Statystyczne własności estymatora MNK & heteroskedastyczność

Ekonometria WNE

Sebastian Zalas

University of Warsaw
s.zalas@uw.edu.pl

KLASYCZNY MODEL REGRESJI LINIOWEJ

1. $\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon$ - model jest liniowy
2. Zmienne losowe $\{(y_1, X_1), ..., (y_i, y_i), ..., (y_n, y_n)\}$ są niezależne oraz wylosowane z tego samego rozkładu (*independently and identically distributed - iid.*)
3. $rz[\boldsymbol{X}_{n\times k}] = k$ - rząd kolumnowy $\boldsymbol{X}$ jest pełny
4. $\mathbb{E}[\varepsilon|\boldsymbol{X}] = 0$ - wartość oczekiwana składnika losowego jest równa 0
5. $\mathbb{E}[\varepsilon\varepsilon'|\boldsymbol{X}] = \boldsymbol{I}\sigma^2$ - sferyczność wariancji
6. $\varepsilon|\boldsymbol{X} \sim \mathcal{N}(0, \boldsymbol{I}\sigma^2)$ składnik losowy ma rozkład normalny

Własności estymatora

- Estymator $\hat{\theta}$ nieznanego parametru θ jest funkcją danych, a więc jest jest on **zmienną losową**.

$$\hat{\theta} = g(X_1, ...X_n)$$

- Istnieje zatem wartość oczekiwana estymatora - $\mathbb{E}[\hat{\theta}]$ oraz jego wariancja - $\mathbb{V}[\hat{\theta}]$ (*sampling variance*)
- Estymator $\hat{\theta}$ parametru θ nazywamy **nieobciążonym** gdy:

$$\mathbb{E}[\hat{\theta}] = \theta$$

- **Obciążenie** estymatora

$$bias(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$$

Własności estymatora

- Estymator $\hat{\theta}$ nieznanego parametru θ jest funkcją danych, a więc jest jest on **zmienną losową**.

$$\hat{\theta} = g(X_1, ... X_n)$$

- Istnieje zatem wartość oczekiwana estymatora - $\mathbb{E}[\hat{\theta}]$ oraz jego wariancja - $\mathbb{V}[\hat{\theta}]$ (*sampling variance*)
- Estymator $\hat{\theta}$ parametru θ nazywamy **nieobciążonym** gdy:

$$\mathbb{E}[\hat{\theta}] = \theta$$

- **Obciążenie** estymatora

$$bias(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$$

Nieobciążoność

- Założenia (1) oraz (4):

$$
\begin{aligned}
\mathbb{E}[\varepsilon|\boldsymbol{X}] = 0 \\
&= \mathbb{E}[\boldsymbol{y} - \boldsymbol{X}\beta|\boldsymbol{X}] \\
&= \mathbb{E}[\boldsymbol{y}|\boldsymbol{X}] - \mathbb{E}[\boldsymbol{X}\beta|\boldsymbol{X}] \\
&= \mathbb{E}[\boldsymbol{y}|\boldsymbol{X}] - \boldsymbol{X}\beta \\
&\Rightarrow \mathbb{E}[\boldsymbol{y}|\boldsymbol{X}] = \boldsymbol{X}\beta
\end{aligned}
$$

Nieobciążoność

- Estymator uzyskany MNK $\hat{\beta}$ wektora parametrów β:

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

- Pokażmy, że estymator $\hat{\beta}$ jest nieobciążony, pw. że założenia (1) - (4) są spełnione:

$$\begin{aligned}
\mathbb{E}[\hat{\beta}|\boldsymbol{X}] &= \mathbb{E}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}|\boldsymbol{X}] \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \underbrace{\mathbb{E}[\boldsymbol{y}|\boldsymbol{X}]}_{\boldsymbol{X}\beta} \\
&= \underbrace{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}}_{=\boldsymbol{I}}\,\beta \\
&= \beta
\end{aligned}$$

Nieobciążoność

- Estymator uzyskany MNK $\hat{\beta}$ wektora parametrów β:
$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$
- Pokażmy, że estymator $\hat{\beta}$ jest nieobciążony, pw. że założenia (1) - (4) są spełnione:
$$\begin{aligned}
\mathbb{E}[\hat{\beta}|\boldsymbol{X}] &= \mathbb{E}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}|\boldsymbol{X}] \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \underbrace{\mathbb{E}[\boldsymbol{y}|\boldsymbol{X}]}_{\boldsymbol{X}\beta} \\
&= \underbrace{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}}_{=\boldsymbol{I}}\beta \\
&= \beta
\end{aligned}$$

Nieobciążoność

- Zdekomponujmy estymator MNK $\hat{\beta}$:

$$\begin{aligned}\hat{\beta} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}\\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\beta + \varepsilon)\\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\beta + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\varepsilon\\ &= \beta + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\varepsilon\end{aligned}$$

- Pokażmy nieobciążony korzystając z def. obciążenia:

$$\begin{aligned}\mathbb{E}[\hat{\beta} - \beta|\boldsymbol{X}] &= \mathbb{E}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\varepsilon|\boldsymbol{X}]\\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\,\mathbb{E}[\varepsilon|\boldsymbol{X}] = 0\end{aligned}$$

Nieobciążoność

- Zdekomponujmy estymator MNK $\hat{\beta}$:

$$\begin{aligned}\hat{\beta} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}\\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\beta + \varepsilon)\\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\beta + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\varepsilon\\ &= \beta + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\varepsilon\end{aligned}$$

- Pokażmy nieobciążony korzystając z def. obciążenia:

$$\begin{aligned}\mathbb{E}[\hat{\beta} - \beta|\boldsymbol{X}] &= \mathbb{E}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\varepsilon|\boldsymbol{X}]\\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\,\mathbb{E}[\varepsilon|\boldsymbol{X}] = 0\end{aligned}$$

Nieobciążoność

Twierdzenie. Nieobciążoność estymatora MNK.

W Klasycznym Modelu Regresji Liniowej (założenia (1) - (4)), estymator uzyskany MNK jest nieobciążony:

$$\mathbb{E}[\hat{\beta} \mid \boldsymbol{X}] = \beta$$

- $\mathbb{E}[\varepsilon \mid \boldsymbol{X}] = 0 \Rightarrow$ nieuwzględnienie ważnego czynnika w modelu $\Rightarrow$ jest obecny w $\varepsilon \Rightarrow$ oszacowania będą obciążone
- Rozkład warunkowy względem $\boldsymbol{X}$ - estymator jest nieobciążony dla każdej realizacji macierzy regresorów $\boldsymbol{X}$
- Warunkowy rozkład $\hat{\beta}$ jest skoncentrowany wokół β

Nieobciążoność

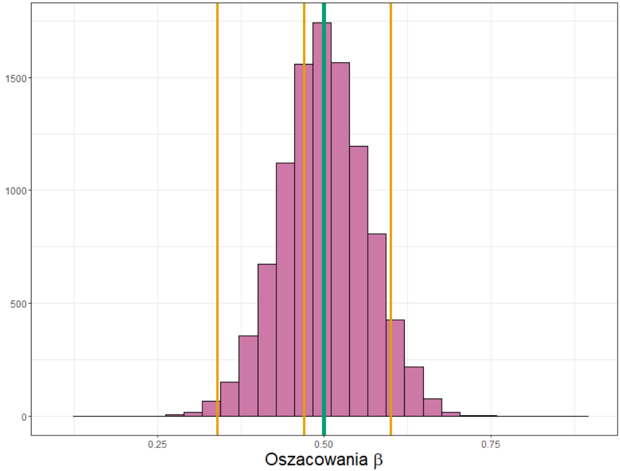
Twierdzenie. Nieobciążoność estymatora MNK.
W Klasycznym Modelu Regresji Liniowej (założenia (1) - (4)), estymator uzyskany MNK jest nieobciążony:

$$\mathbb{E}[\hat{\beta} \mid \boldsymbol{X}] = \beta$$

- $\mathbb{E}[\varepsilon \mid \boldsymbol{X}] = 0 \Rightarrow$ nieuwzględnienie ważnego czynnika w modelu $\Rightarrow$ jest obecny w $\varepsilon \Rightarrow$ oszacowania będą obciążone
- Rozkład warunkowy względem $\boldsymbol{X}$ - estymator jest nieobciążony dla każdej realizacji macierzy regresorów $\boldsymbol{X}$
- Warunkowy rozkład $\hat{\beta}$ jest skoncentrowany wokół β

NIEOBCIĄŻONOŚĆ: $\hat{\beta}$ VS. β

Symulacja modelu $y = 1 + 0.5x + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, 4)$



Wariancja estymatora MNK

- Definicja wariancji. Niech $\boldsymbol{Z}$ będzie wektorem losowym:

$$\mathbb{V}[\boldsymbol{Z}] = \mathbb{E}[(\boldsymbol{Z} - \mathbb{E}[\boldsymbol{Z}])(\boldsymbol{Z} - \mathbb{E}[\boldsymbol{Z}])^{\mathsf{T}}]$$

- Wariancja warunkowa:

$$\mathbb{V}[\boldsymbol{X} \mid \boldsymbol{Z}] = \mathbb{E}[(\boldsymbol{Z} - \mathbb{E}[\boldsymbol{Z}])(\boldsymbol{Z} - \mathbb{E}[\boldsymbol{Z}])^{\mathsf{T}} \mid \boldsymbol{X}]$$

- Wariancja składnika losowego:

$$\mathbb{V}[\varepsilon \mid \boldsymbol{X}] = \mathbb{E}[\varepsilon\varepsilon^{\mathsf{T}} \mid \boldsymbol{X}]$$

Wariancja estymatora MNK

- Przywołajmy fakt: $\hat{\beta} = \beta + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\varepsilon$
- Wyprowadzimy formułę wariancji estymatora MNK, czyli $\mathbb{V}[\hat{\beta}]$:

$$\begin{aligned}
\mathbb{V}[\hat{\beta} \mid \boldsymbol{X}] &= \mathbb{E}[(\beta - \hat{\beta})(\beta - \hat{\beta})^\top \mid \boldsymbol{X}] \\
&= \mathbb{E}[((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\varepsilon)((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\varepsilon)^\top \mid \boldsymbol{X}] \\
&= \mathbb{E}[(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\varepsilon\varepsilon^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} \mid \boldsymbol{X}] \\
&= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathbb{E}[\varepsilon\varepsilon^\top \mid \boldsymbol{X}]\,\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}
\end{aligned}$$

- wariancja estymatora $\hat{\beta}$ zależy od $\mathbb{E}[\varepsilon\varepsilon^\top \mid \boldsymbol{X}] = \mathbb{V}[\varepsilon \mid \boldsymbol{X}]$

Wariancja estymatora MNK

- W KMRL zakładamy sferyczność wariancji - $\mathbb{E}[\varepsilon\varepsilon'|\boldsymbol{X}] = \boldsymbol{I}\sigma^2$, czyli:

$$\mathbb{V}[\varepsilon \mid \boldsymbol{X}] = \mathbb{E}[\varepsilon\varepsilon' \mid \boldsymbol{X}] = \boldsymbol{\Omega} = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix} = \boldsymbol{I}\sigma^2$$

- W takim przypadku, wariancja **estymatora MNK** przyjmuje postać:

$$\mathbb{V}[\hat{\beta} \mid \boldsymbol{X}] = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\sigma^2\boldsymbol{I}\,\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$$

WARIANCJA ESTYMATORA MNK

- Sferyczność wariancji - co to oznacza?

$$\mathbb{V}[\varepsilon \mid \boldsymbol{X}] = \mathbb{E}[\varepsilon\varepsilon' \mid \boldsymbol{X}] = \boldsymbol{\Omega} = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix} = \boldsymbol{I}\sigma^2$$

- **homoskedastyczność** $\Rightarrow$ stałość wariancji $\Rightarrow$ takie same elementy na diagonali macierzy $\boldsymbol{\Omega}$
- brak autokorelacji $\Rightarrow$ zera poza diagonalą macierzy $\boldsymbol{\Omega}$

Gauss-Markov

Twierdzenie Gaussa Markova.

W Klasycznym Modelu Regresji Liniowej (założenia (1) - (5)), nieobciążony estymator uzyskany MNK ma najniższą wariancję:

$$\mathbb{V}[\tilde{\beta} \mid \boldsymbol{X}] \geq \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}$$

- Żaden inny nieobciążony estymator nie może mieć niższej wariancji niż $\sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}$
- Estymator MNK jest najefektywniejszy w klasie liniowych nieobciążonych estymatorów - **BLUE** - *Best Linear Unbiased Estimator*

Wariancja ε

- $\mathbb{V}[\varepsilon] = \sigma^2$ - nie znamy $\sigma^2 \Rightarrow$ nie znamy $\mathbb{V}[\hat{\beta}]$
- Należy oszacować σ^2:
$$s^2 = \frac{1}{n-k}\sum_{i=1}^{n} e_i^2 = \frac{\boldsymbol{e}'\boldsymbol{e}}{n-k}$$
nieobciążony estymator σ^2
- Estymator wariancji $\hat{\beta}$:
$$\widehat{\mathbb{V}[\hat{\beta} \mid \boldsymbol{X}]} = s^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1} = \frac{\boldsymbol{e}'\boldsymbol{e}}{n-k}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} = \widehat{\boldsymbol{V}}_{\hat{\beta}}$$

Heteroskedastyczność ε

- Analizowane dane mogą jednak nie spełniać założenia o sferyczności składnika losowego. Wtedy jego wariancja ma postać:

$$\mathbb{V}[\varepsilon \mid \boldsymbol{X}] = \mathbb{E}[\varepsilon\varepsilon^\top \mid \boldsymbol{X}] = \boldsymbol{\Omega} = \begin{bmatrix} \sigma_1^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_n^2 \end{bmatrix}$$

  - heteroskedastyczność $\Rightarrow$ brak stałości wariancji $\Rightarrow$ różne elementy na diagonali macierzy $\boldsymbol{\Omega}$
  - utrzymujemy założenia o braku autokorelacji $\Rightarrow$ zera poza diagonalą macierzy $\boldsymbol{\Omega}$

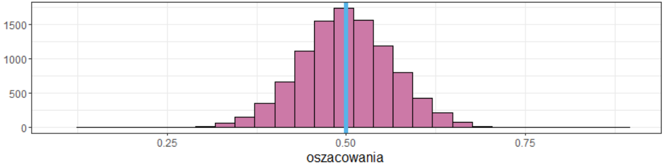
- W takim przypadku, wariancja **estymatora MNK** przyjmuje postać:

$$\mathbb{V}[\hat{\beta} \mid \boldsymbol{X}] = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Omega}\,\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$$

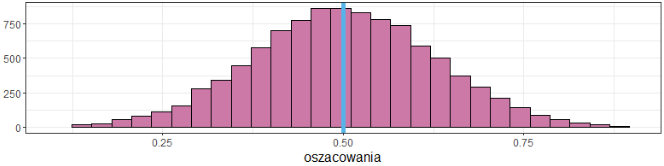
Heteroskedastyczność ε - konsekwencje

- Nieobciążoność $\hat{\beta}$ pozostaje nienaruszona.
- Jednak estymator $\hat{\beta}$ będzie **nieefektywny** tzn. możemy znaleźć estymator o niższej wariancji w klasie liniowych estymatorów
- Oznacza to, że precyzja estymatora się zmniejsza, co wypływa także na jakość wnioskowania statystycznego z oszacowanego modelu.

Homo- vs hetero- skedastyczność



(a) Rozkład $\hat{\beta}$ w modelu z **homoskedastycznym** składnikiem losowym.



(b) Rozkład $\hat{\beta}$ w modelu z **heteroskedastycznym** składnikiem losowym.

Heteroskedastyczność - estymator odporny

- Gdy występuje heteroskedastyczności, zwykły estymator wariancji może być obciążony $\Rightarrow$ musimy skonstruować taki estymator wariancji, który będzie odporny na heteroskedastyczność

- Idealnie byłoby, gdybyśmy mogli mieć:

$$\begin{aligned}\widehat{\boldsymbol{V}}_{\hat{\beta}}^{ideal} &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\,\mathbb{E}[\varepsilon\varepsilon^\top \mid \boldsymbol{X}]\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\\ &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top \begin{bmatrix}\varepsilon_1^2 & \dots & 0\\ \vdots & \ddots & \vdots\\ 0 & \dots & \varepsilon_n^2\end{bmatrix}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\end{aligned}$$

- Otrzymanie takiego estymatora nie jest możliwe...

Heteroskedastyczność - estymator White'a

- White (1980) pokazał, że poniższy estymator wariancji estymatora MNK:

$$\widehat{\boldsymbol{V}}_{\hat{\beta}}^{HC0} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \left(\sum_{i=1}^{n} X_i X_i' e_i^2 \right) (\boldsymbol{X}^\top \boldsymbol{X})^{-1}$$

  jest zgodnym estymatorem $\mathbb{V}[\hat{\beta} \mid \boldsymbol{X}]$, przez co jest odporny na heteroskedastyczność.

- HC - *Heteroscedasticity consistent*, czasami mówi się *heteroscedasticity robust*

DLACZEGO $\mathbb{V}[\varepsilon \mid \boldsymbol{X}]$ JEST TAK WAŻNA?

- Wariancja składnika losowego, $\mathbb{V}[\varepsilon]$ decyduje o wariancji estymatora, $\mathbb{V}[\hat{\beta}]$
- Błąd standardowy $\hat{\beta}$:

$$se[\hat{\beta}] = \sqrt{\mathbb{V}[\hat{\beta}]} = \begin{bmatrix} \sqrt{\beta_0} & 0 & \dots & 0 \\ 0 & \sqrt{\beta_1} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sqrt{\beta_k} \end{bmatrix}$$

- Na podstawie $\hat{\beta}$ będziemy testować istotność oszacowań $\Rightarrow$ niepoprawna wariancja ε prowaddzi do niepoprawnego wnioskowania

Test Breusch'a - Pagan'a

- Hipoteza zerowa

$$H_0 : \mathbb{E}[\varepsilon^2 \mid x_1, x_2, \ldots, x_k] = \mathbb{E}[\varepsilon^2] = \sigma^2.$$

- Oszacuj model, uzyskaj kwadraty reszt, e^2:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + e$$

- Oszacuj model, policz $R^2_{e^2}$ z tej regresji:

$$e^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \ldots + +\delta_k x_k + u$$

- Oblicz statystykę testową, uzyskaj *p-value*:

$$F = \frac{R^2_{e^2} \frac{1}{k}}{(1 - R^2_{e^2}) \frac{1}{n-k-1}} \sim F_{k,n-k-1}$$

lub skorzystaj ze statystyki:

$$LM = n \times R^2_{\widehat{u}^2} \sim \chi^2_k$$

Test White'a

- Hipoteza zerowa

$$H_0 : \mathbb{E}[\varepsilon^2 \mid x_1, x_2, \ldots, x_k] = \mathbb{E}[\varepsilon^2] = \sigma^2.$$

- Oszacuj model, uzyskaj kwadraty reszt, e^2:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + e$$

- Oszacuj model (z kwadratami i interakcjami), policz $R^2_{e^2}$ z tej regresji:

$$e^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \ldots + \delta_k x_k + \sum_{i=1}^{k} \sum_{j=1}^{k} x_i x_j + u$$

- Oblicz statystykę i uzyskaj *p-value* tak jak w przypadku testu BP.

Pytania? Wątpliwości?

Dziękuję!

e: s.zalas@uw.edu.pl