# Przewidywanie wskaźnika INDPRO z wykorzystaniem metod statystycznych

Presentation by **Filip Kin**
January 2022

pwc

# Badanie zbioru danych

Ponieważ nazwy zmiennych są nieopisane, nie jesteśmy w stanie posłużyć się żadnym twierdzeniem ekonomicznym, dlatego wykorzystamy metody matematyczno-statystyczne. Z analizy eksploracyjnej wynikło, że:

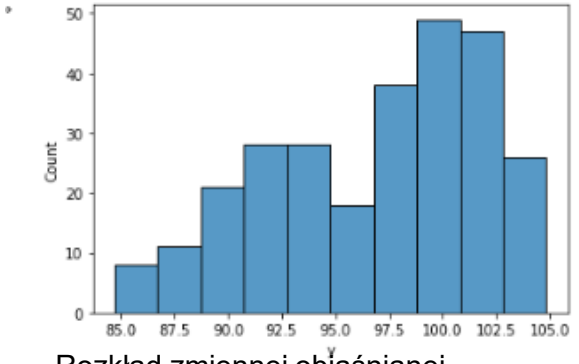
- Mamy do czynienia z szeregiem czasowym, na który wpływa wiele parametrów.
- Cały zbiór liczy 275 rekordów, mamy jedną zmienną objaśnianą oraz 5 zmiennych objaśniających.
- Po zbadaniu rozkładu zmiennej objaśnianej ukazuje się rozkład leptokurtyczny lewostronny.
- Testy korelacji wykluczają zależność zmiennych objaśniających od siebie i wykazują korelację tychże zmiennych ze zmienną objaśnianą.
- Dzięki użyciu testu KPSS oraz ADF możemy dojść do wniosku, że model nie jest stacjonarny.
- Za pomocą testu ACF dochodzimy do wniosku, że mamy do czynienia z szeregiem autoregresyjnym.

Dzięki powyższym informacjom dobieramy 3 modele, które mają największe szanse powodzenia w prognozowaniu:

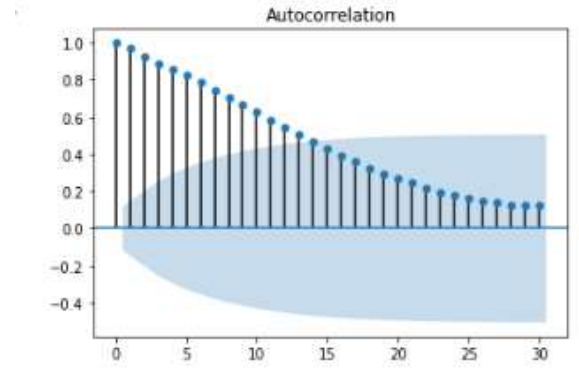
1. Model AR(1)
2. Model regresji liniowej
3. Model regresji wielomianowej

| | y | x1 | x2 | x3 | x4 | x5 |
|---|---|---|---|---|---|---|
| 0 | NaN | 2.456 | 4.00 | 1386.96 | 6.4 | 406426 |
| 1 | 104.7474 | 2.837 | 4.04 | 1462.59 | 4.4 | 405948 |
| 2 | 104.8129 | 2.578 | 3.56 | 1736.47 | 4.6 | 405466 |
| 3 | 104.4135 | 2.793 | 3.13 | 1688.85 | 18.0 | 403957 |
| 4 | 104.4852 | 3.283 | 3.10 | 1822.57 | 4.4 | 400415 |

Pięć pierwszych rekordów




Rozkład zmiennej objaśnianej




Test ACF dla zmiennej objaśnianej

# Porównanie modeli statystycznych/uczenia maszynowego

## Model AR

Z uwagi na zanikającą wykładniczo autokorelację można użyć modelu autoregresyjnego AR(1). Z wydruku wynika, że wyraz wolny jest nieistotny a kryteria informacyjne są dość niskie. Niestety, przy próbie prognozowania model nie może poszczycić się zbyt duża dokładnością, ponieważ mimo zmieniania liczby opóźnień, ciągle prognozuje 0.

## Regresja liniowa

Najpopularniejsza forma prognozowania. Z wydruku wiadomo, że

1. wszystkie zmienne są statystycznie istotne
2. test Durbina-Watsona pokazuje, że mamy do czynienia z autokorelacją dodatnią
3. test Jarque-Berra udowadnia, że rozkład zmiennej objaśnianej skośnością i kurtozą przypomina rozkład normalny

Błąd RMSE wyniósł w przybliżeniu 2.505 Ponadto adjusted R^2 wynosi 66% na danych treningowych i 73% na danych testowych. Walidacja krzyżowa nie wykazała, aby niższa liczba parametrów znacząco poprawiała wyniki. Niezły wynik, jednak można spróbować użyć innego modelu.

## Regresja wielomianowa

Istnieje szansa, że nie mamy do czynienia z zależnością liniową. Z tego powodu używamy regresji wielomianowej stopnia drugiego, czyli funkcji kwadratowej. RMSE jest niższy niż w przypadku regresji liniowej i wynosi w przybliżeniu 2.21. Współczynnik determinacji 83 % na danych treningowe oraz 79% na danych jest zadowalający. Walidacja krzyżowa tak jak w przypadku regresji liniowej nie wykazała, aby niższa liczba parametrów poprawiała wyniki. Błąd

# Dobór ostatecznego modelu i podsumowanie

Ostatecznie decydujemy się na użycie regresji wielomianowej, ponieważ przyniosła ona najlepsze rezultaty z wybranych modeli, nie wykazując przy tym przetrenowania ani niedotrenowania.

Z tego wynika, że trend w szeregu nie jest liniowy, jednakowoż mała ilość danych oraz zbyt krótki horyzont czasowy może zakrzywiać rzeczywistość.

Zważając na powyższy fakt, współczynnik determinacji R^2 wynoszący 79% jest satysfakcjonujący.