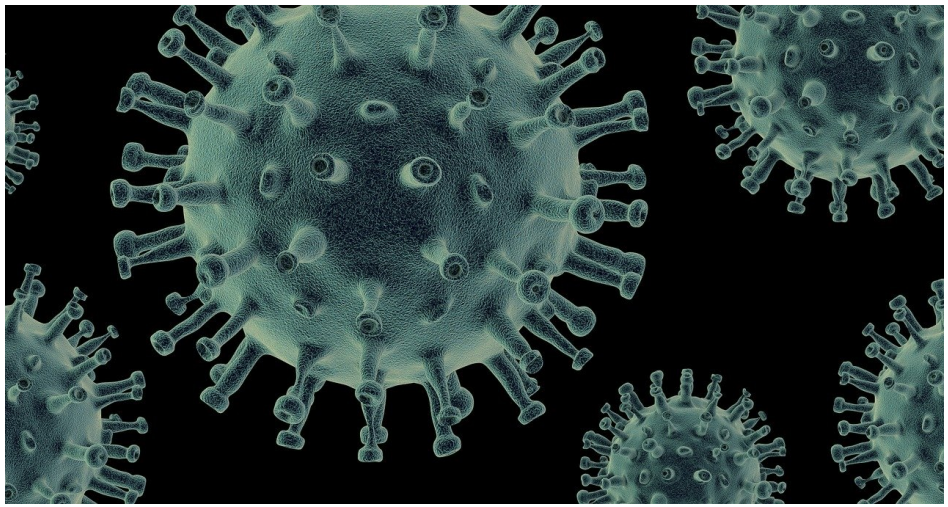
Oczekiwane błędy prognoz-ocena *ex-ante* i przedziały ufności

7 czerwca 2021



Model liniowy

Kontynuujemy pracę nad modelem

$$Y_t = \beta_0 + \beta_1 X_{t,1} + \ldots + \beta_k X_{t,k} + \epsilon_t, \quad t = 1, 2, ..., n.$$

Model można zapisać w formie macierzowej

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

lub równoważnie:

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

Założenia-powtórezenie

Mamy więcej obserwacji niż parametrów tzn. $k+1 \leq n$. Ponadto,

- Z1 Macierz zmiennych objaśniających $\mathbf{X}$ jest deterministyczna (nielosowa), tzn. wartości $[X_{t1}, X_{t2}, \ldots, X_{tk}]^T$ są z góry zadane i nie są wynikiem żadnego losowania dla każdego t;
- Z2 Rząd macierzy $\mathbf{X}$ jast równy $k+1$, innymi słowy kolumny są liniowo niezależne;
- Z3 $E(\epsilon) = \mathbf{0}$, tzn. $E\epsilon_t = 0$ dla każdego t;
- Z4 $D^2(\epsilon) = \sigma^2\mathbf{I}$, tzn.
  - ϵ_t jest ciągiem nieskorelowanych zmiennych losowch ;
  - każda zmienna ϵ_t ma tą samą wariancję σ^2, (σ jest nieznana);
- Z5 Wszystkie ϵ_t mają rozkład normalny $N(0, \sigma^2)$.

Estymacja-powtórzenie

Na podstawie powyższych danych

- konstruujemy estymator nieznanego parametru β za pomocą metody najmniejszych kwadratów:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \approx \beta;$$

- Konstruujemy oszacowania (prognozy) obserwacji:

$$\hat{Y}_t = \hat{\beta}_0 + \hat{\beta}_1 X_{t,1} + ... + \hat{\beta}_k X_{t,k} \approx Y_t.$$

Estymator $\hat{\beta}$ - własności

Oszacowanie wartości oczekiwanej i macierzy kowariancji estymatora $\hat{\beta}$:

- Nieobciążoność (Twierdzenie Gaussa-Markova):

$$E(\hat{\beta}) = \begin{bmatrix} E\hat{\beta}_1 \\ E\hat{\beta}_2 \\ \vdots \\ E\hat{\beta}_k \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} = \beta;$$

- Macierz kowariancji $D^2(\hat{\beta}) := [Cov(\hat{\beta}_i, \hat{\beta}_j)]_{i,j=1,\ldots,k}$:

$$D^2(\hat{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1};$$

Estymator S^2 parametru σ^2

Estymatorem parametru σ^2 jest

$$S^2 = \frac{1}{n-k-1}\sum_{t=1}^{n}(Y_t - \hat{\beta}_0 - \hat{\beta}_1 X_{t,1} - \ldots - \hat{\beta}_{t,k})^2;$$

Równoważny wzór macierzowy:

$$S^2 = \frac{1}{n-k-1}\overbrace{\underbrace{(Y - X\hat{\beta})^T}_{\text{macierz } 1\times n} * \underbrace{(Y - X\hat{\beta})}_{\text{macierz } n\times 1}}^{\text{macierz } 1\times 1 \text{ czyli liczba.}}$$

Estymator S^2 jest nieobciążony tzn:

$$ES^2 = \sigma^2;$$

Prognozowanie

Zatem na podstawie obserwacji z okresów $k = 1, 2, ..., n$ znajdujemy estymator $\hat{\beta}$.

- Dysponujemy już oszacowaniami zmiennych objaśniających w przyszłych okresach $n+1, n+2, ..., n+T$

$$\hat{X}_{n+1} \approx X_{n+1}, \quad \hat{X}_{n+2} \approx X_{n+2}, \ldots, \quad \hat{X}_{n+T} \approx X_{t+T};$$

- Zbudujemy prognozy przyszłych wartości $\hat{Y}_{n+1}, \hat{Y}_{n+2}, ..., \hat{Y}_{n+T}$ wg wzoru:

$$\hat{Y}_{n+\tau} = \hat{\beta}_0 + \hat{\beta}_1 \hat{X}_{t+\tau,1} + ... + \hat{\beta}_k \hat{X}_{t+\tau,k}, \quad \tau = 1, 2, ..., T.$$

Źródła błędów prognoz

Z prognozowaniem wiążą się błędy wynikające z następujących przyczyn:

- **Błąd estymacji:** prawdziwa wartość parametru jest inna niż estymator $\hat{\beta} \neq \beta$;
- **Błąd losowy:** prognozując zaniedbujemy wartość składnika losowego;
- **Błąd struktury modelu:** założenia [Z1-Z5] nie zawsze są spełnione;
- **Błąd specyfikacji:** Niewłaściwy dobór zmiennych objaśniających, lub niewłaściwy wzór analityczny;
- **Błąd warunków endogenicznych:** nieprawidłowe wartości zmiennych objaśniających w okresie obserwacji, lub zmiana warunków kształtowania zmiennej objaśnianej:
- **Błąd warunków egzogenicznych:** nieprawdidłowe określenie zmiennych objaśniających w okresie prognozy;
- **Błędy pomiaru:** Zaobserwowane wartości zmiennych objaśniającej i objaśnianej również są obarczone błędami.

Niech

$$\hat{\Upsilon}_\tau := \begin{bmatrix} 1 \\ \hat{X}_{n+\tau,1} \\ \vdots \\ \hat{X}_{n+\tau,k} \end{bmatrix} \quad \text{oraz} \quad \Upsilon_\tau := \begin{bmatrix} 1 \\ X_{n+\tau,1} \\ \vdots \\ X_{n+\tau,k} \end{bmatrix}$$

Prognozę można wyrazić w formie macierzowej:

$$\hat{Y}_{n+\tau} = \hat{\beta}_0 + \hat{\beta}_1 \hat{X}_{n+\tau,1} + ... + \hat{\beta}_k \hat{X}_{n+\tau,k} = \hat{\Upsilon}_\tau^T * \hat{\beta}$$

Dokładna wartość w przyszłości wyraża się wzorem

$$Y_{n+\tau} = \beta_0 + \beta_1 X_{n+\tau,1} + ... + \beta_k X_{n+\tau,k} + \epsilon_t = \Upsilon_\tau^T * \beta + \epsilon_{t+\tau};$$

Oczekiwane błędy prognoz

Ocena stopnia dokładności *ex-ante* ma za zadanie

- oszacować oczekiwany błąd prognozy zmiennej objaśnianej zanim poznamy jej wartości w przyszłości (w odróżnieniu z oceną *ex-post*);
- wyjaśnia **wkład błędu estymacji** i **błędu losowego** w kształtowaniu błędu prognozy;

Prognozowanie - oczekiwane błędy prognoz

Błąd prognozy wyrażamy wzorem:

$$e_\tau := Y_{n+\tau} - \hat{Y}_{n+\tau}, \quad \text{dla} \quad \tau = 1, 2, ..., T.$$

Na początek przyjmujemy, że jedyne źródła błędów to:

- Błąd estymacji; $\beta \neq \hat{\beta}$;
- Błąd losowy; $\epsilon_{n+\tau} \neq 0$;
- Błąd warunków egzogenicznych: $\Upsilon_\tau \neq \hat{\Upsilon}_\tau$

Ze wzorów

$$\hat{Y}_\tau = \hat{\Upsilon}_\tau^T \hat{\beta} \quad \text{oraz} \quad Y_\tau = \Upsilon_\tau^T \beta + \epsilon_{n+\tau}.$$

otrzymujemy

$$\begin{aligned} e_\tau &= Y_\tau - \hat{Y}_\tau \\ &= \Upsilon_\tau^T \beta + \epsilon_{t+\tau} - \hat{\Upsilon}_\tau^T \hat{\beta} \\ &= \Upsilon_\tau^T (\beta - \hat{\beta}) + (\Upsilon_\tau^T - \hat{\Upsilon}_\tau^T)\hat{\beta} + \epsilon_{n+\tau}. \end{aligned}$$

W dalszej części ignorujemy błąd warunków egzogenicznych i przyjmujemy:

$$\Upsilon_\tau = \hat{\Upsilon}_\tau.$$

Mamy więc:

$$\begin{aligned} e_\tau &= \Upsilon_\tau^T * (\beta - \hat{\beta}) + (\Upsilon_\tau^T - \hat{\Upsilon}_\tau^T) * \hat{\beta} + \epsilon_{n+\tau} \\ &= \Upsilon_\tau^T * (\beta - \hat{\beta}) + \epsilon_{n+\tau} \end{aligned}$$

Obliczamy wartość oczekiwaną błędu prognoz:

$$
\begin{aligned}
E(e_\tau) &= E(\Upsilon_\tau^T * (\beta - \hat{\beta}) + \epsilon_{n+\tau}) \\
&= E(\Upsilon_\tau^T * (\beta - \hat{\beta})) + \underbrace{E(\epsilon_{n+\tau})}_{=0 \text{ z założenia } [Z5].} \\
&= \underbrace{E(\Upsilon_\tau^T * (\beta - \hat{\beta}))}_{\Upsilon_\tau \text{ jest nielosowe.}} \\
&= \Upsilon_\tau^T * \underbrace{E(\beta - \hat{\beta})}_{=\mathbf{0} \text{ (wektor) z Tw. Gaussa-Markova.}} \\
&= \Upsilon_\tau^T * \mathbf{0} = 0;
\end{aligned}
$$

Prognozowanie - oczekiwane błędy prognoz

Otrztmujemy zatem

$$E(e_\tau) = 0,$$

zatem

$$Var(e_\tau) = E(e_\tau - E(e_\tau))^2 = E(e_\tau^2).$$

Dalej mamy

$$
\begin{aligned}
E(e_\tau^2) &= (\Upsilon_\tau^T(\beta-\hat{\beta})+\epsilon_{n+\tau})^2 \\
&= E\left((\Upsilon_\tau^T*(\beta-\hat{\beta}))^2+2\Upsilon_\tau^T(\beta-\hat{\beta})+\epsilon_{n+\tau}^2\right). \\
&= E(\Upsilon_\tau^T*(\beta-\hat{\beta}))^2+\underbrace{2E(\Upsilon_\tau^T*(\beta-\hat{\beta}))}_{\Upsilon_\tau \text{ jest nielosowy}}+\underbrace{E\epsilon_{n+\tau}^2}_{=\sigma^2 \text{ z } [Z4].} . \\
&= E(\Upsilon_\tau^T*(\beta-\hat{\beta}))^2+2\Upsilon_\tau^T*\underbrace{E(\beta-\hat{\beta})}_{=0 \text{ z Tw. Gaussa-Markova.}}+\sigma^2 \\
&= E(\Upsilon_\tau^T*(\beta-\hat{\beta}))^2+\sigma^2.
\end{aligned}
$$

Pozostaje obliczyć $E(\Upsilon_\tau^T*(\beta-\hat{\beta}))^2$.

Zauważmy:

$$\begin{aligned}(\Upsilon_\tau^T * (\beta - \hat{\beta}))^2 &= \underbrace{\Upsilon_\tau^T * (\beta - \hat{\beta})}_{\text{macierz } 1\times 1} * (\Upsilon_\tau^T * (\beta - \hat{\beta}))^T \\ &= \Upsilon_\tau^T * (\beta - \hat{\beta}) * (\beta - \hat{\beta})^T * \Upsilon_\tau\end{aligned}$$

Mamy więc:

$$\begin{aligned}E((\Upsilon_\tau^T * (\beta - \hat{\beta}))^2) &= E(\Upsilon_\tau^T * (\beta - \hat{\beta}) * (\beta - \hat{\beta})^T * \Upsilon_\tau) \\ &= \Upsilon_\tau^T * E((\beta - \hat{\beta}) * (\beta - \hat{\beta})^T) * \Upsilon_\tau \\ &= \Upsilon_\tau^T * \underbrace{E((\hat{\beta} - E(\hat{\beta})) * (\hat{\beta} - E(\hat{\beta}))^T)}_{=D(\hat{\beta}) \text{ z Tw Gaussa-Markova}} * \Upsilon_\tau \\ &= \Upsilon_\tau^T * D^2(\hat{\beta}) * \Upsilon_\tau.\end{aligned}$$

Ze wcześniejszych wzorów

$$D^2(\hat{\beta}) = E(\hat{\beta} - \beta) * (\hat{\beta} - \beta)^T) = \sigma^2(\mathbf{X}^T * \mathbf{X})^{-1},$$

oraz ze wcześniejszego slajdu otrzymujemy

$$E((\Upsilon_\tau^T * (\beta - \hat{\beta}))^2) = \sigma^2 \Upsilon_\tau^T * (\mathbf{X}^T * \mathbf{X})^{-1} * \Upsilon_\tau.$$

Otrzymujemy:

$$E((\Upsilon_\tau^T * (\beta - \hat{\beta}))^2) = \sigma^2 \Upsilon_\tau^T * (\mathbf{X}^T * \mathbf{X})^{-1} * \Upsilon_\tau.$$

Stąd i ze wcześniejszych slajdów

$$\begin{aligned} Ee_\tau^2 &= E((\Upsilon_\tau^T * (\beta - \hat{\beta}))^2) + \sigma^2 \\ &= \sigma^2 \Upsilon_\tau^T * (\mathbf{X}^T * \mathbf{X})^{-1} * \Upsilon_\tau + \sigma^2 \\ &= \sigma^2 \left(1 + \Upsilon_\tau^T * (\mathbf{X}^T * \mathbf{X})^{-1} * \Upsilon_\tau\right) \end{aligned}$$

Otrzymujemy

- **Wartośś oczekiwana** błędu prognoz :

$$E(e_\tau^2) = 0;$$

- **Wariancja** błędu prognoz:

$$Var(e_\tau) = \sigma^2 \left(1 + \Upsilon_\tau^T * (\mathbf{X}^T * \mathbf{X})^{-1} * \Upsilon_\tau\right);$$

- **Odchylenie standardowe** błędu prognoz:

$$s_\tau = \sqrt{Var(e_\tau)} = \sigma \sqrt{1 + \Upsilon_\tau^T * (\mathbf{X}^T * \mathbf{X})^{-1} * \Upsilon_\tau}.$$

Zauważmy, że

- Odchylenie standardowe prognozy s_τ zależy od nieznanego parametru σ:

$$s_\tau = \sqrt{Var(e_\tau)} = \sigma\sqrt{1 + \Upsilon_\tau^T * (\mathbf{X}^T * \mathbf{X})^{-1} * \Upsilon_\tau};$$

- Estymator parametru σ^2 jest S^2 wyrażony wzorem

$$S^2 = \frac{1}{n-k-1}\sum_{t=1}^{n}\left(Y_t - \hat{\beta}_0 - \hat{\beta}_1 X_{t,1} - ... - \hat{\beta}_k X_{t,k}\right)^2 \approx \sigma^2;$$

- Zatem naturalnym estymatorem s_τ jest $\hat{s}_\tau$ wyrażony wzorem:

$$\hat{s}_\tau = S\sqrt{1 + \Upsilon_\tau^T * (\mathbf{X}^T * \mathbf{X})^{-1} * \Upsilon_\tau}.$$

Średni względny błąd predykcji *ex-ante*

Średni względny błąd predykcji *ex-ante* w okresie τ definiujemy w następujęcy sposób:

$$v_\tau = \frac{\hat{s}_\tau}{\hat{Y}_{n+\tau}} * 100\%.$$

Średni względny błąd predykcji *ex-ante*-własności

Średni względny błąd predykcji:

- Wyraża się wzorem:

$$v_\tau = \frac{\hat{s}_\tau}{\hat{Y}_{n+\tau}} = \frac{S}{\hat{Y}_{n+\tau}} \sqrt{1 + \Upsilon_\tau^T * (\mathbf{X}^T * \mathbf{X})^{-1} * \Upsilon_\tau};$$

- Określa w procentach jakich błędów można się spodziewać w okresie prognozy τ w stosunku do wielkości samej prognozy;

Prognoza przedziałowa

Twierdzenie

Przy założeniach [Z1-Z5], dla każdego $\tau = 1, 2, ..., n$, zmienna losowa:

$$U_\tau := \frac{Y_{n+\tau} - \hat{Y}_{n+\tau}}{\hat{s}_\tau}$$

ma rozkład t-Studenta z $n - k - 1$ stopniami swobody ($\mathcal{T}_{n-k-1}$).

Prognoza przedziałowa

Ponieważ U_τ ma rozkład $\mathcal{T}_{n-k-1}$,

- możemy określić kwantyl t_α tego rozkładu rzędu $1-\alpha/2$ spełniający
$$P(|U_\tau| < t_\alpha) = 1-\alpha;$$
- Innymi słowy:
$$P(\hat{Y}_{n+\tau} - t_\alpha \hat{s}_\tau \leq Y_{n+\tau} \leq \hat{Y}_{n+\tau} + t_\alpha \hat{s}_\tau) = 1-\alpha$$
- Wtedy określany $(1-\alpha) * 100\%$ przedział ufności dla błędu jako:
$$[\hat{Y}_{n+\tau} - t_\alpha \hat{s}_\tau, \hat{Y}_{n+\tau} + t_\alpha \hat{s}_\tau];$$
- Istnieje $(1-\alpha)100\%$ szans, że przyszła wartość $Y_{n+\tau}$ znajdzie się w w.w. przedziale (VERTE).

Prognoza przedziałowa

Przedział ufności dla przyszłej wartości

Dla przyszłej wartości $Y_{n+\tau}$ określamy $100(1-\alpha)\%$ przedział ufności jako

$$[\hat{Y}_{n+\tau} - t_\alpha \hat{s}_\tau, \hat{Y}_{n+\tau} + t_\alpha \hat{s}_\tau].$$

Uwaga

Jeśli zbyt dużo przeszłych wartości znajdzie się poza przedziałem ufności, możemy przypuszczać, że model nie został prawidłowo określony, lub zmieniły się warunki kształtowania zmiennych losowych.

Przykład - Temperatura w Tokio 1875-2010








Przykład - Temperatura w Tokio 1875-2010

- Określiliśmy model postaci:

$$Y_t = 13.7287 + 0.000178514 * t^2 + \epsilon_t,$$

  dla $t = 1, 2, ..., 125$ (indeksując lata;
- Model został pozytywnie zweryfikowany, ale prognozy znacznie odbiegły od przyszłych wartości.
- Pozostaje pytanie, czy te prognozy mieszczą się w granicach błędu statystycznego: konstruujemy warości *ex-ante*, a później przedziały ufności:

Przykład - Temperatura w Tokio 1875-2010

Najpierw obliczamy

$$\hat{s}_\tau = S\sqrt{1 + \Upsilon_\tau^T * (\mathbf{X}^T * \mathbf{X})^{-1} * \Upsilon_\tau}$$

dla $\tau = 1, 2, ..., 20$ (odpowiednie dla lat 2001-2020). Mamy

$$\begin{cases} \Upsilon_1^T = [1,\ 126^2] \\ \Upsilon_2^T = [1,\ 127^2] \\ \vdots \\ \Upsilon_\tau^T = [1,\ (125+\tau)^2] \\ \vdots \\ \Upsilon_1^T = [1,\ 146^2] \end{cases}$$

Przykład - Temperatura w Tokio 1875-2010

Na podstawie obliczeń

$$S^2 = 0.199736653 \quad \text{oraz} \quad S = \sqrt{0.199736653} = 0,446919068.$$

Obliczamy poszczególne wartości miar *ex-ante* na lata $2001 - 2020$ ($\tau = 1, 2, ..., 20$):

Przykład - Temperatura w Tokio 1875-2010- ocena *ex-ante*

Lata	Ocena *ex-ante* w %	lata	Ocena *ex-ante* w %
2001	2.7635	2011	2.7165
2002	2.7585	2012	2.7121
2003	2.7536	2013	2.7077
2004	2.7488	2014	2.7034
2005	2.7440	2015	2.6992
2006	2.7393	2016	2.6950
2007	2.7346	2017	2.6909
2008	2.7300	2018	2.6869
2009	2.7254	2019	2.6829
2010	2.7209	2020	2.6790

Można spodziwać się błędów w granicach 2.67% do 2.76%.

Przykład - Temperatura w Tokio 1875-2010- przedziały ufności

- Mając $n = 145$ danych i jedną zmienną objaśniającą szukamy 95% przedziałów ufności odpowiadającym $\alpha = 0.05$.
- Kwantyl $\mathcal{T}_{145-2}(\mathcal{T}_{n-k-1})$ rzędu $1 - \alpha/2$ wynosi $t_\alpha = 1.9767$;
- Sprawdzimy, czy przyszłe wartości znalazły się w przedziałach ufności.

Przykład - Temperatura w Tokio 1875-2010- przedziały ufności na lata 2001-2010

Lata	Temp.roczna	Dln. wart. ufności	Grn. wart. ufności
2001	16.49166667	15.658	17.468
2002	16.70833333	15.702	17.514
2003	16.00833333	15.747	17.560
2004	17.34166667	15.792	17.607
2005	16.21666667	15.837	17.654
2006	16.39166667	15.883	17.701
2007	16.96666667	15.929	17.749
2008	16.425	15.975	17.798
2009	16.69166667	16.022	17.846
2010	16.88333333	16.069	17.895

Przykład - Temperatura w Tokio 1875-2010- przedziały ufności na lata 2011-2020

Lata	Temp.roczna	Dln. wart. ufności	Grn.wart. ufności
2011	16.48333333	16.116	17.945
2012	16.3	16.164	17.995
2013	17.08333333	16.212	18.045
2014	16.58333333	16.260	18.096
2015	16.38333333	16.308	18.147
2016	16.44166667	16.357	18.198
2017	15.80833333	16.407	18.250
2018	16.79166667	16.456	18.302
2019	16.45	16.506	18.355
2020	16.53333333	16.556	18.408

Przykład - Temperatura w Tokio 1875-2010- przedziały ufności na lata 2011-2020

- Z poprzedniego wykładu widzieliśmy, że wszystkie prognozowane temperatury przekroczyły prawdziwe przyszłe wartości;
- Z przedziałów ufności wynika jednak, że większość prawdziwych wartości znalazły się w przedziałach ufności;
- Z drugiej jednak strony, trzy spośród obserwacji, znalazły się poniżej dolnych granic ufności (lata 2017,2019 i 2020, oznaczone na niebiesko).
- Ponieważ 3 spośród 20 obserwacji znalazły się poza przedziałem ufności, tylko $17/20 * 100\% = 85\%$ obserwacji zostało pokrytych przez przedział, a większość obserwacji była blisko dolnej granicy;
- Można uznać, że tych odstępstw nie wyjaśnić błędem statystycznym.