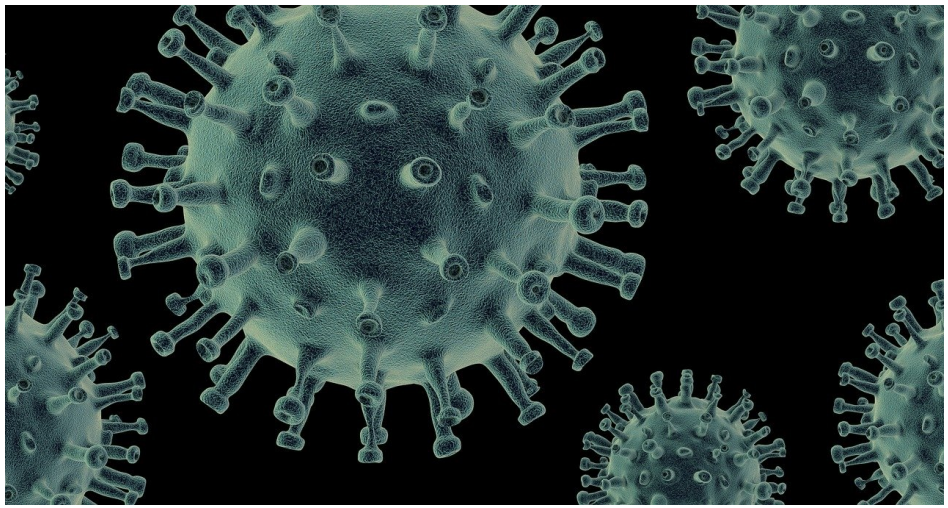
Weryfikacja założeń C.D.

26 kwietnia 2021



Rozkład normalny

Szczególny przypadek $\mathcal{N}(0,1)$ rozkładu normalnego nazywamy *standardowym rozkładem normalym*.

- Gęstość rozkładu normalnego

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \quad \text{for } x \in \mathbb{R};$$

- Rozkład jest symetryczny, tzn.
  - jeśli $X \sim \mathcal{N}(0,1)$ to $-X \sim \mathcal{N}(0,1)$
  - równoważnie gęstość jest funkcją *nieparzystą*, tzn. $\varphi(-x) = \varphi(x)$ dla $x \in \mathbb{R}$;
- *Dystrybuanta* rozkłada normalnego

$$\Phi(x) = \int_{-\infty}^{x} \varphi(s) ds$$

jest stablicowana dla $x \geq 0$, a za pomocą symetrii $\mathcal{N}(0,1)$, dla $x < 0$ wyznaczamy $\Phi(x)$ za pomocą wzoru

$$\Phi(x) = 1 - \Phi(-x).$$

Rozkład normalny

Własności $\mathcal{N}(\mu, \sigma^2)$:

- Jeśli $X \sim \mathcal{N}(0,1)$ wtedy $\mu + \sigma X \sim \mathcal{N}(\mu, \sigma^2)$;
- Odwrotnie, jeśli $X \sim \mathcal{N}(\mu, \sigma^2)$ wtedy $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$;
- Jeśli $X \sim \mathcal{N}(\mu, \sigma^2)$ to $EX = \mu$ oraz $Var(X) = \sigma^2$;
- Gęstość $\mathcal{N}(\mu, \sigma^2)$ wyraża się wzorem

$$\varphi_{\mu,\sigma^2}(x) = \frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}.$$

Definitions

Niech X będzie zmienną losową taką, że $EX = \mu$ oraz $Var(X) = \sigma^2$.

- $p$ty *moment* zmiennej X wyraża się jako EX^p;
- $p$ty *moment centralny* X wyraża się jako $m_p = E(X - \mu)^p$;
- pth *moment standaryzowany* X wyraża się jako $\kappa_p = E\left(\frac{X-\mu}{\sigma}\right)^p$;
- 3-ci moment standaryzowany nazywamy *skośnością*;
- 4-ty moment standaryzowany nazywamy *kurtozą*.

Skośność i kurtoza

Jeśli X jest zmienną losową taką, że $EX = \mu$ oraz $Var(X) = \sigma^2$ wtedy zachodzą warunki:

- Skośność:

$$\kappa_3 = E\left(\frac{X-\mu}{\sigma}\right)^3 = \frac{E\left(X-\mu\right)^3}{\sigma^3} = \frac{m_3}{\left(\sigma^2\right)^{\frac{3}{2}}} = \frac{m_3}{\left(m_2\right)^{\frac{3}{2}}}.$$

- Kurtoza:

$$\begin{aligned} \kappa_4 &= E\left(\frac{X-\mu}{\sigma}\right)^4 = \frac{E\left(X-\mu\right)^4}{\sigma^4} \\ &= \frac{m_4}{\left(\sigma^2\right)^2} = \frac{m_4}{\left(m_2\right)^2}. \end{aligned}$$

Skośność dla rozkładu normalnego

Jeśli $X \sim \mathcal{N}(\mu, \sigma^2)$, to $EX = \mu$, $Var(X) = \sigma^2$, oraz

- $U := \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$, zatem U ma rozkład symetryczny. Zatem,

$$\kappa_3 = E\left(\frac{X-\mu}{\sigma}\right)^3 = EU^3 = \int_{-\infty}^{\infty} \underbrace{s^3 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}s^2}}_{\text{funkcja nieparzysta}} ds = 0;$$

- Wnioskujemy, że **skośność każdego rozkładu normalego wynosi** 0.

Kurtoza dla rozkładu normalego

Jeśli $X \sim \mathcal{N}(\mu, \sigma^2)$, to $EX = \mu$ i $Var(X) = \sigma^2$. Wtedy,

- $U := \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$, zatem U ma rozkład symmetryczny. Stąd,

$$\begin{aligned}
\kappa_4 &= E\left(\frac{X-\mu}{\sigma}\right)^4 = EU^4 = \int_{-\infty}^{\infty} s^4 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}s^2} ds \\
&= \int_{-\infty}^{\infty} s^3 \left(-\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}s^2}\right)' ds \\
&= \underbrace{-\frac{s^3}{\sqrt{2\pi}} e^{-\frac{s^2}{2}}\Big|_{-\infty}^{\infty} + 3\int_{-\infty}^{\infty} s^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}s^2} ds}_{\text{całkujemy przez części.}} \\
&= 3\int_{-\infty}^{\infty} s^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}s^2} ds = 3Var(U) = 3.
\end{aligned}$$

- Mamy więc, $\kappa_4 = 3$ czyli **kurtoza każdego rozkładu normalnego wynosi** 3.

Skośność i kurtoza dla rozkładu normalego

Skośność i kurtoza

Własności *skośności* (κ_3) i *kurtozy* (κ_4):

- **Skośność** miara symmetrii rozkładu wokół średniej;
  - gdy $\kappa_3 = 0$ osią symetrii dla wykresu gęstości jest prosta pionowa $x = \mu$ (w układzie kartezjańskim (x, y));
  - **Skośność dla rozkładu normalego to** 0;
- **Kurtoza** to miara rozproszenia rozkładu normalnego:
  - Wysoka wartość κ_4 oznacza, że gęstość ma *ciężki ogon* w ∞ i $-\infty$ (wartości dalekie od 0 są prawdopodobne);
  - **Kurtoza dla każdego rozkładu normalnego jest jednakowa i wynosi** 3;

Zatem, wszystkie rozkłady normalne są symetryczne wokół średniej i jednakowo rozproszone.

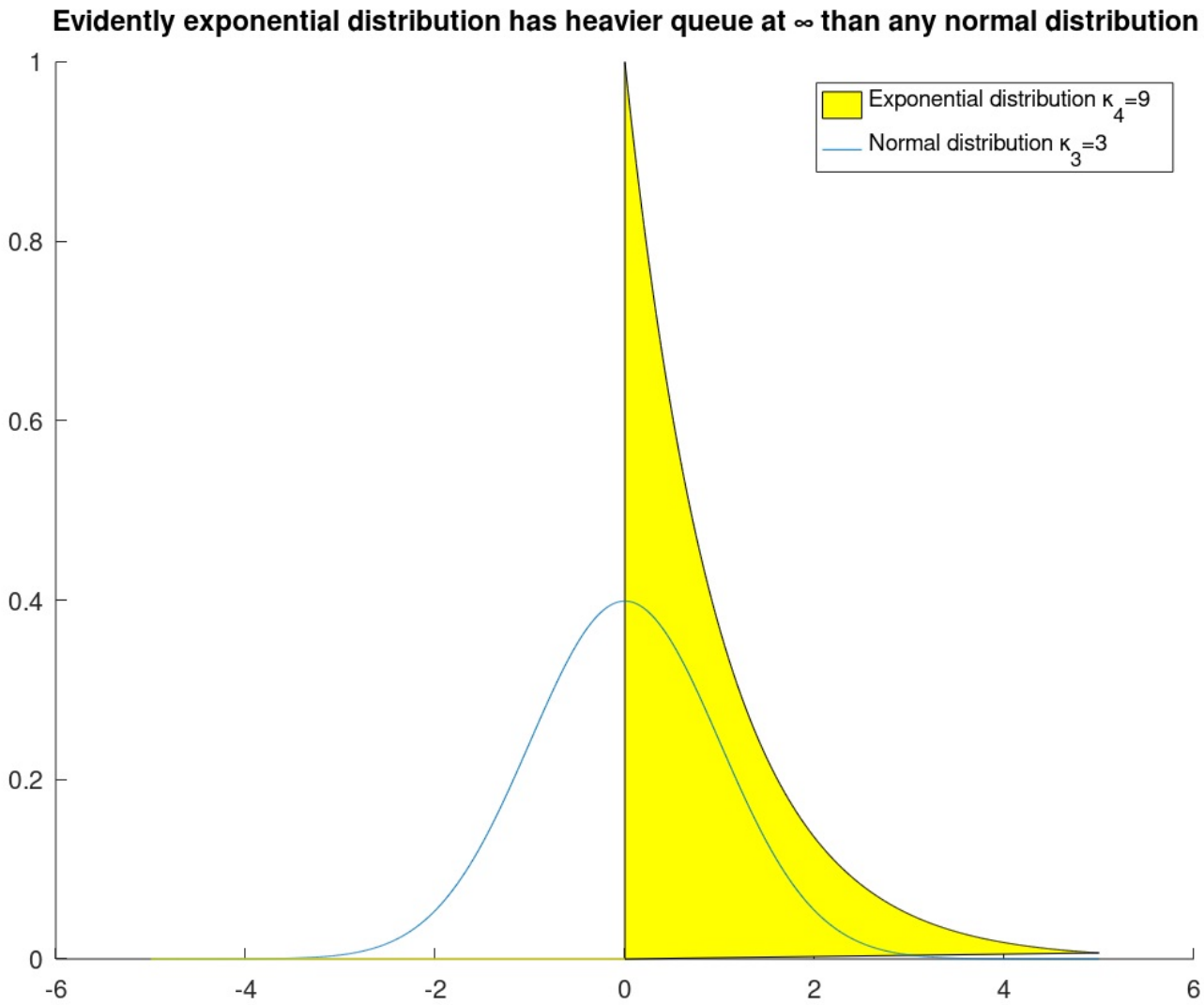
Kurtoza - ilustracja

- Przykładowo rozważmy *rozkład wykładniczy* $\mathcal{E}(\lambda)$ $(\lambda > 0)$ o gęstości

$$f_\lambda(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for} \quad x \geq 0 \\ 0 & \text{for} \quad x < 0 \end{cases}$$

- Skośność wynosi $\kappa_3 = 2$, a kurtoza $\kappa_4 = 9$;
- Gęstość rozkładu normalnego dąży do nieskończoności szybciej niż rozkład wykładniczy.

Kurtoza - ilustracja



Model regresji wielorakiej-powtórka

Nadal rozważamy model:

$$\underbrace{\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}}_{\mathbf{Y}} = \underbrace{\begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1k} \\ 1 & X_{21} & X_{22} & \dots & X_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{nk} \end{bmatrix}}_{\mathbf{X}} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}}_{\beta} + \underbrace{\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}}_{\epsilon},$$

równoważnie w formie macierzowej:

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon.$$

Założenia - powtórzenie

Przede wszystkim mamy więcej obserwacji niż parametrów, czyli $k + 1 \leq n$. Ponadto,

- Z1 Macierz $\mathbf{X}$ zmiennych objaśniających jest deterministyczna (nielosowa), tzn. macierz zmiennych objaśniających $[X_{t1}, X_{t2}, \ldots, X_{tk}]^T$ jest nielosowa dla t;
- Z2 Rząd macierzy $\mathbf{X}$ jest $k + 1$, zatem kolumny są liniowo niezależne;
- Z3 $E(\epsilon) = \mathbf{0}$, czyli $E\epsilon_t = 0$ dla t;
- Z4 $D^2(\epsilon) = \sigma^2\mathbf{I}$, tzn. ϵ_t jest ciągiem nieskorelowanych zmiennych losowych o wariancji σ^2, (σ jest nieznane);
- Z5 Wszystkie ϵ_t mają rozkład normalny $N(0, \sigma^2)$.

Weryfikacja założenia [Z5]

Zamierzamy przetestować, czy założenie [Z5] o normalności ϵ_t jest spełnione ponieważ:

- Estymator $\hat{\beta}$ otrzymany metodą najmniejszych kwadratów ma rozkład normalny, a ta własność jest potrzebna do konstrukcji wielu innych testów:
- Testy istotności
  - **pojedynczej zmiennej β_j:** statystyka testowa ma rozkład *t-Studenta*;
  - **grupy zmiennych z wektora β:** statystyka testowa ma rozkład *F-Snedecora*;
- Za pomocą własności rozkładu normalego konstrujemy test weryfikujący hipotezę normalności reszt.

Test normalności

Konstruujemy testy:

- **Test Jarque-Bera** - łatwy w interpretacji, ale słaba moc
  - często przyjmuje hipotezę gdy jest ona fałszywa;
- **Test Shapiro-Wilka** - trudny w interpretacji (ominiemy ten temat), ale wysoka moc
  - rzadziej niż inne znane testy przyjmuje hipotezę gdy ona jest fałszywa;
  - jest odporna na wystąpienie autokorelacji reszt.
- W tym celu obliczamy reszty modelu:

$$\hat{\epsilon} = \mathbf{Y} - \mathbf{X}\hat{\beta}, \quad \text{where } \hat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}.$$

Test Jarque Bera - konstrukcja i idea

- Testujemy hipotezę:

  $H_0 : \epsilon_t$ ma rozkład normalny VS $H_1 : H_0$ jest fałszywa.

- Ponieważ wiemy, że dla każdego rozkładu normalnego zachodzi $\kappa_3 = 0$, $\kappa_4 = 3$ formułujemy hipotezę pomocniczą:

  $H_0 : \kappa_3 = 0,$ oraz $\kappa_4 = 3$ VS $H_1 : \kappa_3 \neq 0,$ lub $\kappa_4 \neq 3;$

Test Jarque Bera - konstrukcja i idea

Z konstrukcji wynika, że

$$\sum_{t=1}^{n} \hat{\epsilon}_t = 0,$$

zatem mamy następujące estymatory:

- Estymator parametru σ^2, czyli wariancji ϵ_t to

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{t=1}^{n} \hat{\epsilon}_t^2;$$

- Estymator parametru $m_3 = E(\epsilon_t - E(\epsilon_t))^3 = E\epsilon_t^3$ to

$$\hat{m}_3 = \frac{1}{n} \sum_{t=1}^{n} \hat{\epsilon}_t^3;$$

- Estymator parametru $m_4 = E(\epsilon_t - E(\epsilon_t))^4 = E\epsilon_t^4$ to

$$\hat{m}_4 = \frac{1}{n} \sum_{t=1}^{n} \hat{\epsilon}_t^4;$$

Test Jarque Bera - konstrukcja i idea

- Z definicji skośności $\kappa_3 = \frac{m_3}{\sigma^3}$, definiujemy następujący estymator skośności

$$\hat{\kappa}_3 = \frac{\hat{m}_3}{\hat{\sigma}^3} = \frac{\frac{1}{n}\sum_{t=1}^{n} \epsilon_t^3}{\left(\frac{1}{n}\sum_{t=1}^{n} \hat{\epsilon}_t^2\right)^{3/2}};$$

- Z definicji kurtozy $\kappa_4 = \frac{m_4}{\sigma^4}$, definiujemy następujący estymator kurtozy

$$\hat{\kappa}_4 = \frac{\hat{m}_4}{\hat{\sigma}^4} = \frac{\frac{1}{n}\sum_{t=1}^{n} \epsilon_t^4}{\left(\frac{1}{n}\sum_{t=1}^{n} \hat{\epsilon}_t^2\right)^{2}};$$

Test Jarque Bera - konstrukcja i idea

- Statystyka testowa Jarque-Bera ma postać:

$$JB = n\left(\frac{1}{6}\hat{\kappa}_3^2 + \frac{1}{24}(\hat{\kappa}_4 - 3)^2\right);$$

- Jeśli H_0 jest prawdziwa i ϵ_t ma jakikolwiek rozkład normalny, to JB ma rozkład $\chi^2(2)$ (chi -kwadrat z 2 stopniami swobody);

Test Jarque Bera - konstrukcja i idea

Odrzucamy hipotezę H_0 na rzecz H_1 na poziomie istotności α (domyślnie $\alpha = 0.05$)

- JB przekroczy poziom krytyczny, kwantyl $\chi^2(2)$ na poziomie $1 - \alpha$:

$$JB > Q_{\chi^2(2)}(1 - \alpha);$$

- lub równoważnie p_{value} nie przekroczy poziomu istotności, tzn.

$$p_{value} = 1 - CDF_{\chi^2(2)}(JB) \leq \alpha,$$

gdzie $CDF_{\chi^2(2)}$ to *dystrybuanta* rozkładu $\chi^2(2)$.

Test Jarque Bera - zalety i wady

Zalety:

- **Łatwa implementacja:** łatwo obliczyć JB;
- **Łatwo dostrzegalna idea:** przesłanki do odrzucenia hipotezy są jasne:
  - rozkład jest zbyt skośny
  - rozkład jest zbyt mocno rozproszony, czyli gęstość ma zbyt ciężki ogon w ∞ lub $-\infty$;
- **dobrze znany rozkład statystyki** JB**:** dokładniej $\chi^2(2)$

Wady:

- **słaba moc testu:** często przyjmuje fałszywą hipotezę;

Test Shapiro-Wilka - konstrukcja

Testujemy hipotezę:

$$H_0 : \epsilon_t \text{ ma rozkład normalny} \quad \text{VS} \quad H_1 : H_0 \text{ jest fałszywa.}$$

Przy obliczonych resztach $\hat{\epsilon}$ konstruujemy statystykę testu Shapiro-Wilka:

- Porządkujemy wartości $\{\hat{\epsilon}_1, \hat{\epsilon}_2, \ldots, \hat{\epsilon}_n\}$ od najmniejszej do największej:
- Niech $\hat{\epsilon}_{t:n}$ będzie t- najmniejsza wartość $\hat{\epsilon}$ nazywamy $t-$ *statystyką porządkową*:
  - Przykładowo, jeśli $\hat{\epsilon} = [-3, 8, -1, 5]^T$, wtedy $\hat{\epsilon}_{1:4} = -3$, $\hat{\epsilon}_{2:4} = -1$, $\hat{\epsilon}_{3:4} = 5$, $\hat{\epsilon}_{4:4} = 8$;

Test Shapiro-Wilka - konstrukcja

Statystyka testowa:

- Niech $(U_1, U_2, \ldots, U_n)$ będzie próbą ze standardowego rozkładu normalnego $\mathcal{N}(0,1)$, i niech $U^O = (U_{1:n}, U_{2:n}, \ldots, U_{n:n})$ będzie wektorem statystyk porządkowych. Definiujemy $\mathbf{V}$ jako macierz kowariancji U^O tzn. $\mathbf{V} = [Cov(U_{i:n}, U_{j:n})]$, oraz $m_i = EU_{i,n}$;
- Definiujemy

$$[a_1, a_2, \ldots, a_n]^T = \frac{\mathbf{m}^T \mathbf{V}^{-1}}{\sqrt{\mathbf{m}^T \mathbf{V}^{-1} \mathbf{V}^{-1} \mathbf{m}}},$$

- Statystyka testowa wyraża się wzorem:

$$SW = \frac{\left(\sum_{t=1}^{n} a_t \hat{\epsilon}_{t:n}\right)^2}{\sum_{t=1}^{n} (\hat{\epsilon}_t - \bar{\hat{\epsilon}})^2} = \frac{\left(\sum_{t=1}^{n} a_t \hat{\epsilon}_{t:n}\right)^2}{\sum_{t=1}^{n} (\hat{\epsilon}_t)^2}.$$

- jeśli hipoteza H_0 jest prawdziwa, czyli ϵ_t ma jakikolwiek rozkład normalny, wtedy rozkład SW (niezależny od parametrów rozkładu) jest stablicowany.

szczegóły można znaleźć: Hanusz, Z. and Tarasińska, J. (2012), slajd 5;

Test Shapiro-Wilka - zalety i wady

Zalety:

- **Duża moc testu:** jest doświadczalnie udowodnione, że przy hipotezie H_1, ma wyższą moc niż wiele innych testów, przykładowo:
  - Test Andersona-Darling;
  - Test Kolmogorova-Smirnova;
  - Test Lillieforsa.
- **odporność na wystąpienie autokorelacji:**
  - test pozostaje użyteczny nawet wtedy gdy wystąpi autokorelacja reszt ϵ_t.

Wady:

- **Brak jasnego wzoru:** do obliczenia *SW* potrzeba specyficznych tablic, które są użyteczne tylko do tego problemu;
- **Niewidoczna idea:** idea *SW* nie jest widoczna na pierwszy rzut oka;
- **Stablcowany rozkład** *SW*: rozkład jest stablicowany, stąd implementacja wartości krytycznych i p-value wymaga

Wstęp do przykładu (udział OZE, a energochłonność)

Definicja (na marginesie)

Kilogram oleju ekwiwalentnego (kgoe) - stosowana w bilansach międzynarodowych jednostka miary energii. Oznacza ilość energii, jaka może zostać wyprodukowana ze spalenia jednego metrycznego kilograma ropy naftowej. Jedna tona oleju ekwiwalentnego równa jest 41.868 GJ (giga dżuli) lub 11.63 MWh (megawatogodzina).

Definicja (na marginesie)

Energochłonność pierwotna Stosunek zużycia energii pierwotnej do produktu krajowego brutto (PKB).

Przykład - energochłonność, a odnawialne źródła energii

Przykład

Procentowy udział OZE, a energochłonność

Za pomocą danych z Głównego Urzędu Statystycznego przebadamy związek między energochłonnością pierwotną PKB wyrażoną w kgoe na 1000 euro, a udziałem OZE w miksie energetycznym. Dane przedstawia tabelka.

- Znajdź współczynnik korelacji Pearsona;
- Oszacuj model liniowy;
- Za pomocą testu Jarque-Bera i Shapiro-Wilka zweryfikuj hipotezę o normalności reszt;
- Czy można zastosować testy istotności Studenta?

Przykład CD

Rok	Udział energii elektrycznej z OZE w końcowym rozliczeniu brutto (w %)	Energochłonność pierwotna PKB (w kgoe/1000 euro)
2000	1,7	485
2001	2	478
2002	2	462
2003	1,6	458
2004	2,207634071	434
2005	2,675865088	423
2006	3,009405955	420
2007	3,5	393
2008	4,372325204	381
2009	5,8	355
2010	6,848341452	366
2011	8,162195252	352
2012	10,67934145	334
2013	10,7	328
2014	12,40374325	305
2015	13,43270176	298
2016	13,36	302
2017	13,1	302
2018	13,03	292

Przykład CD

Poszukujemy model postaci:

$$Y_t = \beta_0 + \beta_1 X_t + \epsilon_t,$$

gdzie

- X_t -energochłonność w roku $t = 1, 2, ..., 19$ (począwszy od roku 2000 a skończywszy na 2018) w kgoe/1000 euro;
- Y_t - udział OZE w % rozliczeniu energii elektrycznej;

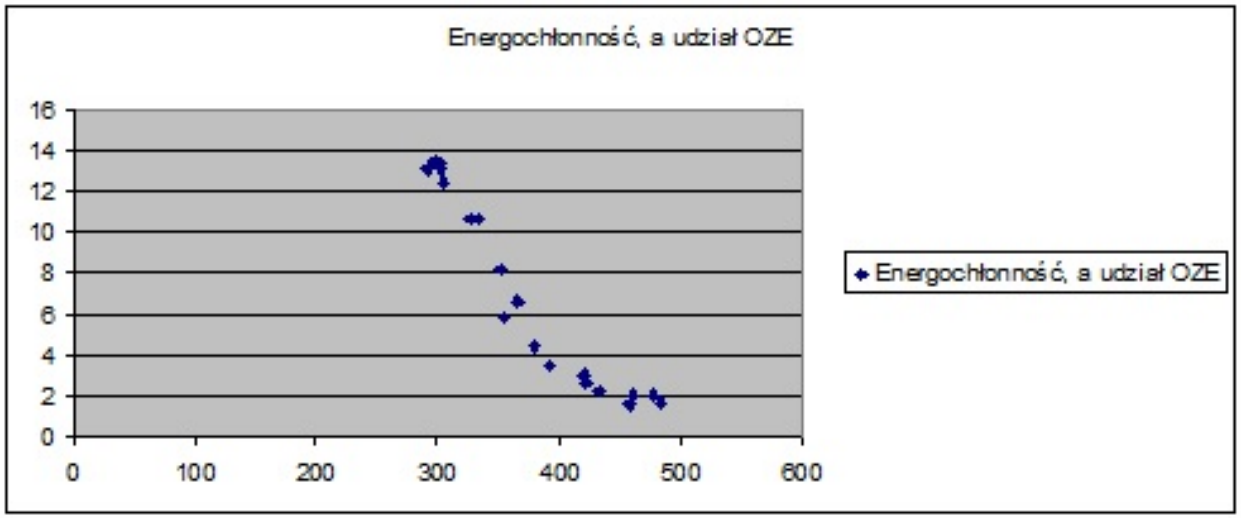
Przykład CD

Jako przymiarkę można policzyć współczynnik korelacji Pearsona o wzorze ogólnym:

$$Corr(X, Y) = \frac{\sum_{t=1}^{n}(X_t - \bar{X})(Y_t - \bar{Y})}{\sqrt{\sum_{t=1}^{n}(X_t - \bar{X})^2 \sum_{t=1}^{n}(Y_t - \bar{Y})^2}}.$$

W naszym przypadku wynosi on $-0,953359349$ co wskazuje na silną ujemną, korelację między energochłonnością, a udziałem OZE.

Przykład CD


Energochłonność, a udział OZE
16
14
12
10
8
6
4
2
0
0
100
200
300
400
500
600
Energochłonność, a udział OZE

Przykład CD

Metodą najmniejszych kwadratów znajdujemy model postaci:

$$Y_t \approx 32.4816 - 0.0679086 X_t + \epsilon_t,$$

Obliczamy reszty ze wzorów:

$$\hat{\epsilon}_t = Y_t - 32.4816 - 0.0679086 X_t.$$

Przykład CD

W celu zweryfikowania hipotezy o normalności reszt na poziomie 0.05, wykorzystam testy Jarque-Bera i Shapiro-Wilka:

Test	p-value	decyzja
Jarque-Bera	0.584651	przyjmujemy hipotezę
Shapiro-Wilka	0.328546	przyjmujemy hipotezę

Oba testy przyjmują hipotezę o normalności reszt. Można więc zastosować standardowy test Studenta do weryfikacji istotności:

	$\hat{\beta}_i$	$\sqrt{d_{ii}}$	T_i	p-value	decyzja
β_0	32.4816	1.99561	16.28	$\frac{8{,}42}{10^{12}} < 0.05$	istotny
β_1	-0.0679086	0.00521456	-13.02	$\frac{2{,}85}{10^{10}} < 0.05$	istotny

Przykład

Zatem przyjmujemy model

$$\text{Udział OZE} \approx 32.4816 - 0.0679086 * \text{Energochłonność}$$