Wnioskowanie w rachunku korelacji i regresji z wykorzystaniem SAS

mgr Maciej Beręsewicz

1 Współczynnik korelacji liniowej Pearsona

Współczynnik korelacji liniowej Pearsona określa poniższy wzór:

$$r = \frac{cov(x,y)}{\sqrt{s_x^2 s_y^2}} = \frac{\sum_{i=1}^{n}((x_i - \bar{x})(y_i - \bar{y}))}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{1}$$

gdzie $cov(x,y)$ oznacza kowariancję między zmienną x oraz y, s_x i s_y oznacza odchylenie standardowe zmiennej x i y. $\bar{x}$ oraz $\bar{y}$ oznaczają wartości średnie, a n oznacza liczbę obserwacji.

2 Przedział ufności dla współczynnika korelacji

Przedział ufności dla współczynnika korelacji należy rozpocząć od transformacji Fishera, którą opisuje poniższy wzór:

$$z_r = tanh^{-1}(r) = \frac{1}{2} log(\frac{1+r}{1-r}) \tag{2}$$

gdzie r określone jest zgodnie ze wzorem (1)[1]. Statystyka z asymptotyczny rozkład normalny normalny o średniej:

$$E(z_r) = \xi + \frac{\rho}{2(n-1)} \tag{3}$$

oraz wariancji

$$V(z_r) = \frac{1}{n-3} \tag{4}$$

gdzie $\xi = tanh^{-1}(\rho)$, zatem przedział ufności dla współczynnika transformowanego współczynnika korelacji z_r określony jest poniższym wzorem:

[1] Przekształcenie możemy zastosować również do współczynnika korelacji rang Spearmana.

$$P\{z_r - z_{1-\alpha/2}\sqrt{\frac{1}{n-3}} < z_r < z_r + z_{1-\alpha/2}\sqrt{\frac{1}{n-3}}\} = 1 - \alpha \tag{5}$$

gdzie $z_{1-\alpha/2}$ jest kwantylem rozkładu normalnego standaryzowanego. W literaturze pojawia się również poprawna na obciążenie przedziału ufności, która wynika z przybliżania statystyki (1) rozkładem normalnym, określona jest poniższym wzorem:

$$B(r_r) = \frac{r}{2(n-1)}. \tag{6}$$

W związku z czym przedział ufności dla transformowanego współczynnika korelacji z_r określony wzorem (5) ma następującą postać:

$$P\{z_r - B(r_r) - z_{1-\alpha/2}\sqrt{\frac{1}{n-3}} < z_r < z_r - B(r_r) + z_{1-\alpha/2}\sqrt{\frac{1}{n-3}}\} = 1 - \alpha \tag{7}$$

gdzie $B(r_r)$ określone jest wzorem (6). Po obliczeniu granic przedziałów ufności dla transformowanego współczynnika korelacji r należy transformować otrzymane granica według następujących wzorów:

$$r_l = tanh(\xi_l) = \frac{exp(2\xi_l) - 1}{exp(2\xi_l) + 1} \tag{8}$$

gdzie ξ_l oznacza dolną granicę przedziału ufności dla transformowanego współczynnika korelacji z_r określonego wzorem (5) lub (7).

$$r_u = tanh(\xi_u) = \frac{exp(2\xi_u) - 1}{exp(2\xi_u) + 1} \tag{9}$$

gdzie ξ_u oznacza górną granicę przedziału ufności dla transformowanego współczynnika korelacji z_r określonego wzorem (5) lub (7). W związku z powyższym przedział ufności dla współczynnika korelacji ma następującą postać:

$$P\{r_l < \rho < r_u\} = 1 - \alpha, \tag{10}$$

który interpretujemy podobnie jak w przypadku przedziału ufności dla średniej czy frakcji.

3 Testowanie istotności współczynnika korelacji

Do testowania poniższego układu hipotez

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$ albo $H_1 : \rho < 0$ albo $H_1 : \rho > 0$

korzystamy ze statystyki testowej określonej wzorem:

$$t = \sqrt{\frac{r^2}{1 - r^2}}\sqrt{n-2}, \tag{11}$$

która ma rozkład t–Studenta z $n - 2$ stopniami swobody. W związku z tym wartość krytyczna dla dwustronnego układu hipotez ma postać $t_{1-\alpha/2}$.

W przypadku gdy weryfikujemy hipotezę o równości współczynnika korelacji z ρ_0 należy zastosować wzór wykorzystujący transformację Fishera. Testowanie następujących hipotez

$H_0 : \rho = \rho_0$

$H_1 : \rho \neq \rho_0$ albo $H_1 : \rho < \rho_0$ albo $H_1 : \rho > \rho_0$

odbywa się z wykorzystaniem poniższej statystyki:

$$z_r - \xi_0 - \frac{\rho_0}{2(n-1)}, \tag{12}$$

gdzie z_r określone jest wzorem (2), a $\xi_0 = tanh^{-1}(\rho_0)$. Statystyka ta ma rozkład normalny o średniej równej zero oraz wariancji równej $\frac{1}{n-3}$. Należy zwrócić uwagę, że $\frac{\rho_0}{2(n-1)}$ oznacza poprawkę na przybliżenie rozkładem normalnym.

4 Procedura PROC CORR w SAS

W SAS do budowania przedziału ufności dla współczynnika korelacji oraz testowania istotności wykorzystywana jest procedura **PROC CORR**. Składnia procedury znajduje się poniżej:

```
PROC CORR data=<zbiór danych>
<pearson> <fisher <(rho0= alpha= type= biasadj=)> >;
VAR <zmienne>;
RUN;
```

gdzie opcje (podane w nawiasach <>) oznaczają:

- pearson – obliczanie współczynnika pearsona (domyślnie, nie jest wymagane podanie). Procedura domyślnie testuje hipotezę $\rho = 0$ według wzoru (11),
- fisher – obliczenie przedziałów ufności dla współczynnika korelacji zgodnie z powyższymi wzorami. Opcja ta ma 4 parametry:
  - rho0 – wskazujemy ρ_0 do testowania hiptez $\rho = \rho_0$ korzystając ze statystyki (12). Domyślnie testuje $\rho_0 = 0$,
  - alpha – określenie poziomu istotności (domyślnie 0.05),
  - type – wskazujemy układ testowanych hipotez. *UPPER* – $H_1 : \rho > \rho_0$, *LOWER* – $H_1 : \rho < \rho_0$, *TWOSIDED* (opcja domyślna) $H_1 : \rho \neq \rho_0$,
  - biasadj – określamy czy chcemy wykorzystać poprawkę określoną we wzorze (12). Do wyboru są opcje *YES* (domyślna) oraz *NO*.