Iwona Mróz,

Instytut Fizyki Doświadczalnej,

Uniwersytet Wrocławski

**- DODATEK – wprowadzenie do metod regresyjnych**

Niniejsze materiały mają charakter roboczy. Bardzo proszę o zgłaszanie zauważonych błędów, braków, niedociągnięć i niejasności. Prośba dotyczy też przypisów. Z góry dziękuję za pomoc☺.

*Tekst opracowano na podstawie podręcznika Andrzeja Stanisza Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny, StatSoft, Kraków 2007, t. 2, s. 21-45 i 99-101 .*

**Wprowadzenie do metod regresyjnych – wstęp**

*UWAGA!!! Tekst nie wyczerpuje tematu dotykając jedynie najważniejszych zagadnień. Osoby zainteresowane zachęcam do zgłębiania literatury przedmiotu.*

Metody regresyjne to szeroka grupa metod pozwalających na ocenę powiązania pewnej cechy statystycznej opisanej przy pomocy zmiennej losowej z innymi cechami, przy czym brana jest pod uwagę siła powiązania, jego kształt i kierunek[[1]](#footnote-1). Zależność pomiędzy zmiennymi losowymi może mieć charakter funkcyjny lub statystyczny[[2]](#footnote-2). W przypadku zależności funkcyjnych, „formalnym zapisem tego wpływu są funkcje regresji, które określają sposób przyporządkowania wartości zmiennej zależnej określonym wartościom zmiennej niezależnej.”[[3]](#footnote-3) Możemy spotkać się z sytuacjami, gdy zmienna zależna, jedno- lub wielowymiarowa zależy od jednej lub wielu zmiennych niezależnych[[4]](#footnote-4).

W praktyce naszym zadaniem jest opracowanie modelu takiej zależności, która może być liniowa lub nieliniowa. Gdy zależność między badanymi zmiennymi ma charakter liniowy, najczęściej używamy tzw. regresji prostej (gdy jednowymiarowa zmienna zależna zależy od jednej zmiennej niezależnej) lub regresji wielorakiej (gdy jednowymiarowa zmienna zależna zależy od wielu zmiennych niezależnych). W praktyce budujemy model w oparciu o dane eksperymentalne będące zbiorem par wartości zmiennych niezależnych i odpowiadającej im wartości zmiennej zależnej. Na podstawie tych wartości wyznaczamy najlepszą postać funkcji opisującą zależność między zmiennymi. Funkcję tę nazywamy funkcją regresji II-go rodzaju[[5]](#footnote-5). **A Stanisz wyróżnia cztery etapy budowy modelu funkcji regresji II-go rodzaju: specyfikację modelu, estymację parametrów modelu, weryfikację modelu oraz użycie modelu do prognozowania[[6]](#footnote-6).**

Specyfikacja modelu polega na zaproponowaniu zależności między zmiennymi wynikających z postawionych hipotez badawczych lub znanych teorii. Liczbowe wartości parametrów modelu wyznaczamy na podstawie danych pomiarowych, na etapie estymacji parametrów modelu. Weryfikacja modelu wymaga sprawdzenia, czy model nie stoi w sprzeczności z teorią oraz czy jest poprawny od strony statystycznej (zob. poniżej). Model uznany za poprawmy może zostać wykorzystany do prognozowania, czyli przewidywania jaka wartość przyjmie zmienna zależna jeżeli zmienna niezależna przyjmie określoną wartość.[[7]](#footnote-7)

Z punktu widzenia niniejszego kursu najważniejsze jest wyznaczenie parametrów modelu oraz statystyczna weryfikacja modelu. Parametry modelu wyznaczamy na podstawie próby. Najczęściej w trakcie studiów będą się Państwo spotykać z wyznaczaniem parametrów modelu przy pomocy metody najmniejszych kwadratów (MNK). Nie jest to jednak metoda jedyna: warto wspomnieć chociażby o metodzie największej wiarygodności, która jest wykorzystywana m.in. do wyznaczania parametrów modeli regresji logistycznej. Aby parametry wyznaczone przy pomocy danej metody spełniały swoje zadanie, muszą być spełnione różne założenia, które dla metody MNK omówimy szczegółowo później (zob. materiały do wykładu nr 14).

Na czym polega filozofia wyznaczania parametrów modelu:

Wyobraźmy sobie, że w rzeczywistości cecha statystyczna opisana zmienną losową Y zależy liniowo od zmiennej losowej X. Nie wiemy jednak, że tak jest i nie znamy postaci związku X i Y. Dysponujemy próbą, która ma postać par pomiarów . Widzimy, ze na płaszczyźnie X0Y punkty pomiarowe układają się tak, ze może to sugerować zależność liniową między zmiennymi.

Wiemy, że ogólna postać modeli regresji liniowej prostej przybiera postać:

(10a.1),

gdzie i to odpowiednio wyraz wolny (*ang. intercept*) i współczynnik określający związek liniowy między X i Y dla populacji, a *ε* oznacza składnik losowy[[8]](#footnote-8).

Idealnie byłoby poznać wartości i , ale nie jest to możliwe i musimy zadowolić się estymatorami tych wartości określonymi na podstawie próby. Korzystając np. z MNK możemy wyznaczyć równanie prostej regresji (wzory pozwalające wykonać to zadanie podano np. w znanej Państwu instrukcji ONP i w materiałach do wykładu nr 14). Dostajemy równanie:

(10a.2),

gdzie *b0* i *b1* to estymatory współczynników i .

Oczywiście chcemy, żeby *b0* i *b1* dobrze szacowały i . Narzędziem służącym do sprawdzenia, czy estymatory są dobre jest tzw. analiza reszt (zob. założenia MNK w materiałach do wykładu nr 14). Rozważmy wartość *xi* zmiennej niezależnej. i-tą resztą *ei*nazywamy różnicę pomiędzy eksperymentalną wartością *yi*, a wartością obliczoną z równania (10a.2) dla wartości *xi*. Zauważmy, że konkretna wartość *ei* jest realizacją pewnej zmiennej losowej: wykonując wiele pomiarów dostaniemy zbiór różnych wartości *ei*.

Model z dobrze oszacowanymi przy pomocy MNK współczynnikami wymaga, aby każdy składnik resztowy spełniał określone założenia. Założenia te, mówiąc obrazowo, pozwalają ocenić, czy model (rozumiany jako zależność funkcyjna między X i Y) wychwycił wszystkie systematyczne zależności pomiędzy badanymi zmiennymi, a w resztach „siedzi” tylko wpływ przypadku. Zauważmy, że reszty możemy analizować dopiero po oszacowaniu parametrów modelu. Jeżeli reszty nie spełniają założeń, należy zbudować inny model lub zastosować inną metodę szacowania wartości parametrów.[[9]](#footnote-9)

Kolejne zagadnienie, które należy rozważyć przy weryfikacji modelu polega na ocenie, czy parametry modelu są istotne statystycznie, czyli czy dla populacji nie są zerami. Wyobraźmy sobie, że po w wyniku obliczeń dostaliśmy: i . Ponieważ wartości te obliczono na podstawie próby, może się zdarzyć, że tak naprawdę w populacji parametry modelu (lub jeden z nich) są zerami. Zauważmy, że przyjęcie, iż *b1* jest różne od zera ma wpływ jakościowy na model: mówi, że Y zależy od X, podczas gdy w rzeczywistości (dla populacji) takiej zależności nie ma. Dlatego zawsze należy sprawdzić istotność parametrów modelu (tzn. sprawdzić, czy poszczególne parametry nie są zerami dla populacji) testując hipotezy:

H0:

H1: .

Ponadto, testuje się istotność całego modelu, co polega na sprawdzeniu, czy istnieje liniowy związek między zmiennymi X i Y (osoby zainteresowane odsyłam do podręcznika Stanisza, s. 45).

Podsumowując, weryfikacja statystyczna modelu liniowego polega na sprawdzeniu istotności poszczególnych parametrów modelu, sprawdzeniu istotności całego modelu oraz sprawdzenia założeń metody najmniejszych kwadratów. Do tego ostatniego służy m.in. analiza reszt.

Oczywiście weryfikację statystyczną modelu przeprowadza się po oszacowaniu jego parametrów. Jeżeli weryfikacja statystyczna oraz weryfikacja merytoryczna, którą nie zajmujemy się w niniejszym kursie, nie powiodą się, należy zaproponować inny model.

**Klasyczny model regresji liniowej prostej**

Dla populacji prosty model regresji liniowej prostej[[10]](#footnote-11) wyraża się przez:

 (14.2)

gdzie *X* jest zmienną losową niezależną, *Y* oznacza zmienną losową zależną, a *ε* oznacza składnik losowy, który także jest zmienna losową. Składnik losowy opisuje wpływ wszystkich czynników, które, oprócz *X*, wpływają na *Y*.

Współczynniki β0 i β1, dotyczące populacji, nie są znane. Szacujemy je na podstawie *n*-elementowej próby, zawierającej pary realizacji zmiennych losowych *X* i *Y*, czyli i wyznaczamy równanie prostej:

 (14.3)

gdzie:  oznacza wartości leżące na tzw. linii regresji, a to tzw. ***reszta***, będąca także zmienną losową.

Współczynniki b0 i b1 są estymatorami współczynników β0 i β1 w populacji. Najczęściej wyznaczamy je korzystając z tzw. metody najmniejszych kwadratów (MNK). Polega ona na znalezieniu takich wartości b0 i b1, aby wyrażenie:

 (14.4a)

osiągnęło minimum. Wówczas:

 (14.4b)

 (14.4c)

gdzie  i  oznaczają, odpowiednio, średnie wartości zaobserwowane dla zmiennych *X* i *Y.[[11]](#footnote-12)*

W tym miejscu nawiążemy do części 2 wykładu nr 10 przypominając, że właściwe zastosowanie metody najmniejszych kwadratów wymaga spełnienia następujących założeń:

Założenia metody najmniejszych kwadratów zacytujemy za A. Staniszem:[[12]](#footnote-13)

1. „Model jest liniowy względem parametrów, tzn.  dla i = 1,2,...,n.”
2. „Liczba obserwacji n musi być większa lub równa liczbie oszacowanych parametrów (b0, b1), tj. n ≥ 2. „
3. „Składnik losowy ei ma wartość oczekiwaną równą zeru (E(ei) = 0 dla wszystkich i = 1,2,...,n).”
4. „Wariancja składnika losowego ei  (wariancja reszt) jest taka sama dla wszystkich obserwacji ( = 2σ dla wszystkich i = 1,2,...,n).”
5. „Składniki losowe (reszty) są nieskorelowane, czyli ei  oraz ej są od siebie niezależne dla wszystkich par i oraz j, gdzie i, j = 1,2,...,n oraz i ≠ j.”
6. „Każdy ze składników losowych (reszty) ma rozkład normalny.”

Spełnienie tych założeń jest wymagane jeśli estymatory b0 i b1 współczynników β0 i β1 w populacji mają mieć cechy dobrych estymatorów.

Liniowość modelu względem parametrów można ocenić wstępnie sporządzając tzw. wykres rozrzutu (*ang.* *scatterplot*). Wykres taki przedstawia dane (pary obserwacji na płaszczyźnie X0Y. Na podstawie układu punktów możemy zorientować się, czy możemy próbować zbudować dobry model liniowy. Liczba obserwacji jest w praktyce wielokrotnie większa od wymaganej.

Przyjrzyjmy się teraz autokorelacji reszt. Mówiąc obrazowo, polega ona na tym, że i-ta reszta „nie wie niczego” o swoich poprzedniczkach i jest od nich niezależna. Autokorelacja pierwszego rzędu wskazuje na niezależność reszty i-tej od i-1-szej, autokorelacja rzędu drugiego ocenia zależność reszty i-tej od i-2-giej itp. W przypadku optymalnym w układzie nie powinno być żadnych autokorelacji, w praktyce często (choć nie zawsze) ograniczamy się do zbadania, czy istnieje autokorelacja rzędu pierwszego.

Często używanym narzędziem sprawdzającym istnienie takiej właśnie autokorelacji (rzędu pierwszego) jest test Durbina-Watsona którego opis przytaczam za A. Staniszem.

*Poniższy opis opracowano na podstawie podręcznika Andrzeja Stanisza „Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny”, StatSoft Polska Sp. z o.o., Kraków 2007, tom 2, str. 103-105. UWAGA – w podręczniku są błędy edytorskie, starałam się je poprawić.*

Analiza reszt pozwalająca ocenić poprawność modelu polega na sprawdzeniu szeregu założeń, z których jedno mówi, że:

„Składniki losowe (reszty) są nieskorelowane, czyli i są ze sobą nieskorelowane dla wszystkich par *i*, *j* = 1,2,…,n oraz .” (A. Stanisz, str.103).

Autokorelacja reszt polega na niespełnieniu powyższego założenia. Najczęściej rozważa się autokorelację pierws:zego rzędu, dla której resztę można opisać jako: .

*ρ* jest współczynnikiem autokorelacji przyjmującym wartości od -1 do 1, a  pewnym składnikiem losowym. Jeżeli, reszty nie są skorelowane. Oczywiście mówienie o autokorelacji ma sens tylko w przypadku uporządkowanych zbiorów danych, tak jak to ma miejsce w przypadku procesów badanych w czasie.

Do badania istnienia autokorelacji (pierwszego rzędu) reszt służy test Durbina-Watsona.

Założenia:

„- analizowany model musi mieć wyraz wolny,

- składniki resztowe mają rozkład normalny (inne źródła podają też „zbliżony do normalnego” – wtrącenie IM),

- w modelu nie występuje opóźniona zmienna zależna w charakterze zmiennej niezależnej (przykładowo ),

- liczba obserwacji jest większa od 15.” (A. Stanisz, str. 104).

Hipoteza zerowa ma postać:

H0: 

Hipotezy alternatywne podano niżej.

Sprawdzian hipotezy (statystyka) ma postać:

 (14.5)

Statystyka *d* przybiera wartości od 0 do 4, przy czym wartość bliska 0 sugeruje autokorelację dodatnią reszt, a wartość bliska 4 – autokorelację ujemną. Wartość bliska 2 sugeruje brak autokorelacji.

Określamy w przybliżeniu tzw. estymator współczynnika autokorelacji  według równania:

 (14.6)

W zależności od znakutestujemy różne zestawy hipotez:

I. Dla przypadku 

H0: 

H1: 

Z tablic rozkładu statystyki *d* odczytujemy dla danego poziomu istotności wartości *dL* i *dU*. Następnie sprawdzamy, czy dla wyliczonej z naszej próby wartości statystyki *d* zachodzi:

1. - nie ma podstaw do odrzucenia H0 (brak autokorelacji),

2. - odrzucamy H0 i przyjmujemy H1 (autokorelacja dodatnia),

3. - test nie przynosi rozstrzygnięcia.

II. Dla przypadku 

H0: 

H1: 

Z tablic rozkładu statystyki *d* odczytujemy dla danego poziomu istotności wartości *dL* i *dU*. Następnie sprawdzamy, czy dla wyliczonej z naszej próby wartości statystyki *d* zachodzi:

1. - nie ma podstaw do odrzucenia H0 (brak autokorelacji),

2. - odrzucamy H0 i przyjmujemy H1 (autokorelacja ujemna),

3. - test nie przynosi rozstrzygnięcia.

**Ocena dopasowania linii regresji do danych[[13]](#footnote-14)**

Dla modeli skonstruowanych w oparciu o regresję liniową często wylicza się tzw. współczynnik determinacji *R2*. Ocenia on dopasowanie linii regresji do danych i pozwala ocenić siłę związku liniowego między rozpatrywanymi zmiennymi. Pamiętajmy jednak, że współczynnik ten, często wyliczany standardowo, nie jest wolny od wad. Dlatego w wielu przypadkach lepiej jest korzystać z tzw. poprawionego *R2* (*ang.* *adjusted R2*).

Wnikając w istotę współczynnika determinacji nieco głębiej zauważamy, że mówi jaka część zmienności danych jest wyjaśniana przez model. Mianowicie, *R2* wyraża się jako:

(14.7),

(oznaczenia wyjaśniono pod Rys. 14.1).

Aby zrozumieć wyprowadzenie wzoru (14.7), zauważmy, że dla każdego punktu , *i* = 1,2,…,*n* jest spełniona równość:

,

co ilustruje Rys. 14.1:

***x***

***y***

(*xi,yi*)

|  |  |
| --- | --- |
| Rys. 14.1. Niebieska linia - | wartość średnia , |
| czarna linia - | linia regresji , |
| czerwony odcinek **-** |  |
| zielony odcinek **-** |  |
| fioletowy odcinek **-** |  |

SST to tzw. całkowita suma kwadratów (*ang. Total Sum of Squares*), definiowana jako:

 (14.8a),

*n* jest liczbą obserwacji, każdy z *n* pomiarów jest reprezentowany przez punkt  , *i* = 1,2,...,*n* ; 

SSE to tzw. suma kwadratów błędów określona przez:

 (14.8b),

oznacza wartość teoretyczną leżącą na linii regresji dla danego *xi*,

SSR to tzw. suma kwadratów odchyleń regresyjnych, wyznaczana według :

 (14.8c).

Podobnie jak w analizie wariancji, zachodzi:

 (14.9)

*SSR* nazywamy zmiennością wyjaśnioną, która wynika ze zmienności zmiennej losowej *X*. *SSE* jest zmiennością niewyjaśnioną, wynikającą z błędu. Współczynnik zmienności określa zatem, jaka część całkowitej zmienności występującej w danych jest wyjaśniona. Im wyższa wartość *R2*, tym lepsze dopasowanie linii regresji do danych.

1. Stanisz, A., Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny, StatSoft, Kraków 2007, t. 2, s. 21. [↑](#footnote-ref-1)
2. Ibidem, s. 22. [↑](#footnote-ref-2)
3. Ibidem, s. 22. [↑](#footnote-ref-3)
4. Ibidem, s. 22. [↑](#footnote-ref-4)
5. Ibidem, s. 23. [↑](#footnote-ref-5)
6. Ibidem, s. 24. [↑](#footnote-ref-6)
7. Ibidem, s. 24-25. [↑](#footnote-ref-7)
8. Ibidem, s. 25. [↑](#footnote-ref-8)
9. Ibidem, s. 101. [↑](#footnote-ref-9)
10. Model zdefiniowano na podstawie: Stanisz, A., Przystępny kurs statystyki z zastosowaniem STATISTICA.PL na przykładach z medycyny, StatSoft, Kraków 2007, t.2, s. 25-26. [↑](#footnote-ref-11)
11. Tamże, s. 26-27. [↑](#footnote-ref-12)
12. Tamże, s. 33-35. [↑](#footnote-ref-13)
13. Opracowano na podstawie: Aczel, A.D., Statystyka w zarządzaniu, Wydawnictwo Naukowe PWN, Warszawa 2000, s. 490-493. [↑](#footnote-ref-14)