Klasyczny Model Regresji Liniowej

Wykład 8

Natalia Nehrebecka Stanisław Cichocki

7 grudnia 2014

Plan zajęć

1. $y_i = x_{1i}\beta_1 + \cdots + x_{Ki}\beta_K + \varepsilon_i$, zapisie macierzowym: $\mathbf{y} = \mathbf{X}\beta + \varepsilon$
2. Zmienne objaśniające $x_{1i}, \ldots, x_{ki}$ są nielosowe dla $i = 1, \ldots, N$
3. $E(\varepsilon_i) = 0$ dla $i = 1, \ldots, N$, zapisie macierzowym: $E(\varepsilon) = \mathbf{0}$
4. $Cov(\varepsilon_i, \varepsilon_j) = 0$ dla $i \neq j$ (brak *autokorelacji* błędu losowego)
5. $Var(\varepsilon_i) = \sigma^2$ dla $i = 1, \ldots, N$ (*homoskedastyczność* $\neq$*heteroskedastyczność* błędu losowego)

1. $y_i = x_{1i}\beta_1 + \cdots + x_{Ki}\beta_K + \varepsilon_i$, zapisie macierzowym: $\mathbf{y} = \mathbf{X}\beta + \varepsilon$
2. Zmienne objaśniające $x_{1i}, \ldots, x_{ki}$ są nielosowe dla $i = 1, \ldots, N$
3. $E(\varepsilon_i) = 0$ dla $i = 1, \ldots, N$, zapisie macierzowym: $E(\varepsilon) = \mathbf{0}$
4. $Cov(\varepsilon_i, \varepsilon_j) = 0$ dla $i \neq j$ (brak *autokorelacji* błędu losowego)
5. $Var(\varepsilon_i) = \sigma^2$ dla $i = 1, \ldots, N$ (*homoskedastyczność* $\neq$*heteroskedastyczność* błędu losowego)

1. $y_i = x_{1i}\beta_1 + \dots + x_{Ki}\beta_K + \varepsilon_i$, zapisie macierzowym:
   $\mathbf{y} = \mathbf{X}\beta + \varepsilon$
2. Zmienne objaśniające $x_{1i}, \dots, x_{ki}$ są nielosowe dla $i = 1, \dots, N$
3. $E(\varepsilon_i) = 0$ dla $i = 1, \dots, N$, zapisie macierzowym: $E(\varepsilon) = \mathbf{0}$
4. $Cov(\varepsilon_i, \varepsilon_j) = 0$ dla $i \neq j$ (brak *autokorelacji* błędu losowego)
5. $Var(\varepsilon_i) = \sigma^2$ dla $i = 1, \dots, N$ (*homoskedastyczność* $\neq$*heteroskedastyczność* błędu losowego)

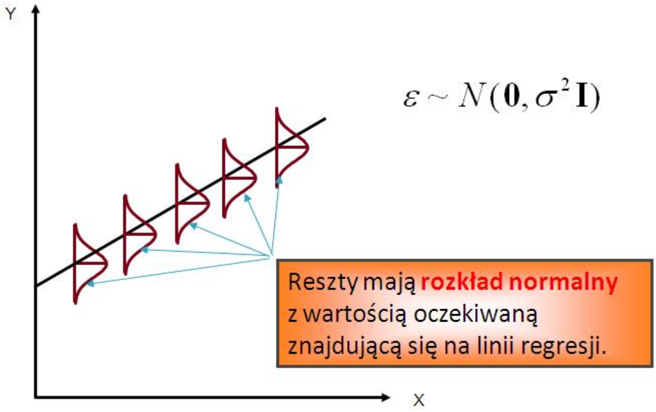
1. $y_i = x_{1i}\beta_1 + \cdots + x_{Ki}\beta_K + \varepsilon_i$, zapisie macierzowym: $\mathbf{y} = \mathbf{X}\beta + \varepsilon$
2. Zmienne objaśniające $x_{1i}, \ldots, x_{ki}$ są nielosowe dla $i = 1, \ldots, N$
3. $E(\varepsilon_i) = 0$ dla $i = 1, \ldots, N$, zapisie macierzowym: $E(\varepsilon) = \mathbf{0}$
4. $Cov(\varepsilon_i, \varepsilon_j) = 0$ dla $i \neq j$ (brak *autokorelacji* błędu losowego)
5. $Var(\varepsilon_i) = \sigma^2$ dla $i = 1, \ldots, N$ (*homoskedastyczność* $\neq$*heteroskedastyczność* błędu losowego)

1. $y_i = x_{1i}\beta_1 + \cdots + x_{Ki}\beta_K + \varepsilon_i$, zapisie macierzowym: $\mathbf{y} = \mathbf{X}\beta + \varepsilon$
2. Zmienne objaśniające $x_{1i}, \ldots, x_{ki}$ są nielosowe dla $i = 1, \ldots, N$
3. $E(\varepsilon_i) = 0$ dla $i = 1, \ldots, N$, zapisie macierzowym: $E(\varepsilon) = \mathbf{0}$
4. $Cov(\varepsilon_i, \varepsilon_j) = 0$ dla $i \neq j$ (brak *autokorelacji* błędu losowego)
5. $Var(\varepsilon_i) = \sigma^2$ dla $i = 1, \ldots, N$ (*homoskedastyczność* $\neq$*heteroskedastyczność* błędu losowego)

- Zaburzenia losowe ε są sferyczne.

$$Var(\varepsilon) = \sigma^2 I = \begin{bmatrix} Var(\varepsilon_1) & Cov(\varepsilon_2, \varepsilon_1) & \cdots & Cov(\varepsilon_N, \varepsilon_1) \\ Cov(\varepsilon_1, \varepsilon_2) & Var(\varepsilon_2) & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ Cov(\varepsilon_1, \varepsilon_N) & Cov(\varepsilon_2, \varepsilon_N) & \cdots & Var(\varepsilon_N) \end{bmatrix}$$

$$= \begin{bmatrix} \sigma^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma^2 \end{bmatrix}$$



$$\varepsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$

1. **Nieobciążoność**: $E(b) = \beta$
2. **Efektywność**:
   Twierdzenie (Gaussa-Markowa) *Dla spełnionych założeń KMRL estymator MNK jest najlepszym estymatorem parametrów β w klasie liniowych i nieobciążonych estymatorów tego parametru.*

- Wektor oszacowań (estymator) parametrów jest wektorem **losowym**, może więc odbiegać od prawdziwych wartości parametrów
- Precyzję oszacowania mierzy się za pomocą miar rozrzutu, dyspersji
- Najpopularniejszą miarą rozrzutu jest macierz **wariancji-kowariancji**

Wyprowadzenie wzoru

- Ważne założenia KMRL: *brak autokorelacji* ($cov(\varepsilon_i, \varepsilon_j) = 0$ dla $i \neq j$) i *homoskedastyczność* ($var(\varepsilon_i) = \sigma^2$ dla $i = 1, \ldots, N$)

$$Var(\mathbf{b}) = Var((X'X)^{-1}X'y) = Var((X'X)^{-1}X'(X\beta + \varepsilon)$$

$$= Var(\beta + (X'X)^{-1}X'\varepsilon) = (X'X)^{-1}X' \underbrace{Var(\varepsilon)}_{\sigma^2 I} X(X'X)^{-1}$$

$$= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} = \sigma^2(X'X)^{-1} = \Sigma$$

$$\Sigma = \begin{bmatrix} Var(b_1) & \cdots & cov(b_p, b_q) \\ \vdots & \ddots & \vdots \\ cov(b_q, b_p) & \cdots & Var(b_k) \end{bmatrix}$$

- Oszacowanie wariancji estymatora parametrów Σ uzyskamy, jeśli uda się znaleźć estymator wariancji błędu losowego σ^2

- Oszacowanie błędów losowych w MNK - reszty
- Dlatego estymator wariancji błędów losowych buduje się na podstawie wariancji reszt

Nieobciążony estymator wariancji

Ponieważ

$$E(\frac{\mathbf{e}'\mathbf{e}}{N-K}) = \sigma^2$$

Dlatego nieobciążonym estymatorem wariancji błędu losowego jest

$$s^2 = \frac{\mathbf{e}'\mathbf{e}}{N-K}$$

$$\hat{\Sigma} = s^2(\mathbf{X}'\mathbf{X})^{-1}$$

```
. xi: regres ln_dochod wiek plec i.wyksztalcenie
i.wyksztalcenie   _Iwykształc_1-3    (naturally coded; _Iwyksztalc_1 omitted)
```

Source	SS	df	MS
Model	63.4299911	4	15.8574978
Residual	306.34918	1078	.284182913
Total	369.779172	1082	.341755242

Number of obs	=	1083
F(4, 1078)	=	55.80
Prob > F	=	0.0000
R-squared	=	0.1715
Adj R-squared	=	0.1685
Root MSE	=	.53309

ln_dochod	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
wiek	.004746	.0016325	2.91	0.004	.0015429	.0079492
plec	-.3240732	.0324819	-9.98	0.000	-.3878081	-.2603383
_Iwykształ~2	.2725003	.0521834	5.22	0.000	.1701078	.3748929
_Iwykształ~3	.715203	.0668967	10.69	0.000	.5839405	.8464654
_cons	6.181451	.0842574	73.36	0.000	6.016124	6.346778

- Nieobciążonym estymatorem wariancji błędu losowego jest:
- $s^2 = \frac{\mathbf{e}'\mathbf{e}}{N-K} = \frac{306{,}34918}{1083-5} = 0,284183$ (Mean Squared Error)
- $s = \sqrt{s^2} = \sqrt{0,284183} = 0,53309$ (Root of **M**ean **S**quared **E**rror)

```
. matrix list e(V)

symmetric e(V)[5,5]
                    wiek          plec  _Iwykszta1~2  _Iwykszta1~3         _cons
        wiek   2.665e-06
        plec  -3.327e-06     .00105507
_Iwykszta1~2   .00001157     -.0000324     .00272311
_Iwykszta1~3   4.674e-06    -.00003634     .00235018     .00447517
       _cons     -.00011    -.00035369    -.00279848    -.00250806     .00709931
```

- estymator macierzy wariancji kowariancji estymatora b: $\hat{\Sigma} = s^2(\mathbf{X}'\mathbf{X})^{-1}$

- $se(b_k)$ - oszacowanie odchylenia błędu standardowego b_k, które wykorzystuje się do mierzenia precyzji oszacowań parametrów.

$$\widehat{se}(b_k) = \sqrt{\left[\hat{\Sigma}_b\right]_{kk}}$$

- Model dla dochodu:

$$log(dochod) = \beta_1 + \beta_2 wiek + \beta_3 plec + \beta_4 srednie + \beta_5 wyzsze + \varepsilon$$

```
. matrix list e(V)

symmetric e(V)[5,5]
                    wiek         plec  _Iwykszta1~2  _Iwykszta1~3        _cons
        wiek   2.665e-06
        plec  -3.327e-06    .00105507
_Iwykszta1~2   .00001157   -.0000324     .00272311
_Iwykszta1~3   4.674e-06  -.00003634     .00235018     .00447517
       _cons    -.00011   -.00035369    -.00279848    -.00250806    .00709931
```

- Oszacowanie β_2 równe $b_2 = 0{,}004746$
- Oszacowanie $Var(b_2) = 2,665e - 06$
- Oszacowanie $se(b_2) = 0,0016325$

- Poniewa estymator b jest liniowy, to analogiczne własnosci posiada również dowolna kombinacja liniowa wektora b.
- Weźmy na przykład wektor złożony ze stałych v o nie wszystkich elementach jednoczesnie równych 0:
  - $v = \begin{bmatrix} v_1 \\ \vdots \\ v_k \end{bmatrix}$
- i utwórzmy kombinację liniową wektora v i b, a więc:
  - $v'b = v_1 b_1 + ... + v_k b_k$.
- Ta kombinacja liniowa jest również najlepszym liniowym i nieobciążonym estymatorem kombinacji liniowej $v'\beta$.

1. Wymienić założenia Klasycznego Modelu Regresji Liniowej (KMRL).
2. Udowodnić, że, w KMRL estymator **b** jest nieobciążony.
3. Wyprowadzić postać macierzy wariancji kowariancji **b** i podać interpretację jej elementów.
4. Podać (słowami) treść twierdzenia Gaussa-Markowa i wyjaśnić jego znaczenie.
5. (*) Udowodnić twierdzenie Gaussa-Markowa.
6. Pokazać, że s^2 jest nieobciążonym estymatorem σ^2.
7. Udowodnić, że $s^2(X'X)^{-1}$ jest nieobciążonym estymatorem $\mathbf{Var(b)}$.