

ĆWICZENIE 6 REGRESJA LINIOWA

Cel

Zapoznanie z modelem regresji liniowej. Wykonanie przykładowych obliczeń dotyczących modelu regresji liniowej dla danych ciągłych za szczególnym uwzględnieniem poprawnej interpretacji wyników.

Wprowadzenie teoretyczne

W pewnym uproszczeniu **modelowanie statystyczne** może być rozumiane jako ciąg kolejno następujących po sobie procedur, których wykonanie prowadzi do wyniku, jakim jest **model statystyczny**. W praktyce modelowania zdarza się często, że wiele z tych procedur należy powtórzyć wielokrotnie. Jeżeli bowiem skonstruowany model nie przejdzie pomyślnie weryfikacji statystycznej, to może się okazać, że badane zjawisko lepiej opisuje inna funkcja lub inny układ zmiennych. Wymusza to ponowną konstrukcję modelu i jego weryfikację.

Algorytm budowy modelu statystycznego jest następujący:

- ✓ dobór zmiennych do modelu regresji,
- ✓ wybór analitycznej postaci modelu (akcent jest tu położony głównie na modele liniowe i modele transformowalne do liniowych),
- ✓ estymacja parametrów modelu,
- ✓ weryfikacja modelu.

Model regresji liniowej można zapisać w następujący sposób:

$$y = \alpha_0 + \alpha_1 x_1 + \dots + \alpha_k x_k + \varepsilon,$$

gdzie y jest **zmienną objaśnianą (zależną)**, x_1, x_2, \dots, x_k są **zmiennymi objaśniającymi (niezależnymi)**, $\alpha_1, \alpha_2, \dots, \alpha_k$ są **parametrami** modelu, ε jest **składnikiem losowym** modelu. Parametry modelu podlegają szacowaniu (estymacji) klasyczną metodą najmniejszych kwadratów.

Zastosowanie tej metody wymaga przyjęcia następujących założeń:

- postać modelu jest liniowa lub sprowadzalna do liniowej,
- zmienne objaśniające są wielkościami nielosowymi,
- zmienne objaśniające są niezależne i wolne od współliniowości, czyli nie występuje między zmiennymi dokładna zależność liniowa,
- liczba obserwacji jest co najmniej równa liczbie szacowanych parametrów,
- składniki losowe dla wszystkich obserwacji mają wartości oczekiwane równe zero ($E(\varepsilon) = 0$)
- składniki losowe mają skończoną wariancję równą σ^2 ,
- kowariancje pomiędzy składnikami losowymi są równe zero, tzn. nie występuje autokorelacja składnika losowego,
- składniki losowe nie są skorelowane ze zmiennymi objaśniającymi,
- składnik losowy ma rozkład normalny $N(0, \sigma)$.

W metodzie najmniejszych kwadratów współczynniki $\hat{\alpha}_i$ dobiera się tak, aby suma kwadratów odchyleń estymowanych wartości zmiennej objaśnianej \hat{y} od jej rzeczywistych wartości y była minimalna:

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \rightarrow \min.$$

Powyższa funkcja przyjmuje minimum w punkcie

$$\hat{\alpha} = (X^T X)^{-1} X^T y,$$

gdzie:

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix} \text{ jest macierzą obserwacji zmiennych objaśniających,}$$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} \text{ jest wektorem obserwacji zmiennej objaśnianej,}$$

$$\hat{\alpha} = \begin{bmatrix} \hat{\alpha}_0 \\ \hat{\alpha}_1 \\ \dots \\ \hat{\alpha}_k \end{bmatrix} \text{ jest wektorem estymatorów parametrów równania regresji,}$$

n jest liczbą obserwacji,

k jest liczbą zmiennych objaśniających w modelu.

Za **estymator wariancji składnika losowego** ε równania regresji przyjmuje się

$$S_\varepsilon^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - k - 1} = \frac{\sum_{i=1}^n e_i^2}{n - k - 1},$$

a za **estymatory wariancji i kowariancji współczynników regresji** elementy leżące odpowiednio na i poza główną przekątną macierzy

$$S^2(\hat{\alpha}) = \begin{bmatrix} d_{00} & d_{01} & \dots & d_{0k} \\ d_{10} & d_{11} & \dots & d_{1k} \\ \dots & \dots & \dots & \dots \\ d_{k0} & d_{k1} & \dots & d_{kk} \end{bmatrix} = S_\varepsilon^2 (X^T X)^{-1}.$$

Wyznaczony metodą najmniejszych kwadratów model regresji liniowej

$$\hat{y} = \hat{\alpha}_0 + \hat{\alpha}_1 x_1 + \hat{\alpha}_2 x_2 + \dots + \hat{\alpha}_k x_k$$

musi być poddany weryfikacji statystycznej pod względem dopasowania modelu do danych empirycznych oraz pod względem istotności współczynników modelu.

Podstawowe miary **dopasowania modelu do danych rzeczywistych** to **błąd standardowy składnika losowego** równania regresji $S_\varepsilon = \sqrt{S_\varepsilon^2}$ oraz **współczynnik determinacji** R^2 , gdzie

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

Im mniejsza wartość S_ε , tym model lepiej opisuje rzeczywistość. Wartości współczynnika znajdują się w przedziale [0,1]. Im wartość R^2 bliższa jedynki, tym model lepiej opisuje rzeczywistość.

W procesie weryfikacji modelu regresji liniowej, w pierwszej kolejności należy sprawdzić, czy zachodzi zależność liniowa między zmienną objaśnianą y , a którąkolwiek ze zmiennych objaśniających x_i modelu. W tym celu należy wykonać **test istotności układu współczynników regresji**. Stawiane hipotezy:

$$H_0 : \sum_{j=1}^k \alpha_j^2 = 0$$

$$H_1 : \sum_{j=1}^k \alpha_j^2 \neq 0$$

Sprawdzianem tak postawionych hipotez jest statystyka:

$$F = \frac{R^2}{1 - R^2} \cdot \frac{n - k - 1}{k},$$

która przy prawdziwości hipotezy zerowej ma rozkład F Snedecora o k stopniach swobody licznika oraz o $(n-k-1)$ stopniach swobody mianownika. Obszar krytyczny jest postaci

$$K = \{F : F \geq F_{k, n-k-1, 1-\alpha}\}.$$

W poprawnym modelu regresji liniowej zmienna objaśniana y musi istotnie zależeć od każdej ze zmiennych objaśniających x_i modelu. Test weryfikujący ten fakt to **test istotności poszczególnych współczynników regresji**. Dla każdego parametru równania regresji α_j ($j = 0, 1, \dots, k$) stawiane są hipotezy:

$$H_0 : \alpha_j = 0$$

$$H_1 : \alpha_j \neq 0$$

Sprawdzianem tak postawionych hipotez jest statystyka:

$$t = \frac{\hat{\alpha}_j}{S(\hat{\alpha}_j)},$$

która przy prawdziwości hipotezy zerowej ma rozkład t Studenta o $(n-k-1)$ stopniach swobody. Obszar krytyczny jest postaci

$$K = \left\{ t : |t| \geq t_{n-k-1, 1-\frac{\alpha}{2}} \right\}.$$

Zadania do wykonania

1. Oszacowano model regresji liniowej zmiennej Y względem X_1 oraz X_2 . Który ma postać $\hat{y} = 2.2 - 1.8x_1 - 0.4x_2$. Wskazać zmienną objaśnianą i zmienne objaśniające, zinterpretować współczynniki modelu, obliczyć wartości składników losowych, jeżeli wiadomo, że:

X_1	X_2	Y
1	0	0
0	1	1
1	2	0
0	0	3

2. Skonstruować, zweryfikować, (w razie potrzeby) poprawić oraz zinterpretować model regresji liniowej dla danych w pliku *dane.xls*.

Źródła:

- Magiera R. „Modele i metody statystyki matematycznej” GiS, Wrocław 2002
- Walesiak M., Gatnar E. i inni „Statystyczna analiza danych z wykorzystaniem pakietu R”, PWN, Warszawa 2009
- Gładysz B., Mercik J. „Modelowanie ekonometryczne. Studium przypadku” Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław 2004
- Dziechciarz J. „Ekonometria. Metody, przykłady, zadania”, Wydawnictwo Akademii Ekonomicznej Im. Oskara Lange we Wrocławiu, Wrocław 2003