MIERNIKI DOPASOWANIA

MODELU REGRESJI LINIOWEJ DO DANYCH

Łukasz Kwiatkowski

Katedra Ekonometrii i Badań Operacyjnych

Plan wykładu

1) Mierniki dopasowania – wykaz i nazewnictwo
2) Odchylenie resztowe
3) Współczynnik zmienności resztowej
4) Współczynnik zbieżności
5) Współczynnik determinacji
6) Przykłady
7) Uwagi końcowe

Myśl przewodnia

→ Jak ocenić to, na ile dobrze model regresji jest dopasowany do danych (na podstawie których został oszacowany). Jak to zmierzyć? Za pomocą jakich mierników? Jak je zinterpretować? Jakie mają wady (ograniczenia)?

Mierniki dopasowania – wykaz i nazewnictwo

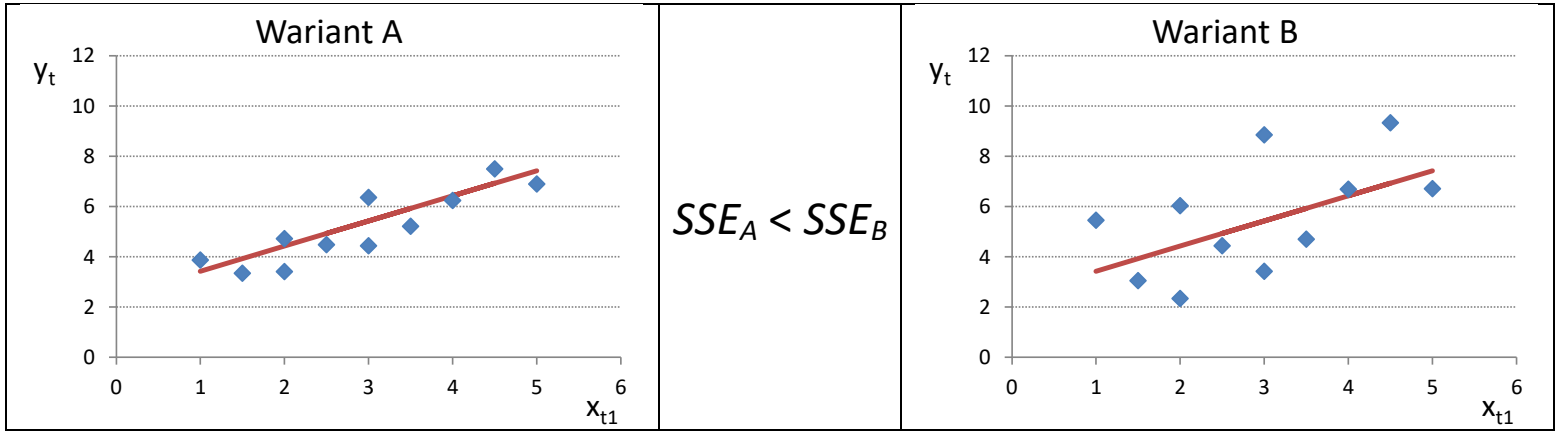
➢ Cztery podstawowe mierniki:

- s – odchylenie resztowe (ang. *residual standard deviation*)
- V_ε – współczynnik zmienności resztowej (ang. *residual coefficient of variation*)
- φ^2 – współczynnik zbieżności (ang. *fraction of variance unexplained, FVU*)
- R^2 – współczynnik determinacji (ang. coefficient of determination)

→ Wszystkie w zasadzie bazują w jakiejś mierze na **sumie kwadratów reszt**, czyli

$$SSE = \underbrace{\sum_{t=1}^{T} \hat{\varepsilon}_t^2}_{zapis\ skalarny} = \underbrace{\hat{\varepsilon}'\hat{\varepsilon}}_{zapis\ macierzowy},$$

gdzie: "SSE" jest skrótem anglojęzycznego zwrotu *Sum of Squared Errors*, $\hat{\varepsilon} = y - \hat{y}$ jest *T*-elementową kolumną **reszt**, $\hat{y} = X\hat{\beta}$ jest *T*-elementową kolumną wartości **wartości teoretycznych**. Nieformalnie (ale za to bardzo intuicyjnie :) rzecz ujmując: im mniejsza SSE (które mierzy sumaryczną odległość punktów danych od dopasowanej prostej), tym lepsze dopasowanie:



!!! Bardzo ważna uwaga !!!

Ogólnie, zmienną objaśnianą (a ściślej, jej pojedynczą wartość o nr t) oznaczamy symbolem y_t (to nie nowość). ALE we wszystkich poniższych rozważaniach musimy pamiętać, że pod terminem / oznaczeniem *zmiennej objaśnianej* możemy rozumieć – w zależności od specyfikacji modelu – pewną zmienną ekonomiczną (np. przychody ze sprzedaży, np. w tysiącach zł) ALBO jej *logarytm*. Jeśli wspomniane przychody ze sprzedaży oznaczymy jako S_t (ang. sales), wówczas:

- W pierwszym przypadku (tj. gdy y_t jest bezpośrednio zmienną ekonomiczną) zapiszemy $y_t \equiv S_t$
  → tym samym y_t jest wyrażony w tej samej co S_t jednostce, czyli w tysiącach zł.

- W drugim przypadku (tj. gdy y_t jest *logarytmem* zmiennej ekonomicznej) zapiszemy $y_t \equiv \ln S_t$
  → jednostką y_t jest wtedy *logarytm* jednostki zmiennej S_t, czyli *logarytm* tysiąca zł (jakkolwiek dziwnie to brzmi :)

Podsumowując, "samo" y_t oznacza zawsze *zmienną objaśnianą* (i zarazem wartości tejże), a tą ostatnią może być pewna zmienna ekonomiczna lub jej logarytm. W kontekście oceny dopasowania modelu regresji do danych ma to o tyle znaczenie, że w tym drugim przypadku oceniamy, w istocie, dopasowanie modelu regresji do wartości będących *logarytmem* oryginalnej zmiennej ekonomicznej (a nie do wartości samej zmiennej ekonomicznej). Będzie to musiało być uwzględnione na etapie interpretacji wartości obliczanych mierników, o czym później...

Uwaga przy okazji: zapis "$A \equiv B$" (relacja równoważności) należy czytać: "*A* jest tym samym, co *B*"

Mierniki dopasowania: odchylenie resztowe, *s*

➢ Formuła: $s = \sqrt{s^2}$,

gdzie s^2 – wariancja resztowa (ang. *residual variance*): $s^2 = \frac{1}{T-k} SSE$

➢ Uwagi:

1) s^2 i s niosą, w zasadzie, tę samą informację – o **wielkości rozproszenia reszt**, czyli rozrzucie wartości empirycznych (in.: zaobserwowanych) zmiennej objaśnianej (y_t) wokół dopasowanej płaszczyzny MNK (w przypadku tylko $K = 1$ zmiennej objaśniającej: płaszczyzna = prosta)
2) s^2 nie posiada jednak interpretacji, gdyż wyrażona jest w *kwadracie* jednostki zm. objaśnianej
3) s – zauważmy – jest wyrażone w tej samej jednostce co y_t
4) s – to **przeciętna różnica** (wyrażona w jednostkach zmiennej objaśnianej) **pomiędzy zaobserwowanymi** (in.: empirycznymi, y_t) **a teoretycznymi** (wynikającymi z oszacowanego modelu, $\hat{y}_t$) **wartościami zmiennej objaśnianej**
5) Model jest tym **lepiej** dopasowany do danych, im **niższa** wartość s; ponadto, $s > 0$ (zawsze)
6) s jest wielkością mianowaną (patrz uwaga nr 3), więc nie wiemy jednak, czy np. $s = 1000$ zł to dużo czy mało $\Rightarrow$ chcemy "uwolnić" s od jednostki, poprzez rozważenie jakiegoś *relatywnego* miernika – takiego, w którym s byłoby odniesione do jakiejś "typowej" wartości zmiennej objaśnianej, również wyrażonej w tej samej co s jednostce, przez to jednostki obydwu wielkości się skrócą. Zwykle za tą "typową", najbardziej reprezentatywną wartość zmiennej objaśnianej przyjmuje się średnią arytmetyczną jej wartości empirycznych, tj. $\bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t$. Stąd kolejny miernik: współczynnik zmienności resztowej – patrz kolejny slajd.

Mierniki dopasowania: współczynnik zmienności resztowej, V_ε

- Formuła: $V_\varepsilon = \frac{s}{|\bar{y}|} \cdot 100,$

  $\bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t$ (średnia arytmetyczna z wartości zmiennej objaśnianej – czy jest nią pewna zmienna ekonomiczna czy też jej logarytm); moduł – w celu uniknięcia jej potencjalnie ujemnych wartości

- Uwagi:

  1) V_ε jest wielkością niemianowaną, wyrażoną w procentach (dzięki czynnikowi "100" we wzorze), choć nieograniczoną od góry wartością 100% (s może być większe od $|\bar{y}|$)
  2) V_ε – zgodnie ze swoim wzorem – informuje o tym, ile procent średniej arytmetycznej zaobserwowanych wartości zmiennej objaśnianej stanowi odchylenie resztowe; zwykle jego interpretację "dołącza się" do tej formułowanej dla odchylenia resztowego, jako jej kontynuację, w tym samym zdaniu (przykłady znajdziesz pod koniec tego materiału)
  3) Model jest tym **lepiej** dopasowany do danych, im **niższa** wartość V_ε; ponadto, $V_\varepsilon > 0$ (zawsze)
  4) Przyjmuje się – choć czysto arbitralnie(!) – następującą skalę pozwalającą określić jakość dopasowania modelu do danych za pomocą V_ε:
     - $V_\varepsilon \leq 10\%$ → model bardzo dobrze dopasowany do danych
     - $10\% < V_\varepsilon \leq 20\%$ → model dobrze dopasowany do danych
     - $20\% < V_\varepsilon \leq 30\%$ → model przeciętnie dopasowany do danych
     - $30\% < V_\varepsilon$ → model słabo dopasowany do danych

  Powtórzmy jednak, że poszczególne progi w tej skali są tylko umowne (arbitralne)…

Mierniki dopasowania: współczynnik zbieżności, φ^2; współczynnik determinacji, R^2

➢ Wprowadzenie: Jeżeli w modelu regresji liniowej występuje wyraz wolny (czyli praktycznie zawsze[1]), wówczas można dowieść prawdziwości następującej tożsamości:

$$\underbrace{\sum_{t=1}^{T}(y_t-\bar{y})^2}_{\substack{całkowita\ zmienność\\ zm.\ objaśnianej}} = \underbrace{\sum_{t=1}^{T}(\hat{y}_t-\bar{y})^2}_{\substack{zmienność\ wyjaśniona\\ za\ pomocą\ modelu}} + \underbrace{\sum_{t=1}^{T}(y_t-\hat{y}_t)^2}_{\substack{zmienność\ NIEwyjaśniona\\ za\ pomocą\ modelu}}, \qquad (*)$$

przy czym:

- $\sum_{t=1}^{T}(y_t-\bar{y})^2$ – utożsamia całkowitą zmienność empirycznych wartości zm. objaśnianej (którą to zmienność chcemy wyjaśnić, wychwycić za pomocą modelu regresji); wyrażenie to – zauważmy – jest "trzonem" wzoru na wariancję (podstawowa miara zmienności/rozrzutu) z próby zmiennej y_t
- $\sum_{t=1}^{T}(\hat{y}_t-\bar{y})^2$ – utożsamia całkowitą zmienność / rozrzut wartości teoretycznych zm. objaśnianej wokół średniej arytmetycznej wartości empirycznych; można zatem powiedzieć, że jest to ta część (choć jeszcze nie w postaci ułamku) zmienności wartości empirycznych zm. objaśnianej, która została "wychwycona"/wyjaśniona w ramach oszacowanego modelu
- $\sum_{t=1}^{T}(y_t-\hat{y}_t)^2 = (zauważmy) = \sum_{t=1}^{T}\hat{\varepsilon}_t^2 = SSE$ – czyli suma kwadratów reszt; utożsamia całkowitą zmienność reszt, czyli różnic pomiędzy empirycznymi a teoretycznymi wartościami zm. objaśnianej; można zatem powiedzieć, że jest to ta część (choć jeszcze nie w postaci ułamku) zmienności wartości empirycznych zm. objaśnianej, która NIE została "wychwycona"/wyjaśniona w ramach oszacowanego modelu

[1] Wyraz wolny praktycznie zawsze powinien być obecny w modelu regresji liniowej. Istnieje po temu kilka powodów, m.in. ten, że pominięcie wyrazu wolnego może skutkować obciążeniem („nietrafnością") ocen pozostałych parametrów.

Mierniki dopasowania: współczynnik zbieżności, φ^2; współczynnik determinacji, R^2

- Dzieląc obydwie strony tożsamości (*) (patrz poprzednia strona) przez wyrażenie po lewej stronie, otrzymujemy:

$$1 = \underbrace{\frac{\sum_{t=1}^{T}(\hat{y}_t - \bar{y})^2}{\sum_{t=1}^{T}(y_t - \bar{y})^2}}_{R^2} + \underbrace{\frac{\sum_{t=1}^{T}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{T}(y_t - \bar{y})^2}}_{\varphi^2}$$

- Zatem: $R^2 + \varphi^2 = 1$ i $R^2 \in [0,1], \quad \varphi^2 \in [0,1]$

- R^2 – informuje, jaka część (w %, po przemnożeniu przez czynnik 100) zaobserwowanej zmienności zmiennej objaśnianej (czyli zmienności jej wartości empirycznych) została wyjaśniona w ramach oszacowanego modelu

  $\rightarrow$ Im **wyższa** (bliższa 1) wartość R^2, tym model jest **lepiej** dopasowany do danych

- φ^2 – informuje, jaka część (w %, po przemnożeniu przez czynnik 100) zaobserwowanej zmienności zmiennej objaśnianej (czyli zmienności jej wartości empirycznych) **NIE** została wyjaśniona w ramach oszacowanego modelu

  $\rightarrow$ Im **niższa** (bliższa 0) wartość φ^2, tym model **lepiej** dopasowany do danych

Mierniki dopasowania: przykłady

Przykład 1. Rozważmy pewien model:

$$E_t = -34{,}2 + 7{,}1I_t + 27{,}3H_t + \hat{\varepsilon}_t,$$

oszacowany na podstawie 10 gospodarstw domowych, celem modelowania wielkości miesięcznych ich wydatków na transport (zmienna E_t, w zł/osobę w g.d.) w zależności od wielkości miesięcznych dochodów gospodarstwa domowego (zmienna I_t, w setkach zł / osobę w g.d.) i typu gospodarstwa: zmienna H_t typu 0-1, przy czym $H_t = 1$ dla gospodarstw pracowniczych, zaś $H_t = 0$ dla gospodarstw emerytów i rencistów. Celem oceny dobroci dopasowania rozważanego modelu do danych obliczono wartości stosownych mierników i otrzymano wyniki: $s = 15{,}29$ (← Pytanie: jaka jest jednostka?); $V_\varepsilon = 5{,}67\%$; $R^2 = 0{,}78$; $\varphi^2 = 0{,}22$ (Przypomnijmy: zachodzi $R^2 + \varphi^2 = 1$, gdyż w modelu występuje wyraz wolny – jego ocena to $\hat{\beta}_0 = -34{,}2$).

Interpretacje: (bierzemy tu pod uwagę konkretne znaczenie zmiennej objaśnianej, którą jest tu $y_t \equiv E_t$)

- $s = 15{,}29$; $V_\varepsilon = 5{,}67\%$ (obydwie wartości interpretujemy razem, w jednym zdaniu): Zaobserwowane wartości miesięcznych wydatków gospodarstw domowych na transport różnią się od ich wartości teoretycznych przeciętnie o ±15,29 zł/os. w g.d. [do tego miejsca mamy interpretację s; dalsza część zdania dotyczy V_ε], co stanowi 5,67% średniej arytmetycznej zaobserwowanych wartości tych wydatków.

  Ponadto, z uwagi na to, że $V_\varepsilon \leq 10\%$ możemy stwierdzić, że rozważany model charakteryzuje się bardzo dobrym dopasowaniem do danych.

Mierniki dopasowania: przykłady interpretacji

Przykład 1. - c.d.

- $R^2 = 0{,}78$ (w dwóch równoważnych brzmieniach – do wyboru):

  → Wersja 1: Całkowita zmienność wielkości miesięcznych wydatków gospodarstw domowych na transport została wyjaśniona w ramach oszacowanego modelu w 78%.

  → Wersja 2: 78% całkowitej zmienności wielkości miesięcznych wydatków gospodarstw domowych na transport zostało wyjaśnione w ramach oszacowanego modelu.

- $\varphi^2 = 1 - R^2 = 0{,}22$ (analogicznie jak współczynnik determinacji, tyle tylko, że dopisując „nie”…):

  → Wersja 1: Całkowita zmienność wielkości miesięcznych wydatków gospodarstw domowych na transport NIE została wyjaśniona w ramach oszacowanego modelu w 22%.

  → Wersja 2: 22% całkowitej zmienności wielkości miesięcznych wydatków gospodarstw domowych na transport NIE zostało wyjaśnione w ramach oszacowanego modelu.

Mierniki dopasowania: przykłady interpretacji

Przykład 2. Przykład ten stanowi subtelną (acz znaczącą :) modyfikację poprzedniego, przy czym psikus polega na tym, że tu zmienną objaśnianą jest *logarytm* owej wielkości miesięcznych wydatków gospodarstw domowych na transport:

$$\ln E_t = -34{,}2 + 7{,}1 I_t + 27{,}3 H_t + \hat{\varepsilon}_t.$$

Załóżmy, że wszystkie wyniki pozostają bez zmian (Pytanie: w jakiej jednostce jest teraz $s = 15{,}29$?). Jak logarytmiczna postać zmiennej objaśnianej ($y_t \equiv \ln E_t$) rzutuje na brzmienie poszczególnych interpretacji?

Interpretacje:

- $s = 15{,}29$; $V_\varepsilon = 5{,}67\%$ (interpretujemy razem, w jednym zdaniu):
  Zaobserwowane wartości logarytmu miesięcznych wydatków gospodarstw domowych na transport różnią się od ich wartości teoretycznych przeciętnie o ±15,29 logarytmu zł/os. w g.d. [do tego miejsca mamy interpretację s; dalsza część zdania dotyczy V_ε], co stanowi 5,67% średniej arytmetycznej zaobserwowanych wartości logarytmu tych wydatków.

  Ponadto, z uwagi na to, że $V_\varepsilon \leq 10\%$ możemy stwierdzić, że rozważany model charakteryzuje się bardzo dobrym dopasowaniem.

Mierniki dopasowania: przykłady interpretacji

Przykład 2. - c.d.

- $R^2 = 0{,}78$ (w dwóch równoważnych brzmieniach – do wyboru):

  → Wersja 1: Całkowita zmienność logarytmu wielkości miesięcznych wydatków gospodarstw domowych na transport została wyjaśniona w ramach oszacowanego modelu w 78%.

  → Wersja 2: 78% całkowitej zmienności logarytmu wielkości miesięcznych wydatków gospodarstw domowych na transport zostało wyjaśnione w ramach oszacowanego modelu.

- $\varphi^2 = 1 - R^2 = 0{,}22$ (analogicznie jak współczynnik determinacji, tyle tylko, że dopisując „nie”…):

  → Wersja 1: Całkowita zmienność logarytmu wielkości miesięcznych wydatków gospodarstw domowych na transport NIE została wyjaśniona w ramach oszacowanego modelu w 22%.

  → Wersja 2: 22% całkowitej zmienności logarytmu wielkości miesięcznych wydatków gospodarstw domowych na transport NIE zostało wyjaśnione w ramach oszacowanego modelu.

Uwagi końcowe, czyli zauważmy jednak, że...

- Wszystkie cztery mierniki są oparte na sumie kwadratów reszt (SSE) – im jest ona niższa, tym na lepsze dopasowanie modelu one wskazują (niższe wartości przyjmują: s, V_ε i φ^2, zaś R^2 – wyższą), i jest to jak najbardziej zgodne z intuicją (patrz wykresy z początku prezentacji). Pozornie wszystko jest z tym w porządku, poza jednym szczegółem... Otóż można pokazać, że włączając do modelu regresji kolejną, dowolną zmienną objaśniającą (nawet nie mającą totalnie żadnego związku z modelowanym zjawiskiem) suma kwadratów reszt zmaleje (mniej lub bardziej, ALE zmaleje). Co to oznacza w praktyce? Oznacza to, że mierniki te – same w sobie – nie nadają się do rozwiązania innego, ważnego problemu w modelach regresji: doboru zmiennych objaśniających do modelu, ponieważ dodając nawet "byle jakie" zmienne mierniki te zawsze będą wskazywały na poprawę dopasowania... (Nie ich wina :)) Zatem można by "ponawrzucać wszelkiego barachła" do modelu tylko po to, by uzyskać R^2 praktycznie równe 1 (co oczywiście byłoby szczytem statystycznej przewrotności ;), tyle tylko, że ten model nie dawałby się do niczego, ponieważ:

  a) ponieważ wysoka liczba parametrów podlegających estymacji skutkuje znaczącym wzrostem jej niepewności (ta sama liczba obserwacji przypadająca na estymację większej liczby parametrów z reguły skutkuje wzrostem wartości błędów średnich szacunku); z kolei duża niepewność estymacji parametrów przekłada się na dużą niepewność wnioskowania o zależnościach pomiędzy zm. objaśnianą a objaśniającymi, jak i dużą niepewność prognoz uzyskanych za pomocą takiego modelu

  b) po drugie, duża liczba parametrów przyczynia się do zwiększenia "elastyczności" modelu (trochę jak w przypadku funkcji wielomianowych: im wyższy stopień wielomianu, tym bardziej "giętki" wykres funkcji). W przypadku "nadmiernej elastyczności" modelu, może się okazać, że dopasowujemy się nim do zakłóceń losowych, do tego, co raczej przypadkowe, co przynależy do składnika losowego. To tzw. problem nadmiernego dopasowania / przeuczenia / przetrenowania modelu (ang. *model overfitting*).

Uwagi końcowe, czyli zauważmy jednak, że...

- Są na to pewne remedia – pewne alternatywne mierniki, w których na ogólne dopasowanie (jakość) modelu składają się dwa elementy: *stricte* dopasowanie modelu do danych (mierzone np. właśnie sumą kwadratów reszt), i dodatkowo także "cena" uzyskania tego dopasowania, rosnąca wraz z liczbą szacowanych parametrów. Ta "cena" ("kara") za każdy parametr modelu jest tak ujęta w mierniku, że jeśli wprowadzenie dodatkowej zmiennej objaśniającej do modelu (zawsze skutkujące kolejnym parametrem) nie prowadzi do stosownie wysokiego spadku sumy kwadratów reszt (takiego, który przeważyłby "cenę" za jej wprowadzenie), wówczas mierniki te wskażą ogólne pogorszenie (a nie polepszenie) dopasowania.
- Jednym z takich mierników jest tzw. *skorygowany* współczynnik determinacji (ang. *adjusted* R^2), zwykle oznaczany jako R^2_{adj} lub $\bar{R}^2$, i idefiniowany wzorem:

$$R^2_{adj} = 1 - \underbrace{\frac{\sum_{t=1}^{T}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{T}(y_t - \bar{y})^2}}_{\varphi^2} \cdot \frac{T-1}{T-k} = 1 - (1 - R^2) \cdot \frac{T-1}{T-k}$$

(UWAGA: ostatni czynnik – ten z "n-ami" – przemnaża tylko φ^2, a nie całe wyrażenie $1 - \varphi^2$). W odróżnieniu od "klasycznych" mierników, skorygowany R^2 może być wykorzystywany do wyboru modelu spośród specyfikacji różniących się liczbą zmiennych objaśniających. Jego "wadą" jest jednak to, że może przyjmować wartości ujemne (czyli nie jest unormowany tak, jak "zwykły" R^2), więc raczej nie poddaje się interpretacji.

- Obecnie najszerzej stosowaną grupą mierników wykorzystywanych w problemie wyboru zmiennych objaśniających do modelu są tzw. *kryteria informacyjne*, np. kryterium Akaike, kryterium Schwarza (zwane także bayesowskim kryterium informacyjnym), i in. Omówimy je przy innej okazji...