Wstęp do prognozowania

ASC 2023

Piotr Żoch

Plan

- Jakość prognozy
- Prognoza naiwna
- Średnia ruchoma
- Wygładzanie wykładnicze

Jakość prognozy

Interpolacja i ekstrapolacja

- **Interpolacja:** wyznaczenie w pewnym przedziale funkcji, która przyjmuje znane wartości dla danych liczb z tego przedziału
    - Wygładzanie.
- **Ekstrapolacja:** wyznaczanie wartości funkcji na zewnątrz przedziału, w którym wartości tej funkcji są znane.
    - Prognozowanie.

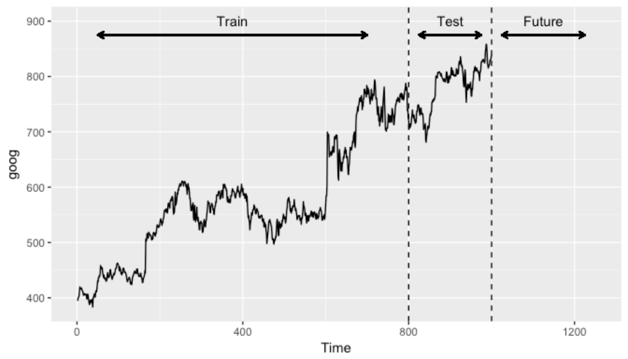
Jakość prognozy

- Niech y_t oznacza faktyczną wartość zmiennej w okresie t, a y_t^f **wartość prognozowaną** (f jak *forecast*, inne częste oznaczenie to $\hat{y}_t$).
- **Błąd prognozy** to

$$e_t := y_t - y_t^f$$

- Uwaga: to błąd **ex post** - musimy znać faktyczną wartość!
- Typowe podejście: dzielimy dane na dwa podzbiory: *training set* i *test set*
    - *training set* wykorzystujemy do estymacji parametrów modelu
    - *test set* wykorzystujemy do oceny jakości prognoz

Jakość prognozy



Źródło: http://uc-r.github.io/ts_benchmarking

Jakość prognozy

- Reszty

$$e_t = y_t - y_t^f$$

  powinny:

  - mieć średnią zero (inaczej prognoza *obciążona*)
  - być nieskorelowane (inaczej nie wykorzystaliśmy części informacji w danych)
- Pożądane właściwości:
  - stała wariancja
  - rozkład normalny

Jakość prognozy

- **Błąd prognozy ex ante:** opisuje *dopuszczalność* prognozy.
    - przed upływem czasu, na który prognoza była ustalona
- **Błąd prognozy ex post:** opisuje *trafność* prognozy.
    - można obliczyć, gdy znana jest realizacja zmiennej prognozowanej
    - to nimi będziemy się zajmować głównie na tych zajęciach

Miary błędów ex post

- Notacja: używamy n obserwacji do prognozy na okresy $n+1, n+2, \ldots, T$.
- **MAE** - mean absolute error
$$\frac{1}{T-n}\sum_{t=n+1}^{T}|e_t|$$
- **MSE** - mean square error
$$\frac{1}{T-n}\sum_{t=n+1}^{T}e_t^2$$
- **RMSE** - *root* mean square error
$$\sqrt{\frac{1}{T-n}\sum_{t=n+1}^{T}e_t^2}$$
- Powyższe *nie* spełniają warunku unormowania.

Miary błędów ex post

- **MAPE -** mean absolute *percentage* error

$$\frac{1}{T-n}\sum_{t=n+1}^{T}\left|\frac{e_t}{y_t}\right| \times 100\%$$

- **AMAPE -** *adjusted* MAPE

$$\frac{1}{T-n}\sum_{t=n+1}^{T}\left|\frac{e_t}{y_t+y_t^f}\right| \times 100\%$$

Przedziały

- Często interesuje nas przedział, w którym y_t będzie znajdować się z określonym prawdopodobieństwem.
- Na przykład, jeśli założymy że błędy prognozy mają rozkład normalny, to

$$y_{t+h} \in \left[y_{t+h}^{f} - 1.96\hat{\sigma}_h, y_{t+h}^{f} + 1.96\hat{\sigma}_h\right]$$

  z prawdopodobieństwem 0.95, gdzie $\hat{\sigma}_h$ to oszacowanie odch. standardowego błędu prognozy o horyzoncie h.

Przedziały

- $\hat{\sigma}_1$ można obliczyć szacując odchylenie standardowe reszt.
- $\hat{\sigma}_h$ zazwyczaj rośnie wraz z h. Obliczenie nieco trudniejsze, zwłaszcza gdy reszty są ze sobą skorelowane.

Metody naiwne

Metody naiwne

- Najprostszy typ metod wykorzystywanych do prognozowania.
- Założenie: brak zmian czynników oddziaływujących na zmienną prognozowaną.
    - stosowane w przypadku niewielkich wahań przypadkowych zmiennej zależnej.
- Zazwyczaj niska jakość prognoz.

Metody naiwne

- Najprostsza prognoza naiwna:

$$y_t^f = y_{t-1}$$

  gdzie y_t^f to prognoza wartości zmiennej zależnej na moment t a y_{t-1} to faktyczna wartość tej zmiennej w momencie $t-1$.
- Możemy też mieć prognozę z horyzontem h:

$$y_{t+h}^f = y_t$$

- Co z trendem, sezonowością?

Metody średniej ruchomej

Metody średniej ruchomej

- Wykorzystywane do
  - **wygładzania** szeregu czasowego
  - **prognozowania** - prognozowana wartość zmiennej to średnia z k poprzednich obserwacji

Metody średniej ruchomej - wygładzanie

$$y_t^s = \frac{1}{2n+1} \sum_{i=-n}^{n} y_{t-i}$$

gdzie y_t^s to wartość zmiennej w okresie t po wygładzaniu a $y_{t-n}, y_{t-n+1}, \ldots, y_{t+n}$ to faktyczne wartości zmiennej.

- Do wygładzania używamy obserwacji w okresie t oraz n poprzedzających i n kolejnych obserwacji.
- Częste oznaczenie: $k = 2n + 1$ (k to *stała wygładzania szeregu*).

Metody średniej ruchomej - prognozowanie

$$y_t^f = \frac{1}{k} \sum_{i=t-k}^{t-1} y_i$$

gdzie y_t^f to prognozowana wartość zmiennej w okresie t a $y_{t-k}, y_{t-k+1}, \ldots, y_{t-1}$ to faktyczne wartości zmiennej.

- Do prognozowania używamy obserwacji w okresie $t-1$ oraz $k-1$ poprzedzających obserwacji.
- Dla $k = 1$ sprowadza się do metody naiwnej.

Jak wybrać k?

- Jakie k jest najlepsze?
- Pomysł: jak dobrze prognozujemy dane, które zaobserowaliśmy?
  - błąd prognozy ***ex post***.
- Szukamy k, które minimalizuje

$$MSE_k := \frac{1}{T-k}\sum_{t=k+1}^{T}\left(y_t - y_t^f\right)^2$$

Ważona średnia ruchoma

- Obserwacje starsze mają mniejsze znaczenie

$$y_t^f = \sum_{i=t-k}^{t-1} w_{i+t+k+1} y_i$$

gdzie

$$\sum_{i=1}^{k} w_i = 1$$

$$0 < w_1 < w_2 < \cdots < w_k \leq 1$$

Metody wygładzania wykładniczego

Proste wygładzanie wykładnicze

- Stosowane w przypadku braku trendu i sezonowości

$$y_t^f = \alpha y_{t-1} + (1 - \alpha)\, y_{t-1}^f$$

alternatywnie

$$y_t^f = y_{t-1}^f + \alpha \left(y_{t-1} - y_{t-1}^f \right)$$

Proste wygładzanie wykładnicze

- Skąd nazwa?

$$
\begin{aligned}
y_t^f &= \alpha y_{t-1} + (1-\alpha)\, y_{t-1}^f \\
&= \alpha y_{t-1} + (1-\alpha)\left[\alpha y_{t-2} + (1-\alpha)\, y_{t-2}^f\right] \\
&= \alpha y_{t-1} + \alpha\,(1-\alpha)\, y_{t-2} + (1-\alpha)^2\, y_{t-2}^f \\
&= \alpha y_{t-1} + \alpha\,(1-\alpha)\, y_{t-2} + \alpha\,(1-\alpha)^2\, y_{t-3} + \cdots
\end{aligned}
$$

- Coraz mniejsze wagi przeszłych obserwacji.

Proste wygładzanie wykładnicze

- Jak wybrać α?
- Najczęściej $\alpha \in [0.2, 0.3]$. Niższa wartość = dłuższa pamięć.
  - Dla $\alpha = 0$ mamy $y_t^f = y_{t-1}^f$
  - Dla $\alpha = 1$ mamy $y_t^f = y_{t-1}$
- Za y_1^f podstawia się y_1 lub średnią z kilku pierwszych obserwacji.
- W przypadku występowania trendu stosowane **podwójne** wygładzanie wykładnicze.

Podwójne wygładzanie wykładnicze - metoda Holta

- Poziom + trend.

$$y^f_{t-1+k} = L_{t-1} + kT_{t-1}$$

gdzie poziom to

$$L_t = \alpha y_t + (1-\alpha)\left(L_{t-1} + T_{t-1}\right)$$

a trend to

$$T_t = \beta\left(L_t - L_{t-1}\right) + (1-\beta)\, T_{t-1}$$

Podwójne wygładzanie wykładnicze - metoda Holta

- Równanie prognozy na okres $t > T$:

$$y_t^f = L_T + (t - T)\, T_T$$

gdzie:

  - y_t^f - prognoza zmiennej wyznaczona na moment t
  - L_T - wygładzona wartość zmiennej prognozowanej na okres T
  - T_T - wygładzona wartość przyrostu trendu na okres T
  - T - liczba obserwacji zmiennej prognozowanej

Podwójne wygładzanie wykładnicze - metoda Holta

- W przypadku metody Holta mamy:

$$L_t = y_t^f + \alpha \left(y_t - y_t^f\right)$$
$$T_t = T_{t-1} + \alpha\beta \left(y_t - y_t^f\right)$$
$$y_t^f = L_{t-1} + T_{t-1}$$

Addytywny model Wintersa

- Poziom + trend + sezonowość.

$$
\begin{aligned}
y^f_{t-1+k} &= L_{t-1} + kT_{t-1} + S_{t+k-m(z+1)} \\
L_t &= \alpha\left(y_t - S_{t-m}\right) + (1-\alpha)\left(L_{t-1} + T_{t-1}\right) \\
T_t &= \beta\left(L_t - L_{t-1}\right) + (1-\beta)\, T_{t-1} \\
S_t &= \gamma\left(y_t - L_{t-1} - T_{t-1}\right) + (1-\gamma)\, S_{t-m}
\end{aligned}
$$

gdzie m oznacza długość cyklu sezonowego (np. 4 albo 12) a $z = \lfloor \frac{(k-1)}{m} \rfloor$.

Multiplikatywny model Wintersa

- Poziom + trend $\times$ sezonowość.

$$
\begin{aligned}
y^f_{t-1+k} &= (L_{t-1} + kT_{t-1}) \times S_{t+k-m(z+1)} \\
L_t &= \alpha \left(\frac{y_t}{S_{t-m}} \right) + (1-\alpha)(L_{t-1} + T_{t-1}) \\
T_t &= \beta (L_t - L_{t-1}) + (1-\beta) T_{t-1} \\
S_t &= \gamma \left(\frac{y_t}{L_{t-1} + T_{t-1}} \right) + (1-\gamma) S_{t-m}
\end{aligned}
$$

gdzie m oznacza długość cyklu sezonowego (np. 4 albo 12) a $z = \lfloor \frac{(k-1)}{m} \rfloor$.