Ekonometria wykład

Tomasz Przechlewski

4 listopada 2023

O czym będzie

Ekonometria – nauka pomocnicza w ramach ekonomii, wykorzystująca narzędzia matematyki i statystyki do badania ilościowych związków zachodzących między zjawiskami ekonomicznymi.

Większość jej metod opracowano poza ekonomią (zaadaptowane z innych nauk).

Celem ekonometrii jest weryfikacja teorii ekonomicznych, przewidywanie procesów ekonomicznych [oraz dostarczanie przesłanek służących sterowaniu tymi procesami.] Podstawowym narzędziem służącym tym celom jest model ekonometryczny.

https://pl.wikipedia.org/wiki/Ekonometria

Econometric models are constructed from economic data with the aid of the techniques of statistical inference.

Modele ekonometryczne to modele statystyczne dotyczące danych ekonomicznych.

Dane (ekonomiczne)

Szeregi czasowe

Badany **obiekt** jest mierzone wielokrotnie, w jednakowych odstępach czasu.

Ale niekoniecznie tak jest: dane dzienne zwykle nie zawierają świąt. Wprawdzie miesięcy jest zawsze 12 w roku ale luty jest 10% krótszy niż styczeń

Time series data is data collected on single unit at regular intervals over time. **Cross-sectional** data is data collected on many units at the same point in time

Obserwacje mogą być skorelowane (zwykle są); problem autokorelacji

Wzrost Stasia w okresie 2010–2020 to szereg czasowy (pozostaje do ustalenia kiedy pomiary są dokonane: raz na miesiąc, raz na kwartał, raz na rok)

Poszczególne obserwacje są skorelowane: jak Staś mial 160cm to pewne za kwartał będzie miał 161cm a nie 190cm…

Przekrojowe

Pomiar jest dokonany w jednym momencie lub okresie czasu ale mierzonych jest wiele obiektów.

Wzrost dzieci w klasie Stasia w dniu 31 grudnia 2020 to dane przekrojowe

Obserwacje raczej nie są skorelowane: jeżeli Jaś ma 170cm to nie wiadomo ile ma Staś albo Gosia…

Panelowe

Połączenie szeregów i danych przekrojowych

Wzrost dzieci w klasie Stasia w okresie 2010–2020

Etapy budowy modelu ekonometrycznego

– specyfikacja zmiennych i wybór analitycznej postaci modelu (najlepiej na podstawie jakieś teorii ekonomicznej a nie *ad hoc*),

– estymacja parametrów,

– weryfikacja modelu,

– praktyczne wykorzystanie oszacowanego modelu w tym a zwłaszcza **prognozowanie**

Model regresji liniowej (Linear regression model; LR model)

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + ... + \beta_k x_{ki} + u_i$$

y jest określane jako zmienna zależna (*dependent variable*);

zmiennie $x_1, ... x_k$ są określane jako zmienne niezależne albo wyjaśniające (*explanatory variables*)

u jest określany jako składnik losowy (*error term*)

Model regresji liniowej z jedną zmienną objaśniającą

$$y_i = \beta_1 + \beta_2 x_i + u_i$$

Założenie 1:

Obserwacje na zmiennej objaśniającej są ustalone (nielosowe)

Założenie 2:

Składniki losowe u_i są niezależnymi zmiennymi losowymi:

$$cov(u_i, u_j) = 0 \qquad \text{dla } : i \neq j$$

o jednakowych rozkładach prawdopodobieństwa. Alternatywnie: brak autokorelacji, tj. korelacja pomiędzy dwoma dowolnymi u_i wynosi zero.

z wartością oczekiwaną równą zero:

$$E(u_i|x) = 0$$

oraz stałą wariancją

$$D^2(u_i|x) = \sigma^2$$

Z powyższych wynika, że funkcja regresji y względem x, tj:

$$\hat{y}_i = E(y_i|x_i) = \beta_1 + \beta_2 x_i$$

$\hat{y}_i$ to warunkowe wartości oczekiwane zmiennej y.

Składnik losowy można interpretować jako odchylenie (błąd pomiaru na przykład) od nieznanej funkcji regresji.

Nieznane parametry funkcji regresji **estymuje się** za pomocą metody najmniejszych kwadratów (MNK albo KMNK; *least squares* albo *ordinary last squares OLS*)

Twierdzenie Gaussa-Markowa

Estymatory KMNK są nieobciążone i mają najmniejszą wariancję w klasie estymatorów liniowych (są efektywne; *efficient*)

Konwencje zapisu oraz fachowe nazewnictwo:

x, y – prawdziwe wartości zmiennych

β – prawdziwe wartości parametrów równania regresji (nieznane)

$\hat{y}$ – teoretyczne wartości zmiennej wynikające z równania regresji

b – estymatory nieznanych wartości parametrów β, czytaj: formuły które po podstawieniu konkretnych wartości dla x i y zwracają **przypuszczalne** wartości β zwane ocenami

$\hat{b}$ – oceny parametrów β. Estymator to funkcja (pomyśl szablon); ocena to realizacja funkcji, konkretna wartość liczbowa dla konkretnych wartości x oraz y

Weryfikacja modelu

Ocena części części stochastycznej modelu

Wariancja składnika losowego (wariancja resztowa)

$$s_e^2 = \sum e_i/(N-k)$$

gdzie: $e = y - \hat{y}$ jest resztą (*residual*);

N jest liczbą obserwacji w próbie

k jest liczbą zmiennych objaśniających łącznie z wyrazem wolnym ($k = 2$)

Błąd standardowy reszt (*standard error*) albo odchylenie standardowe składnika losowego to pierwiastek kwadratowy z wariancji resztowej.

Można udowodnić że s_e^2 jest **nieobciążonym estymatorem** wariancji s^2.

Żmudniejsze rachunki pozwalają wywieść wzory określające estymatory wariancji estymatorów b_1 oraz b_2, które oznaczamy jako s_{b1}^2 oraz s_{b2}^2

Można wykazać że:

$$(b_1 - \beta_1)/s_{b1} \qquad \text{oraz} \qquad (b_2 - \beta_2)/s_{b2}$$

gdzie s_{b1} oraz s_{b2} to średnie błędy szacunku parametrów mają rozkład t-Studenta o $N-k$ stopniach swobody.

Testowanie hipotez o istotności parametrów

$H_0 : \beta_2 = 0$

$H_1 : \beta_2 \neq 0$

Statystyka testu ma rozkład t z $(N-2)$ stopniami swobody

$t = b_2/s_{b2}$

Duże wartości t świadczą przeciwko H_0; małe wartości świadczą przeciwko H_1

Prościej jest patrzeć (na wydrukach komputerowych) na wartość prawdopodobieństwa zrealizowania się wartości równej t i większej

W programie Gretl jest to oznaczone jako **wartość p**

```
          współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------
  const      34404,7         12367,1          2,782      0,0077   ***
  time       174,415           413,926        0,4214     0,6753
```

Jeżeli $p < 0,05$ to H_0 należy odrzucić (oczekiwany rezultat); w powyższym przykładzie z żalem możemy powiedzieć że nie ma podstaw do odrzucenia H_0 czyli model jest do bani ponieważ jeżeli $\beta_2 = 0$ to $y = \beta_1$ czyli nie ma zależności pomiędzy x a y w sensie statystycznym.

Analiza wariancji

Reszta (dla przypomnienia) $e = y - \hat{y}$, albo $y = \hat{y} + e$

Odejmując obustronnie $\bar{y}$, podnosząc do kwadratu oraz sumując otrzymamy:

$$\sum (y - \bar{y})^2 = \sum (\hat{y} - \bar{y})^2 + \sum e^2$$

co oznaczamy jako

$$\text{OSK} = \text{WSK} + \text{RSK}$$

gdzie OSK = ogólna suma kwadratów; WSK = wyjaśniona suma kwadratów; RSK = resztowa suma kwadratów

Dzieląc obustronnie przez OSK

$$1 = \text{WSK/OSK} + \text{RSK/OSK}$$

Wielkość $R^2 = WSK/OSK$ nazywamy **współczynnikiem determinacji**;

Wielkość $\phi^2 = RSK/OSK$ nazywamy **współczynnikiem zbierzności**.

Interpretacja: % zmienności zmiennej y objaśnianej przez model (R^2)

Test F

$H_0 : \beta_2 = \beta_3 = ... = \beta_k = 0$

H_1 : co najmiej jeden współczynnik jest różny od zera

Statystyka testu

$$F = (WSK/(N-k))/(RSK/(N-k))$$

Małe wartości **p** świadczą (jak zwykle) przeciwko H_0 (na czym nam zależy); W programie Gretl wygląda to jakoś tak:

```
F(1, 49)                0,177551   Wartość p dla testu F   0,675329
```

Dla przypadku $k = 2$ test t oraz F dają ten sam wynik (co ilustruje powyższy przykład)

@liveDemo@ dane przekrojowe

Konsumpcja a dochody

Podobno Keynes, ale zależność jest w miarę oczywista:

$$K = a + b \cdot D$$

https://en.wikipedia.org/wiki/Consumption_function Consumption vs Disposable Income (Dochód rozporządzalny; https://pl.wikipedia.org/wiki/Doch%C3%B3d_rozporz%C4%85dzalny)

parametr b jest być interpretowany jako **krańcowa skłonność do konsumpcji**. Pod tą szumną nazwą kryje się udział wydatków na konsumpcję w jednostce dochodów; parametr ten powinien być w przedziale 0–1 Parametr ten ma zdroworosądkową interpretację pn ile dochodu wydamy

Weryfikacja: potrzebujemy konsumpcji i dochodów

Na poziomie indywidualnym (mikroekonomicznym) → Badania budżetów domowych

Na poziomie makroekonomicznym → Bank Danych Lokalnych GUS https://bdl.stat.gov.pl/bdl/metadane/cechy/1870 (przeciętne miesięczne wydatki na 1 osobę/przeciętny dochód rozporządzalny)

@liveDemo@ szeregi czasowe

Wybór modelu

Skorygowany współczynnik determinacji (o liczbę zmiennych objaśniających):

$$\bar{R}^2 = 1 - (1 - R^2)(N-1)/(N-k)$$

Kryteria minimum: istotne wartości parametrów modelu + większa wartość $\bar{R}^2$

Istotna uwaga: Comparing R-squares only makes sense when you don't change the dependent variable: the proportion of variance explained depends both the how much you explain and on how much variance you had to begin with. A non-linear transformation like taking the logarithm will influence the variance of your dependent variable, making the R-squares of the linear model and the log-log model incomparable.

Autokorelacja składnika losowego

Modele wykorzystujące szeregi czasowe są podatne na występowanie **autokorelacji** składnika losowego.

Do wykrywania autokorelacji można wykorzystać: wykres reszt modelu, test Durbina-Watsona oraz test Breutscha-Godfreya

Składniki losowe pozostają w zależności **autoregresyjnej** pierwszego rzędu:

$\xi_i = \rho\xi_{i-1} + \mu_i$

gdzie: μ_i składnik czysto losowy; ρ współczynnik autokorelacji $\rho = \text{cov}(\xi_i, \xi_{i-1})/(\text{var}(\xi_i)\text{var}\xi_{i-1})$

Test Durbina-Watsona (DW) weryfikuje hipotezę o nieistotności autokorelacji pierwszego rzędu:

$H_0 : \rho = 0$

$H_1 : \rho > 0$

Statystyka testu

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

jeżeli $d > 2$ dokonujemy przekształcenia $d' = 4 - d$ a hipoteza alternatywna ma postać (autokorelacja ujemna):

$H_1' : \rho < 0$

Wartość d (albo d') porównuje się z **wartościami krytycznymi** d_l oraz d_u (*upper/lower*):

Jeżeli $d < d_l$ H_0 odrzucamy; występuje autokorelacja składnika losowego

Jeżeli $d > d_u$ nie ma podstaw do odrzucenia H_0

Jeżeli $d_l < d < d_u$ nie można podjąć żadnej decyzji; wynik nierozstrzygnięty

W przypadku weryfikacji modelu regresji pożądany jest wynik $d < d_l$ oczywiście

Program Gretl dla test DW nie drukuje p.

Test Breutscha-Godfreya weryfikuje występowanie autokorelacji wyższych rzędów, przy założeniu że składnik losowy można opisać następującym równaniem:

$$\xi_t = \rho_1 \xi_{t-1} + \rho_2 \xi_{t-2} + \cdots + \rho_p \xi_{t-p} + \nu_t$$

$H_0 : \rho_1 = \rho_2 = \cdots = \rho_p = 0$

Wyniki estymacji modelu uzyskuje się uruchamiając Testy testy autokorelacji. Należy podać rząd opóźnienia p. Nieistotność oszacowanych prarametrów świadczy o niewystępowaniu autokorelacji.

Weryfikacja hipotezy jest oparta na wartość statystyki oznaczonej zwykle jako LM (a w gretlu TR^2). Gretl podaje wartość p dla tej statystyki, jeżeli jest ona mniejsza od przyjętego poziomu istotności, wykle $0,05$, to H_0 nalezy odrzucić.

Test BW ma też alternatywną statystykę testu oznaczoną w Gretlu jako LMF. Podobnie jak dla LM, odrzucamy H_0 jeżeli wartość p jest mniejsza od przyjętego poziomu istotności.

Przykładowy wydruk wygląda jakoś tak:

```
Wsp. determ. R-kwadrat = 0,871434

Statystyka testu: LMF = 6,161940,
z wartością p = P(F(11,10) > 6,16194) = 0,00382

Statystyka testu: TR^2 = 20,042993,
z wartością p = P(Chi-kwadrat(11) > 20,043) = 0,0448
```

Wysoka wartość R^2 oraz niskie wartości p dla statystyk `TR2` oraz `LMF` wskazują na występowanie autokorelacji składnika losowego...

Ocena normalności rozkładu składnika resztowego

Test weryfikuje hipotezę o normalności rozkładu resztowego

$H_0 : u \sim N$

$H_1 : u \nsim N$

Do tego celu opracowano wiele testów w tym test Jarque'a-Bery oraz Doornika-Hansena.

Niskie wartości p (*prob*) świadczą na rzecz hipotezy alternatywnej

W programie Gretl (nowe wersje) test Doornika-Hansena jest dostępny w menu `Testy/test normalności rozkładu reszt`

Ocena jednorodności wariancji składnika resztowego

Test White'a (heteroskedastyczności) weryfikuje hipotezę o występowaniu niejednorodności wariancji składnika losowego

H_0 : wariancja jest jednorodna

Niskie wartości p (*prob*) świadczą na rzecz hipotezy alternatywnej

Ocena liniowości postaci analitycznej

Czy model potęgowy lub wielomianowy daje lepsze rezultaty niż liniowy (test nieliniowości w programie Gretl)

Niskie wartości p (*prob*) świadczą na rzecz hipotezy alternatywnej

Test RESET

Czy to liniowa postać modelu (względem funkcji kwadratowej lub sześciennej) jest najlepszym możliwym do wybrania modelem.

Niskie wartości p (*prob*) świadczą na rzecz hipotezy alternatywnej

Współliniowość zmiennych objaśniających

Współliniowość zmiennych jest cechą nieporządaną: oszacowane błędy standardowe ocen paramatrów (s_b) mają duże wartości, w rezultacie b/S_b ma małe wartości a test na istotność parametru wykazuje że jest on **nieistotny**.

Do oceny współliniowości wykorzystuje się miarę VIF (**variance inflaction factor**):

$$VIF_j = 1/(1 - R_j^2)$$

gdzie R_j^2 jest współczynnikiem korelacji wielorakiej pomiędzy zmienną x_j a pozostałymi zmiennymi. Wartość $VIF_J > 10$ oznacza nadmierną współliniowość.

Zwykle taką nadmiernie współliniową zmienną należy usunąć z modelu

Obserwacje nietypowe

Gretl: Analiza → Ocena wpływowych obserwacji

@liveDemo@

Przykład: determinanty wielkości wynagrodzenia [w USA; Gujarati]

Zbiór danych `Wages.csv`

wage = female + nonwhite + union + edu + exper

gdzie:

wage = hourly wage in USD;

gender = 1 if female ;

nonwhite = 1 if nonwhite;

union = 1 if in union ;

edu = education in yrs ;

exper = experience (staż pracy)

@liveDemo@

Nieliniowe postacie analityczne

Regresja potęgowa

$$Y = \beta_1 X_2^{\beta_2} X_3^{\beta_3} ...$$

Cobb-Douglas production function

$$P = \beta_1 L^{\beta_2} K^{\beta_3}$$

gdzie:

P – wielkość produkcji;

L – nakłady pracy (*labour input*);

K – nakłady kapitałowe (*capital*)

Po obustronnym zlogarytmowaniu otrzymujemy:

$$\ln P = \ln \beta_1 + \beta_2 \ln L + \beta_3 \ln K$$

W modelu potęgowym współczynniki nachylenia są interpretowane jako **elastyczności**: Elastyczność to % zmiana jednej zmiennej podzielona przez % zmianę drugiej zmiennej. Albo: procentowa zmiana zmiennej x skutkuje β% zmianą zmiennej y

W modelu C-D mamy dwie zmienne objaśniające więc mówimy o cząstkowych elastycznościach (*partial elasticities*): β_2 współczynnik elastyczności produkcji ze względu na nakłady pracy, przy założeniu stałych nakładów kapitałowych. Albo: procentowa zmiana produkcji przy jednoprocentowej zmianie w nakładach pracy *ceteris paribus*

W modelu C-D suma współczynników nachylenia jest interpretowania jako **efekt skali produkcji** (*returns to scale*): jeżeli ESP = 1 to stała skala; jeżeli EPS $<$ 1 malejąca skala; jeżeli EPS $>$ 1 rosnąca skala.

Przykład Cobb-Douglas function for USA 2005 [Gujarati]

@liveDemo@ `CDdata.csv`

Przykład krzywa Engla dla światowej konsumpcji mięsa

consumption is under-proportional with income. This can be formally descibed by the power function with parameter b_1 which can be interpreted as an elasticity

$$M = b_0 \mathrm{GDP}^{b_1}$$

@liveDemo@ `meat_and_gni.csv`

Regresja wykładnicza

Najczęściej stosowana jako model tendencji rozwojowej (trendu), patrz przykład.

Jeżeli

$$Y = a_1 a_2^{x_2} a_3^{x_3} ...$$

to po wykonaniu niezbędnych przekształceń

$$\ln Y = \ln a_1 + \ln a_2 x_2 + \ln a_3 x_3 ...$$

Modele trendu

$$Y = a_1 + a_2 t + u$$

Współczynnik a_2 jest interpretowany jako przeciętna zmiana y na jednostkę czasu. Jeżeli jednostką t jest rok to a_2 oznacza przeciętną zmianę roczną

Przykład PKB realny (*real GDP*) dla USA 1960–2007 [Gujarati]

Model wzrostu

$$Y = a_1 a_2^t$$

lub po wykonaniu niezbędnych przekształceń

$$\ln Y = \ln a_1 + \ln a_2 t + u$$

Przykład

Formula of Compound Amount (procent składany):

Kwota końcowa = kwota początkowa × (1+wysokość oprocentowania) ^ liczba okresów

$$GDP_t = GDP_0(1 + r)^t$$

po obustronnym zlogarytmowaniu:

$$\ln GDP_t = \ln GDP_0 + t \ln(1 + r)$$

podstawiając: $B_1 = \ln GDP_0$ oraz $B_2 = \ln(1 + r)$, otrzymujemy ostatecznie:

$$\ln GDP_t = B_1 + B_2 t + u$$

Współczynnik B_2 pomnożony przez 100 jest interpretowany następująco: jednostkowa zmiana x skutkuje $B_2 \times 100\%$ zmianą y.

Roczna przeciętna stopa wzrostu wynosi $B_2 \times 100\%$.

$B_2 \times 100\%$ jest także określany jako półelastyczność (*semielasticity*)

Przykład PKB realny (*real GDP*) dla USA 1960–2007 [Gujarati]

@liveDemo@ `GDP_US.csv`

W okresie 1960–2007 realny GDP USA wzrastał o 3,15% rocznie. Ta stopa wzrostu jest statystycznie istotna.

An R-square comparison is meaningful only if the dependent variable is the same for both models. So the R-square from the linear model cannot be compared with the R-square from the log-log model.

For the log-log model, the way to proceed is to obtain the antilog predicted values and compute the R-square between the antilog of the observed and predicted values. This R-square can then be compared with the R-square obtained from OLS estimation of the linear model

Regresja wielomianowa

EKC

Teorię/hipoteza stworzoną przez Simona Kuznetsa, zgodnie z którą wraz z postępującym rozwojem kraju najpierw nierówności społeczne rosną, a potem spadają. Odwrócone U.

Środowiskowa krzywa Kuznetsa (EKC) postuluje podobną zależność pomiędzy rozwojem a degradacją środowiska naturalnego.

$$CO^2 = a_0 + a_1 PKB + a_2 PKB^2$$

Przykład: konsumpcja mięsa na świecie

$$M = a_0 + a_1 PKB + a_2 PKB^2$$

Aby potwierdzić hipotezę EKC $a_2 < 0$

Można obliczyć wielkość PKB, dla której M osiągnie wartość maksimum; wynosi ona $PKB_{\max} = a_1/(-2 \cdot a_2)$

ta maksymalna wartość M wynosi: $a_0 + a_1 * PKB_{\max} + a_2 * PKB^2_{\max}$

@liveDemo@ `meat_and_gni.csv`

Predykcja ekonometryczna

Uwagi wstępne

Przewidywanie co będzie. Anglicy odróżniają *forecasting* (**prognozowanie**) or *predicting* (**przewidywanie**); w sumie obie rzeczy dotyczą przeszłości ale istnieje subtelna różnica:

jaka będzie wielkość GDP w przyszłym roku (**forecasting**)

jaka będzie wielkość produkcji jeżeli zwiększymy nakłady pracy o x jednostek (**predicting**)

Prognozowanie w modelu regresji

Załóżmy, że:

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x3i + ... + \beta_k x_{ki} + u_i$$

prognozowana wartość:

$$y^* = b_1 + b_2 x_2^* + b_3 x3^* + ... + b_k x_k^* + u^*$$

gdzie: $x_2^* ... x_k^*$ jest wektorem wartości przyjmowanych przez zmienne objaśniające, a u^* jest składnikiem losowym **o takim samym rozkładzie jak składniki losowe z próby** ($N(0, \sigma)$)

Prognozy dokonujemy obliczając

$$\widehat{y}^* = \widehat{b}_1 + \widehat{b}_2 x_2^* + \widehat{b}_3 x3^* + ... + \widehat{b}_k x_k^*$$

gdzie wartości $\widehat{b}_1, ..., \widehat{b}_k$ oznaczają oceny parametrów

$\widehat{y}^*$ nazywamy prognozą punktową (w odróżnieniu od przedziałowej o czym za chwilę)

Wówczas błąd prognozy wynosi:

$$e^* = y^* - \widehat{y}^*$$

Oczywiście wartość y^* jest nieznana, ale można udowodnić, że

$$E(e^*) = 0$$

(średnia wartość błędu wynosi zero)

oraz, że wariancja jest równa:

$$\mathrm{var}(e^*) = \mathrm{var}(\widehat{y}^*) + \mathrm{var}(u) = \mathrm{var}(\widehat{y}^*) + \sigma$$

gdzie $\sigma = \text{var}(u)$ (dla zwięzłości)

Wariancja błędu prognozy jest więc **większa od wariancji składnika losowego**. Prognozujemy z błędem $\text{var}(\hat{y}^*)$ plus zjawisko ma charakter losowy także w przyszłości $\sigma = \sqrt{\text{var}(u)}$. Te dwa rodzaje błędów się sumują…

Ponieważ z założenia u^* jest składnikiem losowym **o takim samym rozkładzie jak składniki losowe z próby** u^*, zatem σjest jedynym nieznanym parametrem (inaczej mówiąc $\text{var}(\hat{y}^*)$ jest **jakąś** funkcją σ; jaką można to analitycznie ustalić ale nie będziemy tego robić)

Wstawiając zamiast σ jego estymator s_e^2 (zwany **błędem standardowy reszt** albo **odchyleniem standardowym składnika losowego** – patrz wyżej) otrzymujemy nieobciążony estymator wariancji błędu prognozy $\text{var}(e^*) = s^{*2}$

Pierwiastek kwadratowy z wariancji błędu prognozy s^* nosi nazwę **średniego błędu predykcji**.

Iterpretacja: ile średnio wyznaczona prognoza może się odchylać od wartości rzeczywistych zmiennej prognozowanej.

Ponieważ u ma rozkład $N(0, \sigma)$, to m.in.duże odchylenia są mniej prawdopodobne niż małe można zatem wyznaczyć dwie wartości:

$$\hat{y}_l^* < y < \hat{y}_u^*$$

pomiędzy którymi – z zadaną dokładnością – znajduje się prognozowana wartość. Ta zadana dokładość jest określona prawdopodobieństwem zwykle 0,95 lub 0,99.

Taka konstrukcja nazywa się **prognozą przedziałową**.

Wniosek: jeżeli model jest słabo dopasowany to prognoza jest jeszcze gorsza a prognoza przedziałowa może być mało przydatna, typu 20 plus/minus 30.

Miary dokładności predycji

Niewątpliwie **średniego błędu predykcji** jest miarą jej dokładności. Im mniejszy tym lepiej.

Średni błąd predykcji należy do miar dokładności predykcji **ex-ante**, tj. wyliczanych bez znajomości prawdziwych wartości (prognoza się jeszcze nie zrealizowała)

Oprócz miary **ex-ante** są miary **ex-post** tj. takie w których prognozy porównuje się z wartościami zrealizowanymi.

Żeby te wartości zrealizowane mieć to albo trzeba poczekać :-) Albo dokonać następującego tricku: dzielimy dane na dwie części zwane zwykle zbiorem uczącym (*training set*) i zbiorem testowym (*test set*)

Na podstawie **zbioru uczącego** szacujemy model.

Na podstawie **zbioru testowego** sprawdzamy właściwości predyktywne modelu.

Ponieważ dysponujemy prognozami oraz realizacjami to możliwe jest oszacowanie jakości prognoz. Są to tego stosowane następujące miary:

Średniokwadratowy błąd prognozy (albo średni błąd kwadratowy; MSE z angielska tj *mean square error*)

$$\text{MSE} = \frac{1}{n^*} \cdot \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Pierwiastek MSE zwany RMSE (root MSE):

$$\text{RMSE} = \sqrt{\text{MSE}}$$

Średni błąd bezwzględny (*Mean Absolute Error*; MAD):

$$\text{MAD} = \frac{1}{n^*} \cdot \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

Interpretacja: ile średnio wyznaczona prognoza odchylała się od wartości rzeczywistych zmiennej prognozowanej (było **może się odchylać** dla miar ex-ante)

@liveDemo@

Koniec