat{\theta}_2^* = T(\boldsymbol{X_2^*}) - \hat{\theta}$,..., $\hat{\theta}_k^* = T(\boldsymbol{X_k^*}) - \hat{\theta}$,
- skonstruować histogram wartości $\hat{\theta}_1^*, \dots, \hat{\theta}_k^*$, który jest przybliżeniem rozkładu statystyki $\hat{\theta} - \theta$.

Uzyskany rozkład jest przybliżeniem rozkładu błędów estymacji parametru θ.

Nazywamy go estymatorem rozkładu $\hat{\theta}$ uzyskanym metodą bootstrap.

Testowanie

Mając rozkład $\hat{\theta}_1^*, \ldots, \hat{\theta}_k^*$ możemy oszacować zmienność estymatora $\hat{\theta} = T(X)$, co nie jest możliwe przy wykorzystaniu tylko wartości $x_1, \ldots, x_n$.

Definicja

Błędem standardowym typu bootstrap estymatora $\hat{\theta}$ nazywamy:

$$S_{\hat{\theta}} = \sqrt{\frac{1}{k-1}\sum_{i=1}^{k}\left(\hat{\theta}_i^* - \bar{\theta}^*\right)^2},$$

$$\bar{\theta}^* = \frac{1}{k}\sum_{i=1}^{k}\hat{\theta}_i^*$$

Testowanie

Metodę bootstrap można wykorzystać do testowania hipotez statystycznych.

W tym przypadku do oceny prawdziwości hipotezy głównej stosujemy rozkład typu bootstrap statystyki testowej przy założonej prawdziwości hipotezy głównej.

Przykład

W tabeli zamieszczono wyniki pewnego egzaminu:

studenci	20.5	18.5	21	30	26	29	26	30.5
studentki	24	29.5	29	8	23.5			

Czy można przyjąć, że wartość oczekiwana punktów uzyskanych z egzaminu przez studentów jest mnijesza niż 30.

Testowanie

Hipotezy mają postać:

$$H_0: m = 30$$

$$H_1: m < 30$$

Gdybyśmy założyli, że badana cecha ma rozkład normalny, to moglibyśmy skorzystać ze statystyki:

$$t = \frac{\bar{X} - 30}{S_{\bar{X}}} = \frac{\bar{X} - 30}{S_X}\sqrt{n}$$

W tym przypadku mamy:

$\bar{x} \approx 25{,}19$, $S_{\bar{X}} \approx 13{,}16$ czyli $t_0 \approx -0{,}366$.

Testowanie

Aby określić czy uzyskana statystyka wskazuje na istotność wartości oczekiwanej wyznaczymy rozkład typu bootstrap statystyki t.

Jednak rozkład ten należy wyznaczyć na podstawie próby bootstrap spełniającej hipotezę H_0.

Dane wyjściowe $x_1, \dots, x_8$ nie muszą jej spełniać. Ale jeżeli je przekształcimy do postaci:

$$y_i = x_i - \bar{x} - m_0 = x_i - \bar{x} - 30,$$

to y_i już spełniają hipotezę H_0.

Na podstawie prób bootstrap $\boldsymbol{Y_1^*}, \dots, \boldsymbol{Y_k^*}$ obliczamy wartości statystyk $t_1^*, \dots, t_k^*$.

Mając te statystyki możemy już wyznaczyć wartość p-value.

Przykład - korelacja

Na podstawie wyników kilkunastu studentów zbadać istotność zależności pomiędzy wynikami z zaliczenia i z egzaminu:

zaliczenie	55	57.5	64.4	66.9	69.3	54.7	60.6	82.5	58.6	64.7	50.7	50.7	51.1	55.3
egzamin	38.9	55.6	53.7	48.1	56.5	44.4	54.6	53.7	43.5	56.5	50	1.9	31.5	27.8

Współczynnik korelacji Pearsona: $r_0 = 0{,}56$.

W celu weryfikacji istotności współczynnika korelacji stosujemy metodę bootstrap, zakładając, że hipoteza główna jest prawdziwa, tzn.:

$$H_0: \rho = 0$$

$$H_1: \rho \neq 0$$

Testowanie

Procedura jest następująca:

1. losujemy **niezależne** próby bootstrapowe z wiersza „zaliczenie” i z wiersza „egzamin”,
2. dla każdej wygenerowanej i-tej próby $(i = 1, \dots, k)$obliczamy współczynnik korelacji r_i^*
3. obliczamy:

$$pv = \frac{\#\{r_i^* : |r_i^*| < r_0\}}{k}$$

Alternatywnie można wyznaczyć dwustronny przedział ufności dla współczynnika korelacji i sprawdzić czy 0 znajduje się w tym przedziale.

Regresja

Rozważmy model liniowy:

$$y = X\beta + \varepsilon$$

Wektor parametrów β można estymować za pomocą MNK. Estymator ma postać:

$$\hat{\beta} = (X'X)^{-1}X'y$$

Jeżeli składniki losowe są i.i.d to jego macierz kowariancji ma postać:

$$var(\hat{\beta}) = \sigma^2(X'X)^{-1}$$

gdzie $\sigma^2 = var(\varepsilon_i)$.

W przypadku, gdy składniki losowe ε_i mają rozkład normalny możemy stosować odpowiednie testy do badania istotności parametrów modelu.

Regresja

Gdy składniki losowe ε_i nie maja rozkładu normalnego lub danych jest mało możemy stosować metody bootstrapowe do oceny istotności parametrów modelu.

Możliwe są dwa sposoby generowania próby bootstrapowej:

- na podstawie błędów losowych
- na podstawie obserwacji.

Można również zastosować metodę Jacknife.

Próba bootstrapowa na podstawie błędów losowych

Po wyestymowaniu parametrów $\hat{\beta}$ modelu obliczamy reszty modelu:

$$e = y - X\hat{\beta}$$

gdzie $e = [e_1, \dots, e_n]$.

Procedura (dla $i = 1, \dots R$):

1. z próby $e_1, \dots, e_n$ losujemy próbę bootstrapową
$$e^* = [e_1^*, \dots, e_n^*]'$$
2. na tej podstawie obliczamy $y^* = X\hat{\beta} + e^*$
3. na podstawie y^* i X obliczamy estymator $\hat{\beta}_i^*$

W ten sposób otrzymujemy próbę bootstrapową $\hat{\beta}_1^*, \dots, \hat{\beta}_R^*$

Możemy teraz wyznaczyć rozkład bottstrapowy parametru β.

Próba bootstrapowa na podstawie obserwacji

Niech:

$$z_i = [y_i, x_{i1}, \dots, x_{ik}]$$

Procedura (dla $i = 1, \dots R$):

1. z próby $z_1, \dots, z_n$ losujemy próbę bootstrapową $z_1^*, \dots, z_n^*$
2. na jej podstawie otrzymujemy wektor y^* i macierz X^*
3. do y^* i X^* stosujemy MNK i obliczamy estymator $\hat{\beta}_i^*$

W ten sposób otrzymujemy próbę bootstrapową $\hat{\beta}_1^*, \dots, \hat{\beta}_R^*$

Możemy teraz wyznaczyć rozkład bottstrapowy parametru β.