Regresja liniowa

Przypadek wielu zmiennych.

Sebastian Zalas

FAME|GRAPE, Uniwersytet Warszawski

Ekonometria 2022/23

Model regresji liniowej

Załóżmy, że zjawisko ekonomiczne można opisać modelem postaci:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \ldots + u_i,$$

gdzie:

- i - indeks obserwacji, $i = 1, ..., n$;
- y_i - zmienna zależna, objaśniana;
- $x_{1,i}$, $x_{2,i}$, . . . ,$x_{k,i}$ - zmienne niezależne, objaśniające;
- β_0, β_1, β_2, . . . β_k - nieznane (prawdziwe) parametry modelu
- u_i - składnik losowy.

Oszacowanie MNK modelu liniowego

Korzystając z MNK otrzymujemy oszacowanie:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \hat{\beta}_2 x_{2,i} + \ldots + \hat{\beta}_k x_{k,i} + u_i,$$

gdzie:

- i - indeks obserwacji, $i = 1, ..., n$;
- $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, ... ,$\hat{\beta}_k$- oszacowania nieznanych parametrów modelu
- u_i - składnik losowy.
- wartości teoretyczne: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \hat{\beta}_2 x_{2,i} + \ldots + \hat{\beta}_k x_{k,i}$
- reszty: $\hat{u}_i = y_i - \hat{y}_i$

MNK - wyprowadzenie

Minimalizujemy sumę kwadratów reszt:

$$\min \sum_{i=1}^{n} (\hat{u}_i)^2$$

czyli różnic pomiędzy wartościami obserwowanymi i teoretycznymi:

$$\min_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1,i} - \hat{\beta}_2 x_{2,i} \ldots - \hat{\beta}_k x_{k,i} \right)^2$$

Warunki pierwszego rzędu:

$$\frac{\partial SST}{\partial \hat{\beta}_0} = \sum_{i=1}^{n} -2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1,i} - \hat{\beta}_2 x_{2,i} \ldots - \hat{\beta}_k x_{k,i}) = 0$$

$$\frac{\partial SST}{\partial \hat{\beta}_k} = \sum_{i=1}^{n} -2x_{k,i}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1,i} - \hat{\beta}_2 x_{2,i} \ldots - \hat{\beta}_k x_{k,i}) = 0 \quad \forall_{k \geqslant 1}$$

Możemy to zapisać w wersji macierzowej!

MNK - notacja macierzowa

- Mamy n zmiennych oraz liczbę zmiennych objaśniających równą k.
- Oznaczmy: macierz zmiennych objaśniających jako $\boldsymbol{X}$ (o wymiarach $(n \times (k+1))$; wektor zmiennej objaśnianej jako $\boldsymbol{y}$ (o wymiarach $n \times 1$), wektor parametrów modelu jako β (o wymiarach $(k+1) \times 1$) oraz wektor składnika losowego jako $\boldsymbol{u}$ (o wymiarach $n \times 1$).
- Wtedy:

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{u}$$

- wektor reszt modelu:

$$\hat{\boldsymbol{u}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\hat{\beta}$$

MNK - wyprowadzenie w wersji macierzowej

- Minimalizujemy sumę kwadratów reszt:

$$\min_{\hat{\beta}}\{\hat{\boldsymbol{u}}^T\hat{\boldsymbol{u}}\}$$

$$\min_{\hat{\beta}}\{(\boldsymbol{y}-\hat{\boldsymbol{y}})^T(\boldsymbol{y}-\hat{\boldsymbol{y}})\}$$

$$\min_{\hat{\beta}}\{(\boldsymbol{y}-\boldsymbol{X}\hat{\beta})^T(\boldsymbol{y}-\boldsymbol{X}\hat{\beta})\}$$

po przemnożeniu:

$$\min_{\hat{\beta}}\{\boldsymbol{y}\boldsymbol{y}^T-2\boldsymbol{y}^T\boldsymbol{X}\hat{\beta}+\hat{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\hat{\beta}\}$$

- Warunek pierwszego rzędu przyrównujemy do zera i otrzymujemy układ równań:

$$-2\boldsymbol{X}^T\boldsymbol{y}+2\boldsymbol{X}^T\boldsymbol{X}\hat{\beta}=0$$

- estymator MNK:

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

Współczynnik determinacji R^2

Współczynnik determinacji R^2:

- jest definiowany jako stosunek wariancji/zmienności zmiennej zależnej wyjaśnionej przez model co całkowitej wariancji zmiennej zależnej $\hat{y}$ do wariancji z próby y
- nie jest wyznacznikiem jakości modelu

Przypomnijmy:

$$y_i = \hat{y}_i + \hat{u}_i$$

Analogiczna zależność dotyczy wariancji:

$$SST = SSE + SSR$$

gdzie:

- SST (*Sum of Squares Total*) - całkowita wariancja zmiennej zależnej (y_i)
- SSE (*Sum of Squares Explained*) - wariancja zmiennej zależnej, wyjaśniona przez model (wariancja $\hat{y}_i$)
- SSR (*Sum of Squares Residual*) - wariancja reszt ($\hat{u}_i$)

Współczynnik determinacji R^2

- Wariancja y_i

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- Wariancja $\hat{y}_i$

$$SSE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

- Wariancja $\hat{u}_i$

$$SSR = \sum_{i=1}^{n} \hat{u}_i{}^2$$

Teraz możemy zdefiniować R^2:

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n} \hat{u}_i{}^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Współczynnik determinacji R^2

Cechy R^2:

- nigdy nie zmniejszy się, jeśli w modelu pojawi się dodatkowa zmienna.
- $R^2 \in (0, 1)$
- model musi zawierać wyraz wolny

Ponieważ standardowy R^2 promuje modele z większą liczbą zmiennych objaśniających, skonstruowano $\bar{R}^2$ - skorygowany R^2:

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1}(1 - R^2)$$

- $\bar{R}^2$ *karze* za dodawanie kolejnych zmiennych objaśniających do modelu, poprzez uwzględnienie liczby stopni swobody.
- $\bar{R}^2$ będzie zawsze mniejszy lub równy R^2.
- nie ma standardowej interpretacji (takiej jak R^2).

Pytania? Wątpliwości?
Dziękuję!

e: s.zalas@uw.edu.pl