Ćwiczenia laboratoryjne

Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie
AGH University of Science and Technology

Justyna Tora

Stacjonarność szeregów czasowych

Stacjonarność

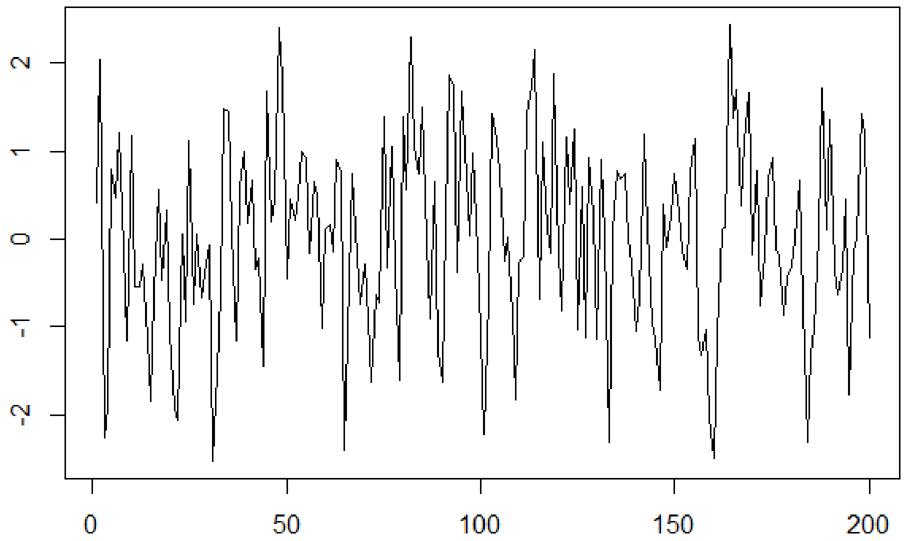
Szereg czasowy jest stacjonarny jeśli jego podstawowe własności nie zmieniają się w czasie. Wyróżnia się słabą stacjonarność (stacjonarność w szerszym sensie) oraz silną stacjonarność.

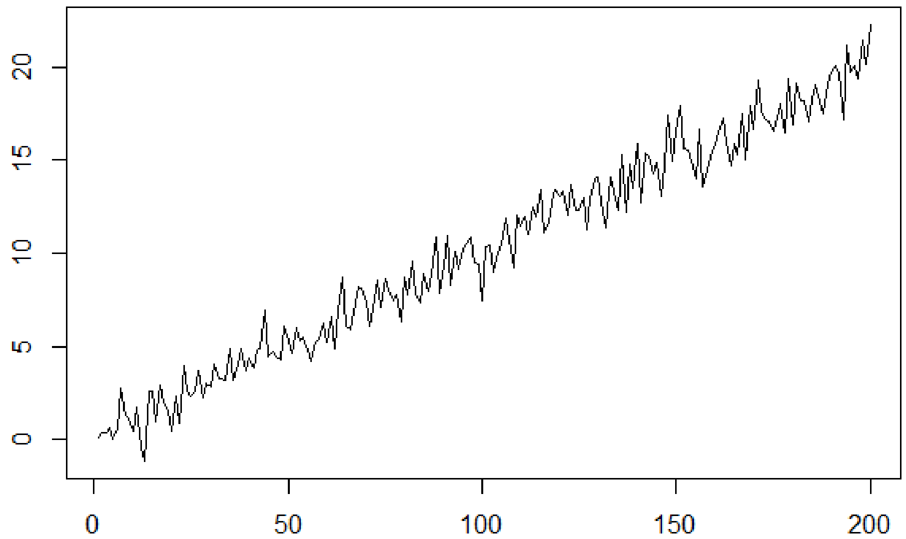
Proces jest **słabo stacjonarny** jeśli spełnione są następujące warunki:

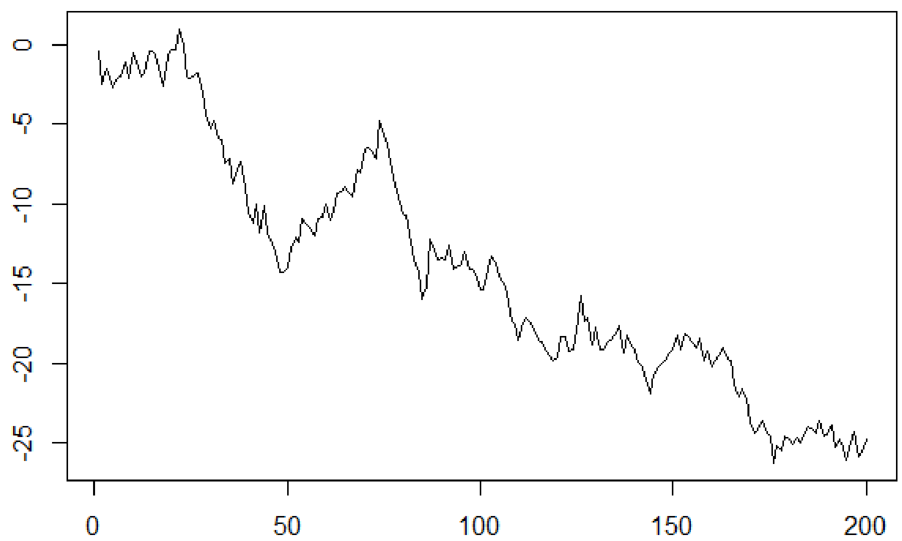
» Wartość oczekiwana procesu stochastycznego $E(X_t) = \mu$ jest stała i skończona.

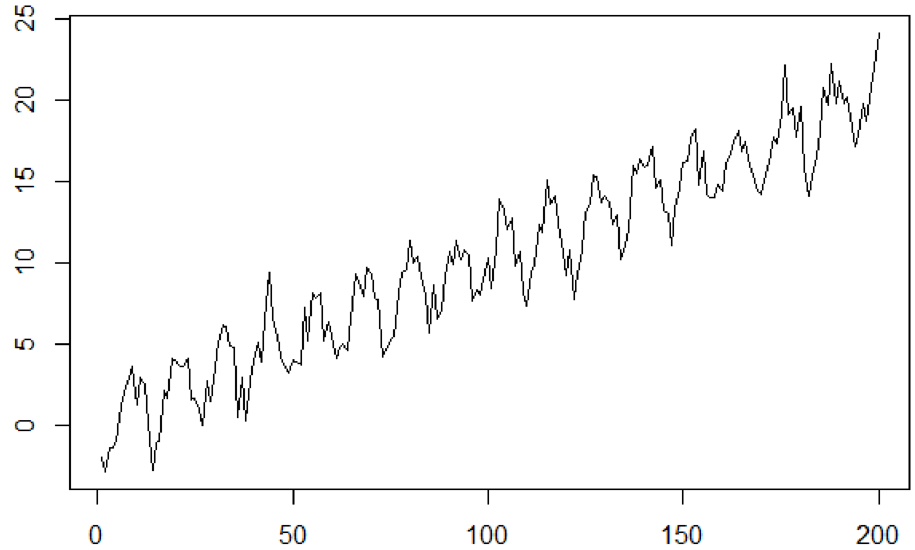
» Wariancja procesu stochastycznego $var(X_t) = \sigma^2$ jest stała i skończona.

» Funkcja autokowariancji zależy wyłącznie od opóźnienia: $cov(X_t, X_{t-k}) = \gamma(k)$


2
1
0
-1
-2
0
50
100
150
200


20
15
10
5
0
0
50
100
150
200


0
-5
-10
-15
-20
-25
0
50
100
150
200


25
20
15
10
5
0
0
50
100
150
200

Testy weryfikujące stacjonarność

- **Test ADF** (augmented Dickey-Fuller test)

Hipoteza zerowa: szereg czasowy nie jest stacjonarny

Hipoteza alternatywna: szereg czasowy jest stacjonarny

Funkcja w R: ***adf.test()*** z pakietu ***tseries***

- **Test KPSS** (Test Kwiatkowskiego, Phillipsa, Schmidta i Shina)

Hipoteza zerowa: szereg czasowy jest stacjonarny

Hipoteza alternatywna: szereg czasowy nie jest stacjonarny

Funkcja w R: ***kpss.test()*** z pakietu ***tseries***

Autokorelacja składnika losowego

Autokorelacja

Jednym z założeń MNK jest brak autokorelacji składnika losowego.

Model: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \varepsilon_i, \mathrm{dla\ i} = 1, \dots, \mathrm{n}$

Macierz wariancji-kowariancji składników losowych (w przypadku spełnionych odpowiednich założeń):

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

Jeśli założenie o braku autokorelacji składnika losowego nie jest spełnione, to macierz wariancji-kowariancji nie jest diagonalna. Wówczas estymator wektora parametrów modelu jest nieefektywny.

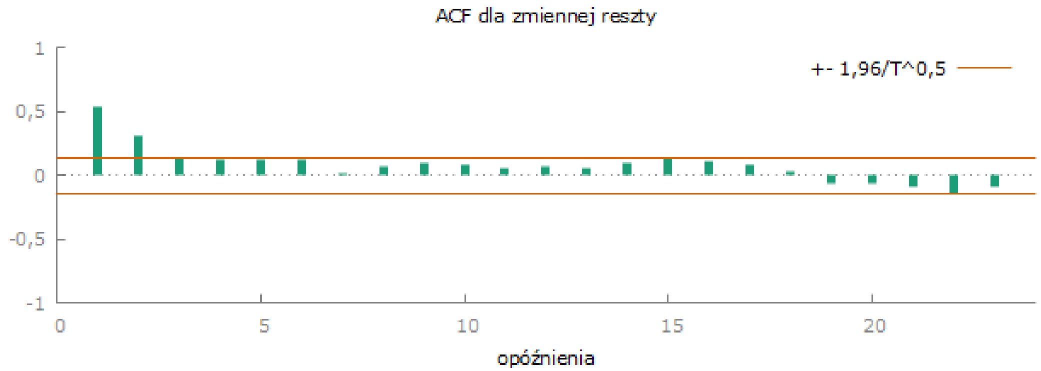
Analiza korelogramu

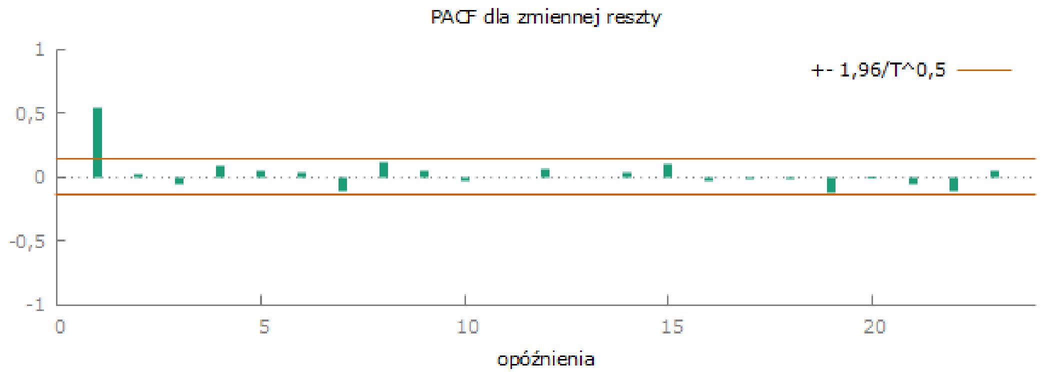
Wykres ACF (Autocorrelation Function) – wykres przedstawiający współczynniki autokorelacji dla kolejnych opóźnień.

Funkcja w R: ***acf()*** z pakietu ***stats***

Wykres PACF (Partial Autocorrelation Function) – wykres przedstawiający autokorelacje cząstkowe (bez zależności od elementów pośrednich).

Funkcja w R: ***pacf()*** z pakietu ***stats***





Testy weryfikujące autokorelację

- Test Durbina-Watsona (tylko do wykrywania autokorelacji I rzędu)

  Funkcja w R: ***dwtest()*** z pakietu ***lmtest***
- Test Breuscha-Godfreya

  Funkcja w R: ***bgtest()*** z pakietu ***lmtest***
- Test Ljung-Boxa

  Funkcja w R: ***Box.test(…, type = "Ljung")*** z pakietu ***stats***

Hipoteza zerowa: brak autokorelacji

Hipoteza alternatywna: występuje autokorelacja

Metoda Cochrane’a-Orcutta

Jeżeli w modelu występuje zjawisko autokorelacji składnika losowego należy skorygować metodę estymacji parametrów lub zmienić postać analityczną modelu.

W przypadku wystąpienia autokorelacji I rzędu:

Krok 1: Szacujemy współczynnik autokorelacji $\hat{\rho}$

Krok 2: Dokonujemy przekształceń zmiennych zgodnie ze wzorami:

$$y_t^* = y_t - \hat{\rho} y_{t-1}$$
$$x_{jt}^* = x_{jt} - \hat{\rho} x_{jt-1}, \qquad j = 1, 2, 3, \ldots, k$$

Krok 3: Budujemy model na nowo utworzonych zmiennych.

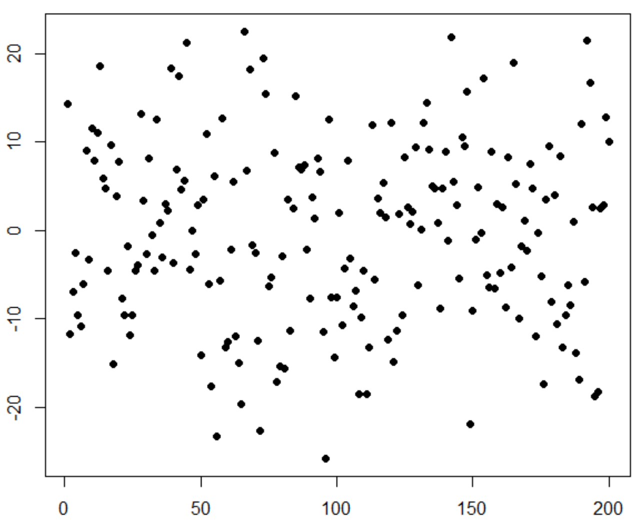
Homoskedastyczność składnika losowego

Homoskedastyczność

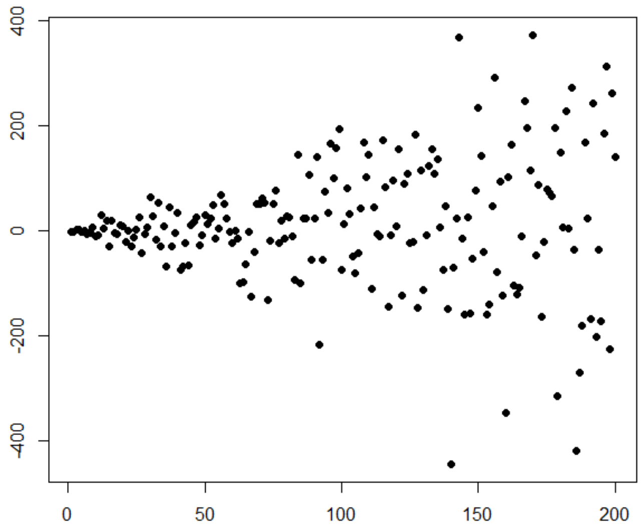
Jednym z założeń MNK jest homoskedastyczność składnika losowego (stałość wariancji błędu losowego).

Przeciwieństwem homoskedastyczności jest heteroskedastyczność.

Homoskedastyczność



Heteroskedastyczność



Homoskedastyczność

Model: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i, \text{dla i} = 1, \ldots, \text{n}$

Macierz wariancji-kowariancji składników losowych:

$$D^2(\varepsilon) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & & \sigma_{2n} \\ \vdots & & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{bmatrix}$$

gdzie $\sigma_i^2 = D^2(\varepsilon_i), \sigma_{ij} = cov(\varepsilon_i, \varepsilon_j) = cov(\varepsilon_j, \varepsilon_i) = \sigma_{ji}$

Założenie homoskedastyczności składników losowych:

$$\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_n^2 = \sigma^2$$

Homoskedastyczność

Konsekwencje niespełnienia założenia dotyczącego homoskedastyczności składnika losowego:

- Estymator MNK pozostaje nieobciążony i zgodny, ale nie jest najefektywniejszy.
- Oszacowania wariancji estymatorów są obciążone.

W związku z powyższym w przypadku występowania heteroskedastyczności testy istotności parametrów mogą zwracać błędne wyniki.

Heteroskedastyczność - testy

Testowanie występowania heteroskedastyczności:

- Test Goldfelda-Quandta – stosowany, gdy podejrzewa się, że wariancja składników losowych zależy od jednej ze zmiennych objaśniających X.
- Test Breuscha – Pagana (ogólniejszy niż test Goldfelda-Quandta) – służy do zbadania przypuszczenia, że wariancje składników losowych zależą od kilku zmiennych objaśniających.
- Test White'a (podobny do testu Breuscha – Pagana) – w teście tym jako czynniki wpływające na wariancję składnika losowego rozważane są zmienne objaśniające, ich potęgi oraz iloczyny.

Heteroskedastyczność – testy w R

- Test Goldfelda-Quandta

```
library(lmtest)
gqtest(model,
       alternative = "two.sided",
       order.by = ~ nazwa_zmiennej_X)
```

- Test Breuscha – Pagana

```
library(lmtest)
bptest(model)
```

- Test White'a (w przypadku dwóch zmiennych objaśniających: x1, x2)

```
library(lmtest)
bptest(model, ~ x1 + I(x1^2) + x2 + I(x2^2) + x1:x2)
```

Rozwiązania problemu

W przypadku, gdy składniki losowe są heteroskedastyczne można zastosować:

- Przekształcenie zmiennych (np. logarytmowanie),
- Ważoną metodę najmniejszych kwadratów (WMNK),
- Odporne błędy standardowe (robust standard errors, heteroskedasticity-consistent standard errors).

Ważona metoda najmniejszych kwadratów

Jeśli w modelu występuje tylko heteroskedastyczność składnika losowego (bez autokorelacji) to do estymacji parametrów modelu można wykorzystać ważoną metodę najmniejszych kwadratów. Polega ona na przypisaniu poszczególnym obserwacjom odpowiednich wag ω_i, których celem jest zrównoważenie wariancji czynnika losowego.

Początkowy model: $y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + u_i$

$$u_i = \omega_i \cdot \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2), \qquad u_i \sim N\left(0, \omega_i^2 \sigma^2\right)$$

Przekształcenie zmiennych: $y_i^* = \frac{y_i}{\omega_i}, x_{1i}^* = \frac{x_{1i}}{\omega_i}, \dots, x_{2i}^* = \frac{x_{2i}}{\omega_i}$

Budowa nowego modelu na zmiennych $y_i^*, x_{1i}^*, \dots, x_{2i}^*$

Ważona metoda najmniejszych kwadratów

Problem: w jaki sposób uzyskać wagi ω_i?

Możliwe rozwiązania:

- Wykorzystanie pomocniczej regresji opisującej logarytm z podniesionych do kwadratu reszt z modelu MNK. Kolejne kroki:
  1) Estymacja parametrów modelu $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ za pomocą MNK
  2) Wyznaczenie $\ln\left(\varepsilon_i^2\right)$, gdzie ε_i to reszty z modelu z kroku 1
  3) Estymacja parametrów modelu $\ln\left(\varepsilon_i^2\right) = \alpha_0 + \alpha_1 x_{1i} + \cdots + \alpha_k x_{ki} + \upsilon_i$
  4) Wyznaczenie $\widehat{\ln\left(\varepsilon_i^2\right)}$
  5) Obliczenie wag: $\omega_i = \sqrt{\exp(\widehat{\ln\left(\varepsilon_i^2\right)}}$
  6) Wykorzystanie ω_i do przekształcenia zmiennych

Ważona metoda najmniejszych kwadratów

Problem: w jaki sposób uzyskać wagi ω_i?

Możliwe rozwiązania:

- Jeśli wiadomo, że j-ta zmienna objaśniająca powoduje zmiany w wariancji składnika losowego wprost proporcjonalnie do swojej wartości, to można wykorzystać ją do wyznaczenia wag, np.:

$$\omega_i = \sqrt{x_{ji}}$$

Powyższe sposoby eliminacji problemu heteroskedastyczności nie zawsze są skuteczne. WMNK opiera się na założeniu, że znane są dokładne wielkości wag. W praktyce określenie wag może być jednak dość trudne.

Przykład w R

Symulacja następującego modelu:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$$

$$u_i = i \cdot \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2), \text{czyli } u_i \sim N(0, i^2 \cdot \sigma^2)$$

Utworzenie danych:

```r
N <- 200
beta_vec <- c(20, 2, -4)
x1 <- seq(from = 0, to = 5, length.out = N)
x2 <- sample(seq(from = 3, to = 17, length.out = 80), size = N, replace = TRUE)
e  <- rnorm(mean = 0, sd = 1:N, n = N)

x <- cbind(1, x1, x2)
y <- x %*% beta_vec + e
dane <- data.frame(y, x1, x2)
```

	y	x1	x2
1	-32.6209313	0.00000000	13.278481
2	-28.7772273	0.02512563	12.215190
3	7.7276732	0.05025126	3.886076
4	-28.3785107	0.07537688	13.278481
5	-2.6475236	0.10050251	7.075949
6	5.7750928	0.12562814	3.354430
7	2.9075944	0.15075377	4.949367
8	-25.4698920	0.17587940	11.329114
9	-9.2448638	0.20100503	8.139241
10	[illegible]	[illegible]	[illegible]

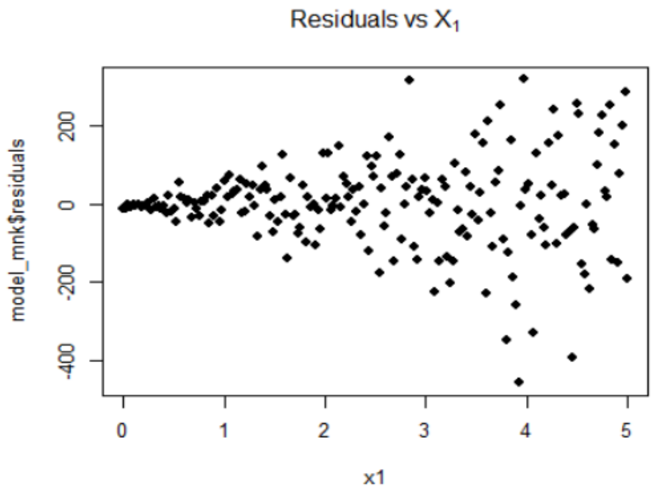
Przykład w R

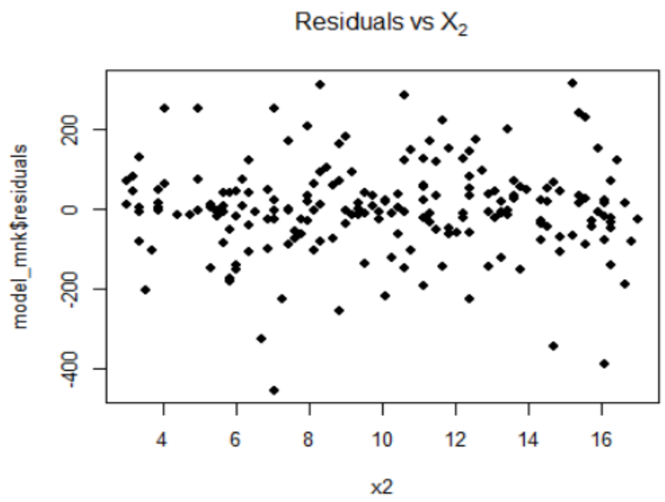
Estymacja parametrów modelu: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$

```
Call:
lm(formula = y ~ ., data = data_mat)

Residuals:
    Min      1Q  Median      3Q     Max 
-455.63  -47.93   -1.89   48.91  316.90 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   23.490     26.030   0.902    0.368
x1            -6.869      5.617  -1.223    0.223
x2            -3.289      2.106  -1.561    0.120

Residual standard error: 115 on 197 degrees of freedom
Multiple R-squared:  0.02078,	Adjusted R-squared:  0.01084 
F-statistic:  2.09 on 2 and 197 DF,  p-value: 0.1264
```

Przykład w R

Wykresy reszt z modelu





Przykład w R

Testy na heteroskedastyczność – test Goldfelda-Quandta:

➢ Dwa warianty: order.by = ~ x1 oraz order.by = ~ x2

```
	Goldfeld-Quandt test

data:  model_mnk
GQ = 8.9024, df1 = 97, df2 = 97, p-value < 2.2e-16
alternative hypothesis: variance changes from segment 1 to 2


	Goldfeld-Quandt test

data:  model_mnk
GQ = 1.06, df1 = 97, df2 = 97, p-value = 0.7747
alternative hypothesis: variance changes from segment 1 to 2
```

Przykład w R

Testy na heteroskedastyczność – test Breuscha – Pagana

```
	studentized Breusch-Pagan test

data:  model_mnk
BP = 35.627, df = 2, p-value = 1.835e-08
```

Testy na heteroskedastyczność – test White'a

```
	studentized Breusch-Pagan test

data:  model_mnk
BP = 37.205, df = 5, p-value = 5.448e-07
```

Wniosek: założenie dotyczące homoskedastyczności składnika losowego nie jest spełnione

Przykład w R

Wykorzystanie ważonej metody najmniejszych kwadratów

- Wersja I: wyznaczenie wag na podstawie dodatkowej regresji opisującej logarytm z podniesionych do kwadratu reszt z modelu MNK
- Wersja II: wykorzystanie kolejnych liczb całkowitych jako wagi, czyli $\omega_i = i$

Znając wagi nie trzeba ręcznie przekształcać każdej zmiennej z osobna. WMNK można przeprowadzić dodając argument *weights* w funkcji *lm():*

```
lm(y~., data = dane, weights = 1/wagi)
```

Przykład w R

Ważne: funkcja lm() mnoży zmienne w regresji razy pierwiastki kwadratowe z podanych wag!

W związku z tym jeśli chce się pomnożyć zmienne w modelu razy $\frac{1}{\sqrt{x_{ji}}}$ to w funkcji lm() należy przypisać do argumentu *weights* wartości $\frac{1}{x_{ji}}$ (wagi podniesione do kwadratu).

Przykład w R

Wersja I (budowa dodatkowego modelu):

1) Estymacja parametrów modelu $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ za pomocą MNK
2) Wyznaczenie $\ln(\varepsilon_i^2)$, gdzie ε_i to reszty z modelu z kroku 1
3) Estymacja parametrów modelu $\ln(\varepsilon_i^2) = \alpha_0 + \alpha_1 x_{1i} + \cdots + \alpha_k x_{ki} + \upsilon_i$
4) Wyznaczenie $\widehat{\ln(\varepsilon_i^2)}$
5) Obliczenie wag: $\omega_i = \sqrt{\exp(\widehat{\ln(\varepsilon_i^2)})}$
6) Wykorzystanie ω_i do przekształcenia zmiennych

```
log_e2 = log(model_mnk$residuals^2)
resid_model  <- lm(log_e2 ~ x1 + x2, data = data_mat)
w <- exp(resid_model$fitted.values)
model_wmnk <- lm(y ~ x1 + x2, data = data_mat, weights =  1 / w)
```

Przykład w R

Wersja I – wyniki:

```
Call:
lm(formula = y ~ x1 + x2, data = data_mat, weights = 1/w)

Weighted Residuals:
    Min      1Q  Median      3Q     Max 
-6.1017 -1.0841 -0.0411  1.1737  6.6646 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)   20.459     10.259   1.994  0.04751 * 
x1            -4.621      4.052  -1.140  0.25549   
x2            -3.233      1.001  -3.228  0.00146 **
---
Signif. codes:  
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.012 on 197 degrees of freedom
Multiple R-squared:  0.05682,	Adjusted R-squared:  0.04724 
F-statistic: 5.934 on 2 and 197 DF,  p-value: 0.003145
```

Przykład w R

Chcąc przedstawić wykres reszt z nowo powstałego modelu konieczne jest dokonanie pewnych przekształceń.

Model uwzględniający heteroskedastyczność: $\mathbf{Y}^* = \boldsymbol{X}^*\beta + \boldsymbol{\varepsilon}^*$

Reszty z nowego modelu: $\hat{\varepsilon}^*_{WMNK} = \mathbf{Y}^* - \boldsymbol{X}^*\hat{\beta}_{WMNK}$

W większości programów przy wykorzystywaniu gotowych funkcji do estymacji WMNK reszty obliczane są na podstawie wzoru:
$\hat{\varepsilon}_{WMNK} = \mathbf{Y} - \mathbf{X}\hat{\beta}_{WMNK}$

Przykład w R

Chcąc ocenić na podstawie wykresów reszt, czy w modelu nadal występuje heteroskedastyczność należy wykorzystać $\hat{\varepsilon}^*_{WMNK}$:

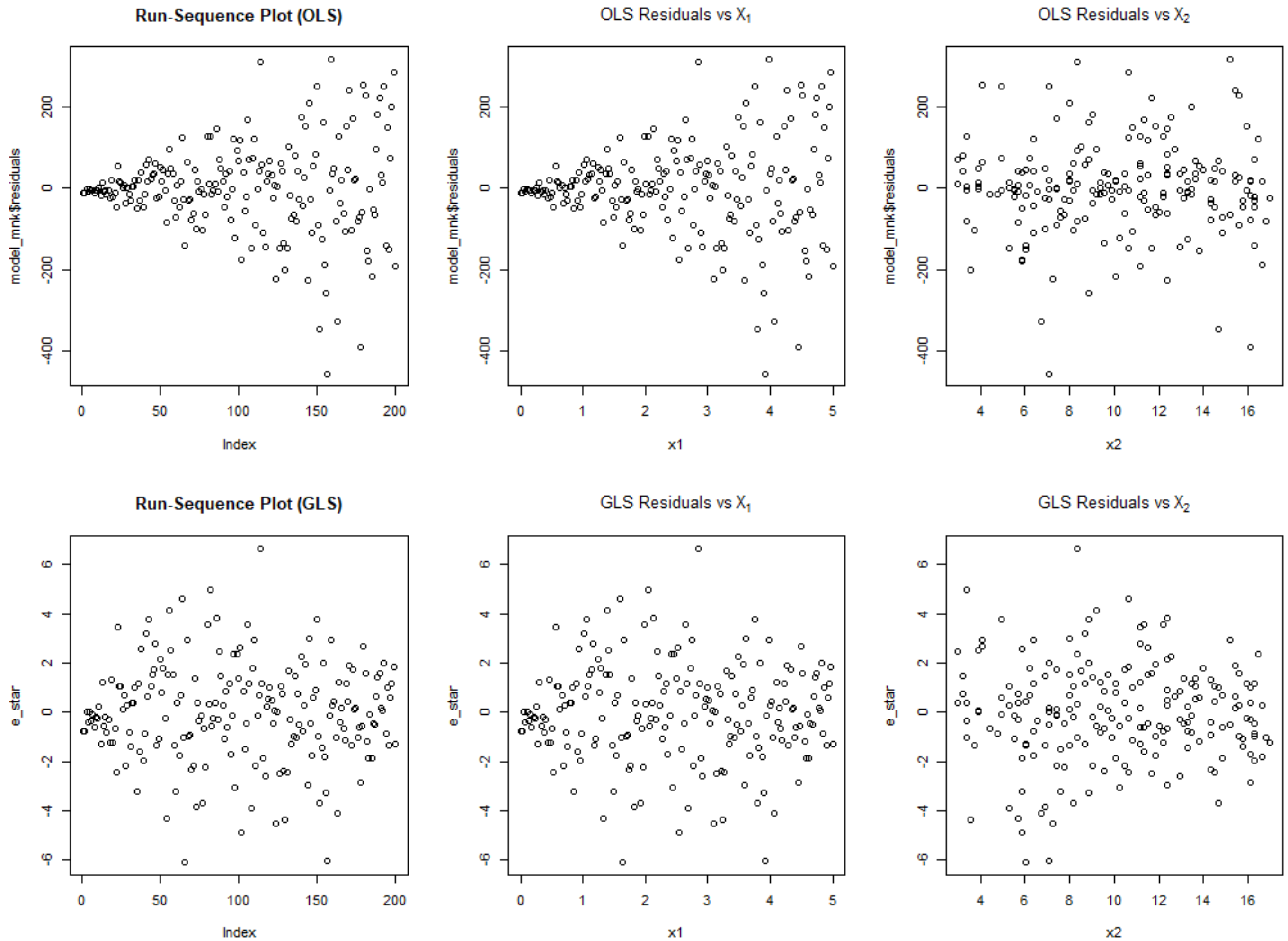
$$\hat{\varepsilon}^*_{WMNK} = \mathbf{Y}^* - \boldsymbol{X}^* \hat{\beta}_{WMNK} = \mathbf{Y}\frac{1}{\omega_i} - \mathbf{X}\frac{1}{\omega_i}\hat{\beta}_{WMNK} =$$

$$= \frac{1}{\omega_i}\left(\mathbf{Y} - \mathbf{X}\hat{\beta}_{WMNK}\right) = \frac{1}{\omega_i}\hat{\varepsilon}_{WMNK}$$

Zapis w kodzie:

```
e_star <- (1 / sqrt(w)) * model_wmnk$residuals
```

Przykład w R



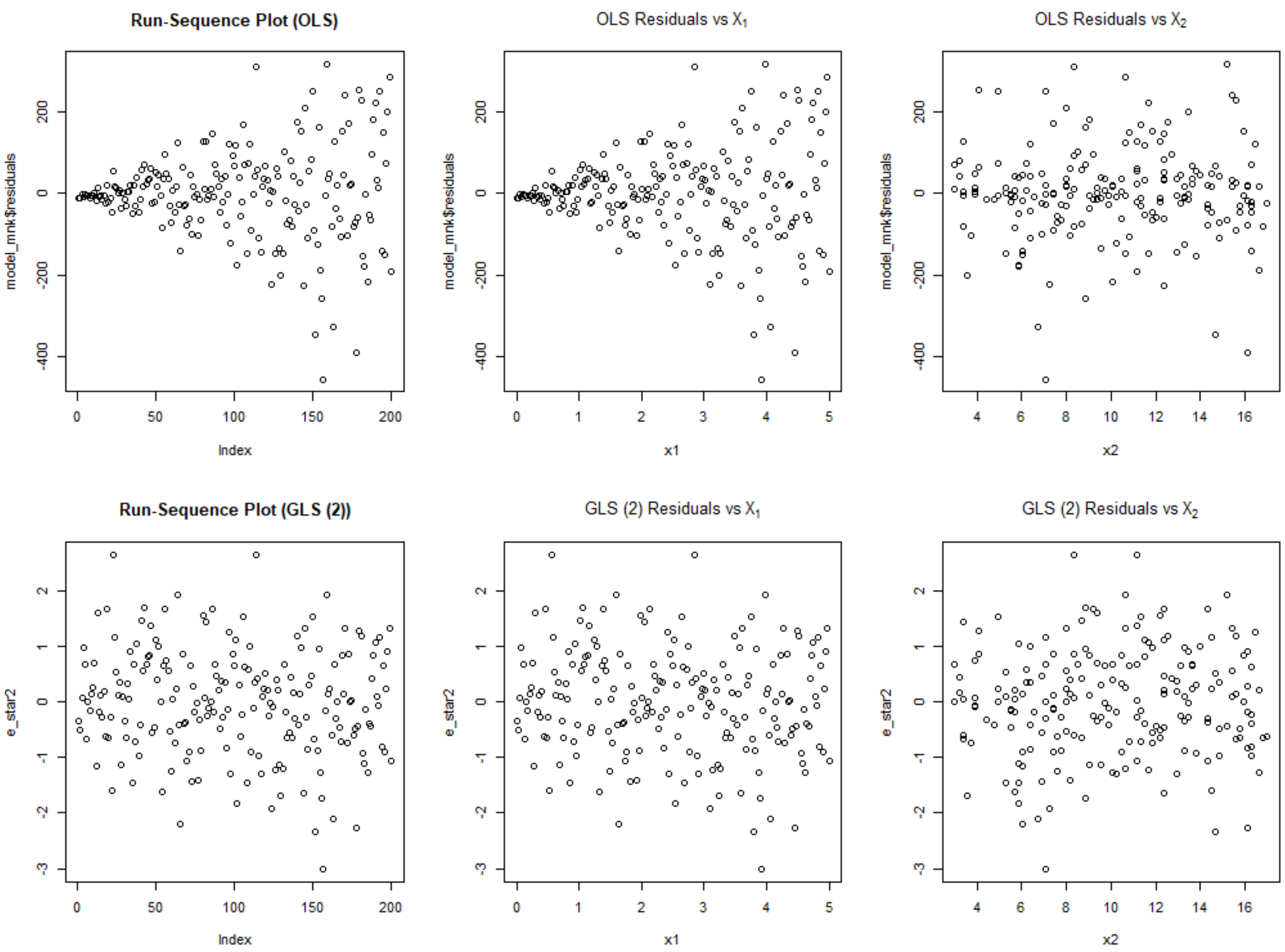
Run-Sequence Plot (OLS)
OLS Residuals vs X1
OLS Residuals vs X2
model_mnk$residuals
Index
x1
x2
Run-Sequence Plot (GLS)
GLS Residuals vs X1
GLS Residuals vs X2
e_star

Przykład w R

Wersja II (wykorzystanie wag: $\omega_i = i$) - wyniki:

```
Call:
lm(formula = y ~ x1 + x2, data = data_mat, weights = 1/(1:N)^2)

Weighted Residuals:
     Min       1Q   Median       3Q      Max 
-3.01335 -0.62702  0.01508  0.66931  2.66602 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  24.0172     3.1892   7.531 1.77e-12 ***
x1           -0.8305     2.9402  -0.282    0.778    
x2           -4.2399     0.2587 -16.387  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9714 on 197 degrees of freedom
Multiple R-squared:  0.5837,    Adjusted R-squared:  0.5795 
F-statistic: 138.1 on 2 and 197 DF,  p-value: < 2.2e-16
```

Przykład w R



Run-Sequence Plot (OLS)
OLS Residuals vs X_1
OLS Residuals vs X_2
model_mnk$residuals
Index
x1
x2
Run-Sequence Plot (GLS (2))
GLS (2) Residuals vs X_1
GLS (2) Residuals vs X_2
e_star2

Przykład w R

Chcąc przetestować, czy reszty z modelu WMNK są homoskedastyczne należy skorzystać z testu Breuscha – Pagana zaimplementowanego w pakiecie ***car*** (funkcja ***ncvTest()***).

Funkcja ***ncvTest()*** rozróżnia, czy w modelu uwzględniono wagi.

```
> ncvTest(model_mnk, var.formula = ~ x1 + x2)
Non-constant Variance Score Test
Variance formula: ~ x1 + x2
Chisquare = 73.83468, Df = 2, p = < 2.22e-16
> ncvTest(model_wmnk, var.formula = ~ x1 + x2)
Non-constant Variance Score Test
Variance formula: ~ x1 + x2
Chisquare = 10.9345, Df = 2, p = 0.0042228
> ncvTest(model_wmnk2, var.formula = ~ x1 + x2)
Non-constant Variance Score Test
Variance formula: ~ x1 + x2
Chisquare = 0.696991, Df = 2, p = 0.70575
```