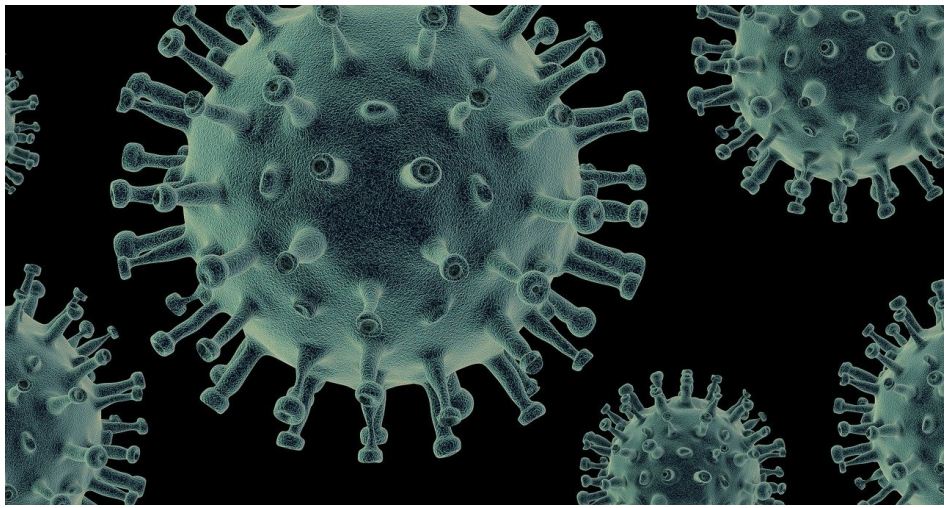
Prognozowanie - indeks Theila i jego rozkład

24 maja 2021



Model regresji wielokrotnej-powtórzenie

Kontynuujemy rozważenia modelu liniowego z wieloma zmiennymi objaśniającymi:

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

Model zapiszemy w formie macierzowej:

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

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 $N(0, \sigma^2)$.

Konstrukcja prognoz

Przypuśmy, że mamy już skonstruowany model i już zweryfikowaliśmy założenia [Z1-Z5] za pomocą poznanych testów. Estymator β ma postać

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

i jest on ostateczny. Pomijając wartości ϵ_t otrzymamy **prognozę zmiennej objaśnianej**:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}.$$

Dokładniej

$$\hat{\mathbf{Y}} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix},$$

gdzie

$$\hat{Y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_{t,1} + \ldots + \hat{\beta}_k x_{t,k}.$$

jest prognozą Y_t. Wartość

$$e_t := Y_t - \hat{Y}_t$$

nazywamy **błędem przybliżenia**.

Interpretacja

- Statystyk prowadzi badania w okresach $1, 2, \ldots, n$ i obserwuje następuje dane $(X_{t,1}, X_{t,2}, \ldots, X_{t,k}, Y_t)_{t=1}^{n}$;
- Bazując na tych danych statystyk buduje model pporzez aproksymację β, i bazując na obserwacjach konstruuje prognozy $\hat{Y}_t$ wartości Y_t for $t = 1, 2, ..., n$;
- Gdy statystyk jest świadomy przyszłych wartości zmiennych objaśniających $(X_{n+\tau,1}, X_{n+\tau,2}, \ldots, X_{n+\tau,k})$, a nieświadomy przyszłych wartości zmiennych objaśnianych $Y_{n+\tau}$ dla $\tau = 1, 2, ..., T$ wtedy statystyk może wyznaczyć *extrapolację prognozy*, czyli obliczyć prognozy $\hat{Y}_{n+\tau}$ przyszłych wartości $Y_{n+\tau}$ jako:

$$\hat{Y}_{n+\tau} = \hat{\beta}_0 + \hat{\beta}_1 X_{n+\tau,k} + \ldots + \hat{\beta}_k X_{n+\tau,k}.$$

Podstawowe miary dokładności prognoz typu *ex-post*

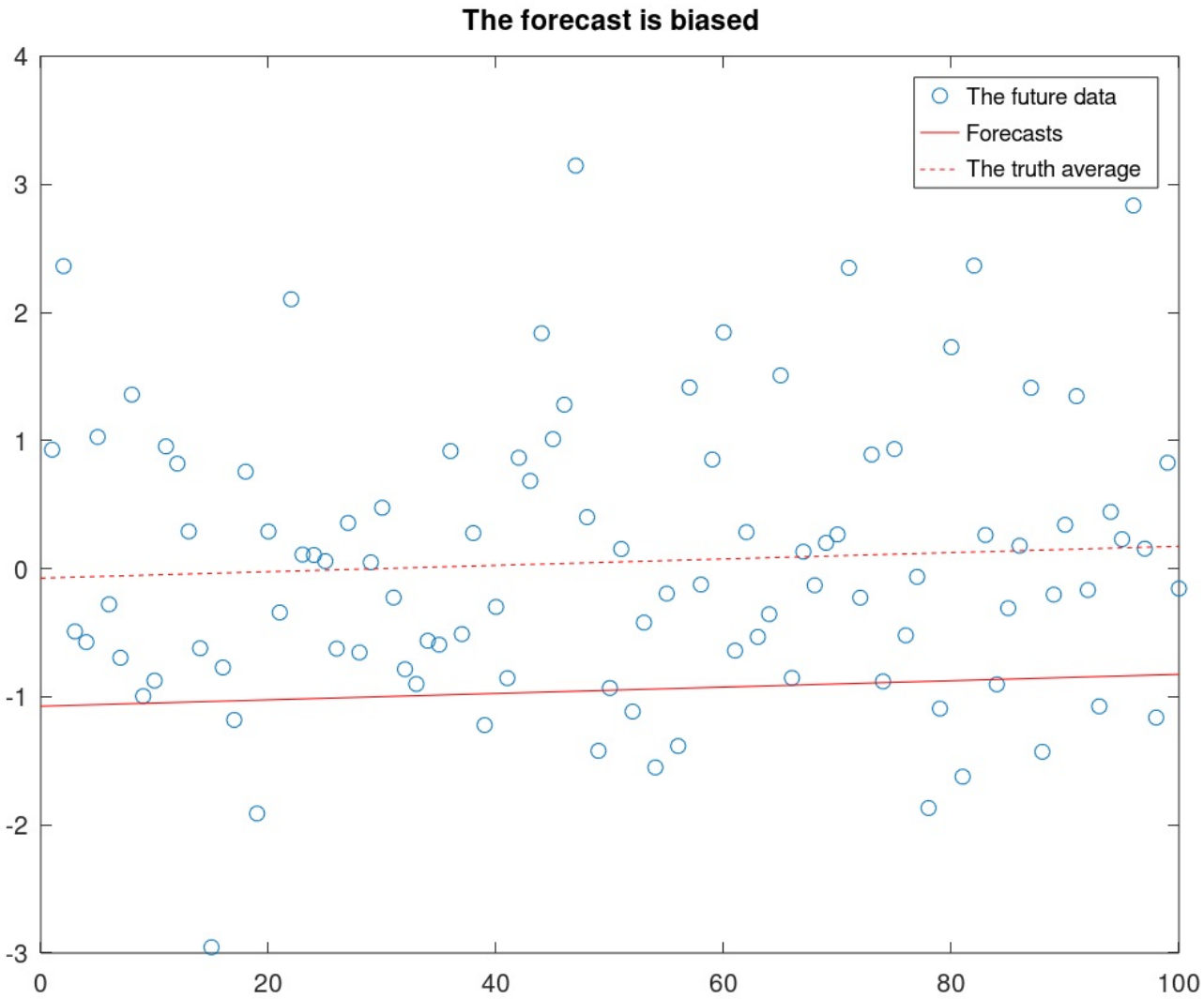
nazwa	nazwa angielska	wzór
Średni błąd	Mean Error	$ME = \frac{1}{T}\sum_{\tau=1}^{T}(Y_{n+\tau} - \hat{Y}_{n+\tau});$
Średni błąd bezwzględny	Mean Absolute Error	$MAE = \frac{1}{T}\sum_{\tau=1}^{T}\lvert Y_{n+\tau} - \hat{Y}_{n+\tau}\rvert;$
Pierwiastek błędu średniokwadratowego	Root Mean Square Error	$RMSE = \sqrt{\frac{1}{T}\sum_{\tau=1}^{T}(Y_{n+\tau} - \hat{Y}_{n+\tau})^2};$
Średni procentowy błąd średniokwadratowy	Mean Absolute Percentage Error	$MAPE = \frac{1}{T}\sum_{\tau=1}^{T}\left\lvert\frac{Y_{n+\tau}-\hat{Y}_{n+\tau}}{Y_{n+\tau}}\right\rvert * 100\%;$
Indeks rozbieżności	Inequality index	$U = \frac{\sqrt{\frac{1}{T}\sum_{\tau=1}^{T}(Y_{n+\tau}-\hat{Y}_{n+\tau})^2}}{\sqrt{\frac{1}{T}\sum_{\tau=1}^{T}Y_{n+\tau}^2}+\sqrt{\frac{1}{T}\sum_{\tau=1}^{T}\hat{Y}_{n+\tau}^2}};$
Indeks Theila	Theil Index	$I^2 = \frac{\frac{1}{T}\sum_{\tau=1}^{T}(Y_{n+\tau}-\hat{Y}_{n+\tau})^2}{\frac{1}{T}\sum_{\tau=1}^{T}Y_{n+\tau}^2};$

Źródła błędów prognoz

Błędy prognoz mają następujące przyczyny:

- *obciążoność prognoz (an. biasness):* złe określenie wartość średniej arytmetycznej prognozy;
- *brak elastyczności (inelasticity):* nieprawidłowe określenie przedziału zmienności prognozy:
- Inne przyczyny jak np. niepełna korelacja między wartością, a prognozą, lub jej brak.

Źródła błędów prognoz-ilustracja obciążoności

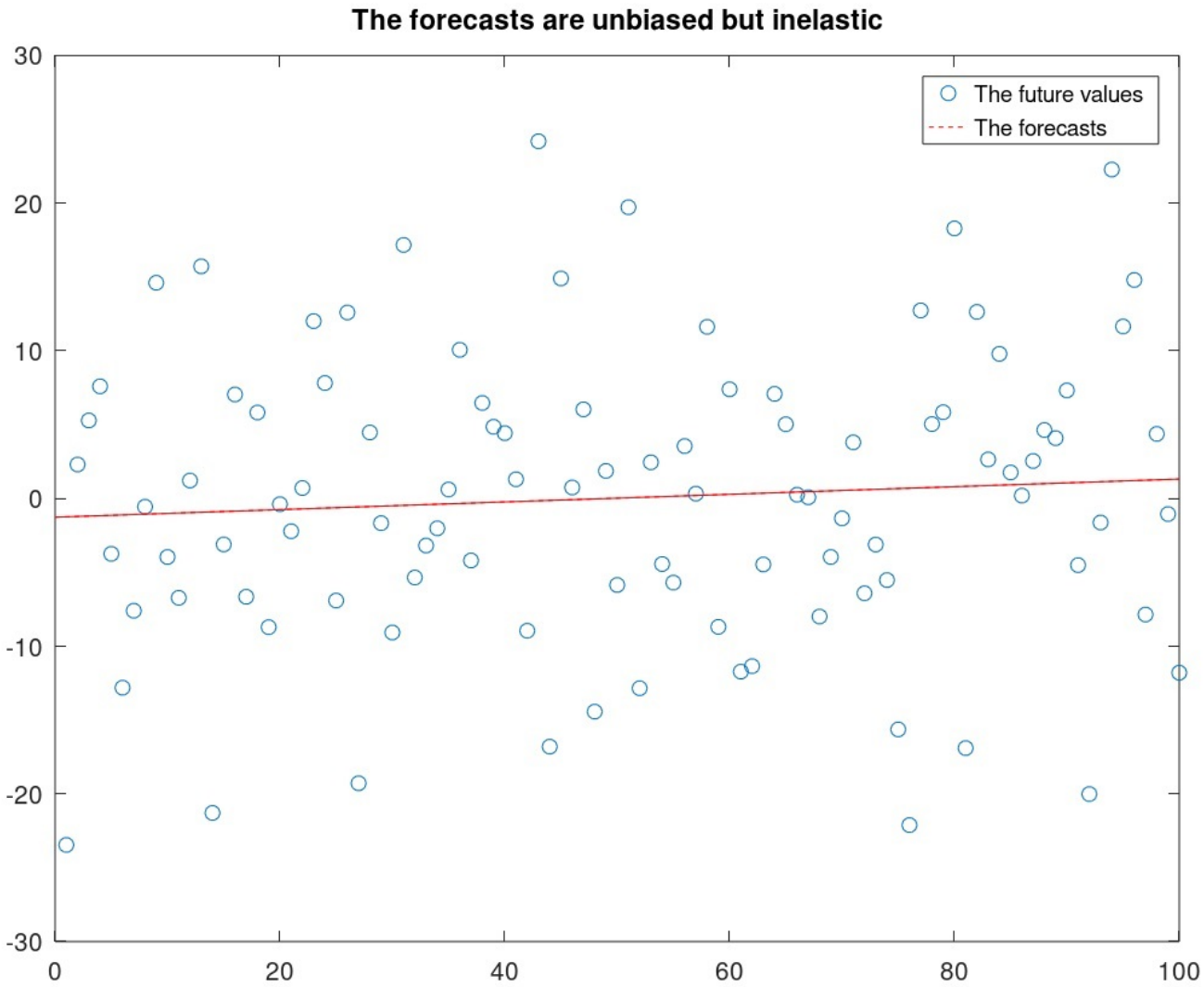


Źródła błędów prognoz-ilustracja obciążoności

Interpretacja wykresu:

- Linia przerywana ilustruje średnią wartość przyszłych wartości;
- Czerwona linia ciągła oznaczająca średnią prognoz jest mocno poniżej średniej dokładnych wartości;
- Wtedy wartość ME jest dodatnia, a i pozostałe miary *ex-post* są dalekie od 0.

Źródła błędów prognoz-ilustracja braku elastyczności



Źródła błędów prognoz-ilustracja braku elastyczności

Interpretacja wykresu:

- Czerwona linia ciągła ilustruje nieobciążoną prognozę;
- Ale wahania dokładnych wartości są bardziej chaotyczne niż prognoza;
- Wtedy ME jest blisko zera, ale pozostałe miary zwrócą wartości dalekie od zera.

Składniki podsumowujące źródła błędów

- **Nieobciążoność:** średnie arytmetyczne przyszłych danych i ich prognoz:

$$\underbrace{\overline{Y} := \frac{1}{T}\sum_{\tau=1}^{T} Y_{n+\tau}}_{\text{średnia przyszłych wartości}} \quad , \quad \underbrace{\overline{\hat{Y}} := \frac{1}{T}\sum_{\tau=1}^{T} \hat{Y}_{n+\tau}}_{\text{średnia prognoz}};$$

- **Brak elastyczności:** błędy standardowe przyszłych danych i ich prognoz:

$$\underbrace{SE(Y) := \sqrt{\frac{1}{T}\sum_{\tau=1}^{T}(Y_{n+\tau} - \overline{Y})^2}}_{\text{błąd standardowy przyszłych danych}} \quad , \quad \underbrace{SE(\hat{Y}) := \sqrt{\frac{1}{T}\sum_{\tau=1}^{T}(\hat{Y}_{n+\tau} - \overline{\hat{Y}})^2}}_{\text{błąd standardowy prognoz}};$$

- **Inne źródła:** korelacja między danymi, a prognozami:

$$\rho = \frac{\frac{1}{T}\sum_{\tau=1}^{T}\left(\hat{Y}_{n+\tau} - \overline{\hat{Y}}\right)\left(Y_{n+\tau} - \overline{Y}\right)}{SE(\hat{Y})SE(Y)}.$$

Parametryzacja źródeł błędów

- **Obciążoność** jest podsumowana przez różnicę między $\overline{Y}$ i $\overline{\hat{Y}}$
  - Niezgodność między średnimi oznacza, że prognozy oscylują między niewłaściwymi wartościami;
- Brak elestyczności podsumowuje różnica między $SE(\overline{Y})$ oraz $SE(\overline{\hat{Y}})$;
  - Niezgodność między błędami oznacza zbyt duże lub zbyt małe wahania prognoz względem przyszłych wartości i to również powoduje błędy nawet jeśli prognozy oscylują wokół prawidłowej wartości;
- Inne źródła błędów podsumowuje wartość ρ.
  - mała wartość ρ zwiększa ryzyko losowego błędu, nawet jeśli powyższe źródła zostały wyeliminowane;

Rozkład współczynnika Theila

Twierdzenie

Indeks Theila może być rozłożony na trzy składniki, z których każdy podsumowuje inne źródło błędów:

$$I^2 = I_1^2 + I_2^2 + I_3^2,$$

gdzie

$$I_1^2 = \frac{\left(\overline{Y} - \overline{\hat{Y}}\right)^2}{\frac{1}{T}\sum_{\tau=1}^{T} Y_{n+\tau}^2}, \quad I_2^2 = \frac{\left(SE(Y) - SE(\hat{Y})\right)^2}{\frac{1}{T}\sum_{\tau=1}^{T} Y_{n+\tau}^2}, \quad I_3^2 = \frac{2(1-\rho)SE(Y)SE(\hat{Y})}{\frac{1}{T}\sum_{\tau=1}^{T} Y_{n+\tau}^2}.$$

Rozkład współczynnika Theila

Zauważmy:

- Wzór I_1^2 podsumowuje nieobciążoność (niewłaściwe oszacowanie średniej);

$$I_1^2 = \frac{\overbrace{\left(\overline{Y} - \overline{\hat{Y}}\right)^2}^{\text{odległość między średnimi}}}{\frac{1}{T}\sum_{\tau=1}^{T} Y_{n+\tau}^2}$$

- I_2^2 podsumowuje brak elastyczności modelu (niewłaściwe oszacowanie błędu średniokwadratowego);

$$I_2^2 = \frac{\overbrace{\left(SE(Y) - SE(\hat{Y})\right)^2}^{\text{odległość między błędami średniokwadratowymi}}}{\frac{1}{T}\sum_{\tau=1}^{T} Y_{n+\tau}^2};$$

Rozkład współczynnika Theila

- I_3^2 podsumowuje inne źródła błędu koncentrując się na braku korelacji między prognozami, a dokładnymi wartościami:

$$I_3^2 = 2\frac{\overbrace{(1-\rho)}^{\text{małe } \rho \text{ jest nowym źródłem błędu}} SE(Y)SE(\hat{Y})}{\frac{1}{T}\sum_{\tau=1}^{T} Y_{n+\tau}^2}.$$

Rozkład współczynnika Theila

Podsumowanie

Współczynnik (indeks) Theila, I^2 jest miarą całkowitego błędu prognozy. Rozkład I^2 wskazuje na rozkład sił poszczególnych źródeł błędu:

- I_1^2 mierzy brak zgodności między średnią przyszłymi wartościami, a średnią ich prognoz:
  - dobry model powinien mieć ten problem pod kontrolą;
- I_2^2 mierzy brak zgodności między wahaniami przyszłych wartości i ich prognoz:
  - dobry model powinien mieć ten problem pod kontrolą;
- I_3^2 mierzy wartości innych źródeł błędu jak braki korelacji między przyszłymi wartościami, a ich prognozami:
  - ponieważ model jest losowy, to źródło jest poza naszą kontrolą, nawet dla najlepszych modeli.

Rozkład współczynnika Theila

Wyrażamy rozkład Theila w następujący sposób:

- Proporcja błędu spowodowana przez nieobciążoność (różnica między średnimi):

$$\tilde{I}_1^2 := \frac{I_1^2}{I^2} * 100\%;$$

- Proporcja błędu spowodowana przez brak elastyczności (różnica wahań):

$$\tilde{I}_2^2 := \frac{I_2^2}{I^2} * 100\%;$$

- Proporcja błędu spowodowana przez inne przyczyny:

$$\tilde{I}_3^2 := \frac{I_3^2}{I^2} * 100\%;$$

Rozkład współczynnika Theila

Rozkład współczynnika Theila dla modelu doskonałego

Model który doskonale szacuje dane (doskonały w skrócie) powinien mieć pełną kontrolę nad nieobciążonością i brakiem elastyczności, ale nie może mieć kontroli nad pozostałymi źródłami błędów. Z tego powodu rozkład Theila dla modelu doskonałego to

$$\tilde{I}_1^2 = 0, \quad \tilde{I}_2^2 = 0, \quad \tilde{I}_3^2 = 100\%.$$

Będziemy usatysfakcjonowani dokładnością prognoz jeśi $\tilde{I}_3^2$ jest blisko 100%.

Roczna temperatura w Japonii na podstawie pomiarów na wyspie Hachiō-jima (Japońska wyspa wulkaniczna)

Roczny rozkład temperatur na japońskiej wyspie Hachiō-jima w latach 1907-2020

Przykład z poprzedniego wykładu

Dysponujemy danymi dotyczącymi rocznej temperatury na wyspie Hachiō-jima, w latach 1907-2020. Za pomocą miar *ex-post* zbadamy dokładność prognozami przyszych wartości. Dokładniej:

- Dzielimy dany na *teraźniejsze dane*, które obejmują lata 1907-2000, oraz *przyszłe dane* które obejmują temperatury w latach 2001-2020;
- Na podstawie *teraźniejszych danych* budujemy model i prognozy;
- Porównamy *przyszłe dane* z ich prognozami, które stworzyliśmy na bazie *teraźniejszych danych*.

Roczny rozkład temperatur na japońskiej wyspie Hachiō-jima w latach 1907-2020

Na podstawie *teraźniejszych danych* 1907-2000 znaleźliśmy model w postaci:

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

gdzie t oznacza indeks roku począwszy od 1907, a kończywszy na 2000, a Y_t oznacza roczną temperaturę w roku t. Stąd $t = 1, 2, \ldots, n$ z $n = 94$.

- Obliczamy estymator $\hat{\beta}_0 = 17.76023868$, oraz $\hat{\beta}_1 = 0.006594147$ i prognozy

$$\hat{Y}_t = 17.76023868 + 0.006594147 * t$$

dla indeksów lat $2001 - 2020$ ($t = 95, ..., 114$);

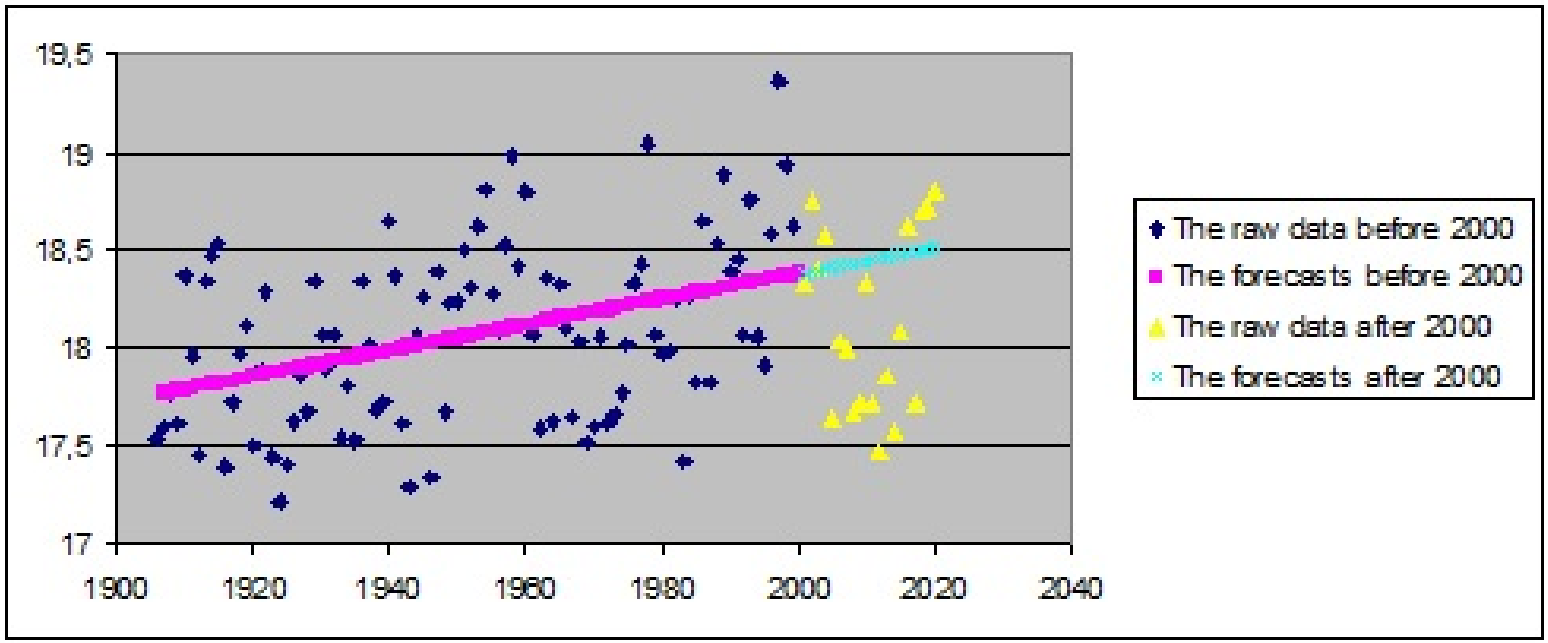
- Za pomocą miar ex-post porównamy prognozy z prawdziwymi "przyszłymi" temperaturami w latach 2001-2020 (bazując na wcześniejszych "teraźniejszych" danych 1907-2000).

Roczny rozkład temperatur na japońskiej wyspie Hachiō-jima w latach 1907-2020

Prognozy i dokładne wartości znajdziemy w poniższej tabeli.

Year	Y_t	$\hat{Y}_t$
2001	18,31667	18,38668
2002	18,75833	18,39328
2003	18,40833	18,39987
2004	18,575	18,40647
2005	17,64167	18,41306
2006	18,03333	18,41965
2007	17,99167	18,42625
2008	17,675	18,43284
2009	17,725	18,43944
2010	18,33333	18,44603
2011	17,725	18,45262
2012	17,46667	18,45922
2013	17,85833	18,46581
2014	17,575	18,47241
2015	18,08333	18,479
2016	18,61667	18,48559
2017	17,725	18,49219
2018	18,70833	18,49878
2019	18,71667	18,50538
2020	18,80833	18,51197

Roczny rozkład temperatur na japońskiej wyspie Hachiō-jima z podziałem na "dane testujące" 1907-2000 i "dane testowane" 2001-2020



Roczny rozkład temperatur na japońskiej wyspie Hachiō-jima

- Niebieska linia ciągła i żółte gwiazdki odzwierciedlają jakość prognoz;
- Widać na pierwszy rzut oka, że prognozy przeszacowały przysze wartości;
- Prognozy wahają się między 18.40, 18.5 stopni Celsjusza, podczas gdy prawdziwe wartości wahają się między 17.50, a $18.80^{o}C$, zatem rola w braku elastyczności powinna być znacząca;
- Model nie powinien być w pełni satysfakcjonujący;
- Rozkład Theila powinien wyjaśnić jakość modelu.

Roczny rozkład temperatur na japońskiej wyspie Hachiō-jima

Mamy

$$\frac{1}{20}\sum_{\tau=1}^{20} Y_{94+\tau}^2 = 329.15, \quad \text{oraz} \quad I^2 = 0.000884841.$$

Dalej mamy,

$$\overline{Y} = 18.14, \quad \overline{\hat{Y}} = 18.45, \quad \text{stąd } I_1^2 = \frac{(18.14 - 18.45)^2}{329.15} \approx 0.000296206,$$

oraz

$$SE(Y) = 0.44, \quad SE(\overline{\hat{Y}}) = 0.04, \quad \text{hence } I_2^2 = \frac{(0.44 - 0.04)^2}{329.15} \approx 0.000496355.$$

Na koniec,

$$I_3^2 = I^2 - I_1^2 - I_2^2 = 9,22799 * 10^{-05}.$$

Roczny rozkład temperatur na japońskiej wyspie Hachiō-jima

Rozkład I^2 jest zaprezentowany w tabeli

obciążoność	brak elastyczności	inne źródła błędu	Błąd całkowity
I_1^2	I_2^2	I_3^2	$I^2 = I_1^2 + I_2^2 + I_3^2$
0.000296206	0.000496355	$9.22799 * 10^{-05}$	0.000884841
$\tilde{I}_1^2$	$\tilde{I}_2^2$	$\tilde{I}_3^2$	$\tilde{I}_1^2 + \tilde{I}_2^2 + \tilde{I}_3^2$
33.476%	56.095%	10.429%	100%

Roczny rozkład temperatur na japońskiej wyspie Hachiō-jima

Z rozkładu I^2 wnioskujemy, że

- Obciążoność jest odpowiedzialne za błąd w 33.476%;
- Brak elastyczności jest odpowiedzialne za błąd w 56.095%;
- Inne źródła są odpowiedzialne za błąd w 10.429%;
- Model jest obciążony dużym ryzykiem błędu, więc jest daleki od doskonałego.