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;
- $x_{1,i}, x_{2,i}, \ldots, x_{k,i}$ - zmienne niezależne, objaśniające;
- $\beta_0, \beta_1, \beta_2, \ldots \beta_k$ - 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:

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

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

Możemy to zapisać w wersji macierzowej!

MNK - notacja macierzowa

- Mamy n zmiennych orazoraz 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:

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

Miara dopasowania R^2

Kiedy oszacowaliśmy model, możemy zastanowić się w jaki sposób ocenić jego dopasowanie do danych $\Rightarrow R^2$

- definiowany jako stosunek wariancji $\hat{y}$ do wariancji z próby y
- interpretacja: jaka część wariancji zmiennej zależnej y została wyjaśniona przez model

Przypomnijmy:

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

Wariancja zmiennej zależnej (*Total Sum of Squares*):

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

Miara dopasowania R^2

Możemy rozłożyć TSS na

- Wariancję $\hat{y}_i$ (*Explained Sum of Squares*)

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

- Wariancję $\hat{u}_i$ (*Residual Sum of Squares*)

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

Teraz możemy zdefiniować R^2:

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

ponieważ $TSS = ESS + RSS$.

Miara dopasowania R^2

- R^2 nigdy nie zmniejszy się, jeśli dołożymy do modelu dodatkową zmienną.
- Nawet w sytuacji, gdy dodana zmienna nie będzie potrzebna.

- $\bar{R}^2$ - skorygowany R^2

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

- $\bar{R}^2$ będzie zawsze mniejszy lub równy R^2
- $\bar{R}^2$ *karze* za dodawanie kolejnych zmiennych obajśniających do modelu

Pytania? Wątpliwości?

Dziękuję!

e: s.zalas@uw.edu.pl