

Analiza danych z użyciem modeli regresyjnych z naruszonymi założeniami

Jarosław Gramacki, Artur Gramacki

Streszczenie: Praca dotyczy modelowania stochastycznego. Pokazano w jakim stopniu takie cechy danych jak autokorelacja oraz zmienna wariancja (niestacjonarność) ich błędów losowych wpływają na jakość modelu regresji użytego do analizowania tych danych. Pokazano w jaki sposób należy poprawnie estymować parametry modelu przy założeniu, że składnik błędu posiada wspomniane cechy. Zademonstrowano wpływ nieuwzględnienia tych cech na błędy estymowanych parametrów.

Słowa kluczowe: regresja, autokorelacja błędów, zmienność wariancji błędów, estymacja parametrów, modelowanie stochastyczne, identyfikacja statyczna

1. WPROWADZENIE

W dziedzinie eksploatacji danych wielokrotnie mamy do czynienia z potrzebą opisywania (modelowania) związków łączących duże ilości danych, które traktujemy jako wyjściowe (zależne, objaśniane), z dużymi ilościami danych, które traktujemy jako wejściowe (niezależne, objaśniające)⁶. Jako przykład można podać duże zbiory danych rejestrowanych w specjalistycznych systemach klasy DCS (ang. *Distributed Control System*), w które wyposażone są praktycznie wszystkie większe obiekty przemysłowe.

Bardzo wygodnym i prostym sposobem opisywania wspomnianych zależności są modele regresyjne, a w szczególności najprostsza ich wersja – regresja liniowa. W literaturze często zwraca się uwagę na fakt, że zastosowanie modelu liniowego do modelowania obiektów nieliniowych może w efekcie prowadzić do niedokładnych (a czasem wręcz błędnych) wyników. Rzadziej zwraca się natomiast uwagę na inne założenia towarzyszące modelom regresyjnym oraz ewentualne negatywne konsekwencje ich niespełnienia.

Struktura artykułu jest następująca: w rozdziale 2 przedstawiono założenia klasycznego modelu regresyjnego i wskazano, które z nich będą dokładniej analizowane. W rozdziale 3 pokazano dwa przykładowe, sztucznie wygenerowane zbiory danych, dla których wspomniane założenia nie są spełnione oraz pokazano ich wpływ na parametry wyestymowanych modeli. W rozdziale 4 wykonano symulacje komputerowe potwierdzające omówione wcześniej teoretycznie problemy. Rozdział 5

zawiera podsumowanie oraz wskazanie kierunków dalszych prac badawczych.

2. MODEL REGRESYJNY

W klasycznym modelu regresji liniowej zakłada się stacjonarność wariancji składników (zmiennych) losowych oraz brak autokorelacji ich wartości. Analiza skutków naruszenia tych dwóch założeń jest przedmiotem rozważań w niniejszej pracy.

Formalnie najprostszą postać modelu regresyjnego⁷ wraz ze wspomnianymi założeniami zapisuje się w postaci równań (1) – (3)

$$y_i = \beta_1 + \beta_2 x_i + e_i \quad (1)$$

$$\text{var}(e_i) = \sigma^2 \quad (2)$$

$$\text{cov}(e_i, e_j) = 0, \quad i \neq j; \quad i, j = 1 \dots n, \quad (3)$$

gdzie β_1, β_2 są nieznanymi parametrami modelu, które należy wyznaczyć (wyestymować), y jest zmienną objaśnianą (zależną), x jest zmienną objaśniającą (niezależną), e jest zmienną losową (składnikiem losowym, błędem losowym) o stałej wariancji σ^2 , n oznacza ilość próbek (obserwacji). Pozostałe założenia klasycznego modelu regresji liniowej: a) wartość oczekiwana błędu losowego $E(e) = 0$, b) nielosowość zmiennej objaśniającej x , c) rozkład normalny błędu losowego e wokół średniej $e \approx N(0, \sigma^2)$, d) brak korelacji składnika losowego ze zmienną objaśniającą, nie są w naszych rozważaniach istotne i zakłada się, że są spełnione.

Zwykła metoda najmniejszych kwadratów (ang. Ordinary Least Squares – OLS) polega na wyborze takich oszacowań (estymatorów) $\hat{\beta}_1, \hat{\beta}_2$ parametrów β_1, β_2 , aby wartość wyrażenia

$$Q = \sum_{i=1}^n \hat{e}_i = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)^2 \quad (4)$$

była jak najmniejsza. Symbol $\hat{e}_i$ oznacza tu estymatę rzeczywistych (ale nieobserwowalnych!) błędów e_i zwanych resztami (residuumi) najmniejszych kwadratów.

W pracy skupimy się na badaniu wartości tzw. błędów standardowych estymatorów w przypadku niespełnienia

⁶ W literaturze angielskiej często używa się również sformułowań *endogenous variables* oraz *exogenous variables*.

⁷ W pracy rozważa się wyłącznie modele z jedną zmienną objaśniającą x oraz jedną zmienną objaśnianą y . Upraszcza to rozważania, niezmieniając przy tym istoty zagadnienia.

założeń wymienionych na początku tego rozdziału. Poniżej przedstawiono więc krótko odpowiednie formuły.

Rzeczywista wariancja składników losowych jest z oczywistych powodów nieznana. Należy ją więc wyestymować na podstawie otrzymanych z modelu wartości reszt

$$\hat{\sigma}^2 = \frac{\sum \hat{e}_i^2}{n-k}, \quad (5)$$

gdzie k jest ilością estymowanych parametrów (u nas 2). Wykazuje się [1], że wartość $n - k$ (zamiast n) powoduje, że estymator (5) rzeczywistej wartości wariancji (2) posiada ważną cechę: jest estymatorem nieobciążonym, czyli $E(\hat{\sigma}^2) = \sigma^2$. Mając dany estymator (5) można [1] wyestymować wariancję estymatorów parametrów:

$$\begin{aligned} \text{var}(\hat{\beta}_1) &= \hat{\sigma}^2 \left[\frac{\sum x_i^2}{n \sum (x_i - \bar{x})^2} \right], \\ \text{var}(\hat{\beta}_2) &= \left[\frac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2} \right]. \end{aligned} \quad (6)$$

Pierwiastki z (6) są ostatecznymi wyrażeniami na błędy standardowe estymatorów $\hat{\beta}_1, \hat{\beta}_2$. Przykładowo:

$$se(\hat{\beta}_1) = \sqrt{\text{var}(\hat{\beta}_1)}. \quad (7)$$

Wartości tych błędów są niezbędne przy określaniu użyteczności modelu regresyjnego w postaci tzw.

przedziałów ufności estymatorów $\hat{\beta}_1, \hat{\beta}_2$ uzyskiwanych na drodze testowania odpowiednich hipotez statystycznych⁸.

W kolejnych rozdziałach zajmiemy się analizą konsekwencji niespełnienia założenia (2) oraz założenia (3).

2.1. NIESTACJONARNOŚĆ WARIANCJI BŁĘDÓW

Z niestacjonarnością wariancji błędów losowych spotykamy się w przypadku analizy danych zarówno przekrojowych (ang. Cross-Sectional Data) jak i mających charakter szeregów czasowych (ang. Time Series). Pierwszy rodzaj danych najczęściej używany jest przy opisie różnorodnych zjawisk ekonomicznych i nie będziemy się takimi danymi zajmowali. Drugi rodzaj danych często pochodzi z obiektów technicznych opomiarowanych wieloma czujnikami, które to dane są rejestrowane w specjalistycznych bazach danych. Z niestacjonarnością wariancji błędów losowych spotkamy się tam na przykład w przypadku analizy obiektu w różnych jego stanach pracy. Przykładowo praca obiektu energetycznego z wyższymi parametrami danego czynników termodynamicznych powoduje, że zdecydowanie trudniej jest przewidzieć zachowanie się jego umownego wyjścia. Przekładając ostatecznie stwierdzenie na język statystyki powiemy, że błąd losowy w modelu (tu) regresyjnym rośnie wraz ze wzrostem wartości zmiennej niezależnej, czyli jest niestacjonarny. Innym przykładem może być zmienna w czasie zbierania danych dokładność przyrządów pomiarowych. Będziemy mieli wtedy do czynienia z tzw. niestacjonarnością przedziałową.

Z analizy przeprowadzonej w poprzednim rozdziale widać, że w przypadku niespełnienia założenia (2) wyrażenia (6) nie będą poprawne, co w konsekwencji wpłynie na (7), które z kolei wpłyną na błędne oszacowanie przedziałów ufności dla estymatorów parametrów określonych wyrażeniami

$$\hat{\beta}_1 \pm t_c se(\hat{\beta}_1), \quad \hat{\beta}_2 \pm t_c se(\hat{\beta}_2), \quad (8)$$

gdzie t_c jest wartością rozkładu t-studenta $t(1-\alpha/2, n-2)$. α oznacza tu założoną przez użytkownika wartość tzw. poziomu istotności, który zwykle przyjmuje wartość 0.05. Obliczenie poprawnych wartości (6) i (7) wymagać będzie zastosowania tzw. uogólnionej metody najmniejszych kwadratów (ang. Generalized Least Squares, GLS), która uwzględnia konkretną formę zaburzenia założenia (2). W pracy przyjmiemy, że zaburzenie ma formę $\text{var}(e_i) = \sigma_i^2 = c h(x_i)$, gdzie c jest ustaloną stałą, h jest pewną funkcją zmiennej niezależnej x_i , rosnącą wraz z x_i .

2.2. AUTOKORELACJA BŁĘDÓW

Podobne problemy napotkamy przy niespełnieniu założenia (3), czyli przy skorelowaniu błędów losowych a w konsekwencji wartości reszt modelu. Skorelowanie takie będzie niemal zawsze obecne w danych o charakterze szeregów czasowych. Przykładowo zakłócenie pracy obiektu energetycznego (ze statystycznego punktu widzenia opisuje to błąd losowy) w czasie t będzie odczuwane nie tylko w czasie jego zaistnienia ale i również w kolejnych okresach czasu $t+1, t+2, t+3$ itd.

W pracy ograniczymy się do autokorelacji składnika losowego rzędu pierwszego AR(1)⁹. W równaniu (1) należy więc doprecyzować postać błędu, czyli

$$e_i = p e_{i-1} + v_i \quad -1 < p < 1, \quad (9)$$

gdzie p jest współczynnikiem autokorelacji, v jest zakłóceniem spełniającym założenia postaci (2), (3) czyli

w praktyce jest to błąd o rozkładzie normalnym $N(0, \sigma^2)$. W tym przypadku również okazuje się, że formuły (6) i (7) nie są poprawne (złudnie dają zbyt optymistyczne wyniki!), co w konsekwencji, tak jak poprzednio, prowadzi do błędnej oceny przedziałów ufności estymatorów. Obliczenie poprawnych wartości (6) i (7) wymagać będzie zastosowania uogólnionej metody najmniejszych kwadratów, która w tym przypadku korygować będzie fakt występowania autokorelacji błędów losowych.

3. ESTYMACJA PARAMETRÓW MODELI ZAKŁÓCONYCH

Z uwagi na ograniczoną objętość pracy, szczegóły konkretnych metod typu GLS nie będą pokazane. Można je znaleźć np. w [1, 2]. Wygenerowane zostaną natomiast sztuczne zbiory danych niespełniające założeń (2) i (3) a następnie wykonana zostanie dla tych danych estymacja parametryczna modelu regresyjnego -- raz klasyczną metodą OLS, a następnie metodą uogólnioną GLS.

⁸ Aby testowanie hipotez było z teoretycznego punktu widzenia możliwe, konieczne jest założenie normalności rozkładu reszt $e \approx N(0, \sigma^2)$. Założenie to, na mocy centralnego twierdzenia granicznego, nie jest jednak krytyczne w przypadku dużej ilości danych pomiarowych (duże wartości n).

⁹ W większości przypadków wykorzystywana dalej korekcja estymatora dla przypadku AR(1) jest wystarczająca.

3.1. NIESTACJONARNOŚĆ WARIANCJI BŁĘDÓW

Rozważony zostanie następujący model

$$\begin{aligned} y_i &= \beta_1 + \beta_2 x_i + e_i \\ \text{var}(e_i) &= \sigma_i^2 = \sigma^2 x_i \quad , \\ x_i &= cx_{i-1} + u_i, \quad i = 1 \dots n \end{aligned} \quad (10)$$

którego charakterystyczną cechą jest niestacjonarność błędu losowego e_i . Wariancja błędu jest proporcjonalna do zmiennej niezależnej x . Zmienna niezależna x tworzy rosnący ciąg wartości¹⁰, gdyż założono, że $E(u_i) > 0$. W modelu uzyskamy więc bardzo typowy wykres niestacjonarności błędu losowego. Składowy losowy u pochodzi z rozkładu normalnego o średniej równej 0.85 oraz wariancji równej 1. Na podstawie specyfikacji (10) wygenerowano przykładowe wartości e oraz x a następnie obliczono y . W modelu mamy: $n = 50, x_0 = 1, c = 1, \beta_1 = 0.5, \beta_2 = 2$. Po zastosowaniu metody OLS zademonstrowany zostanie negatywny wpływ niestacjonarności błędów losowych na wynikowy model regresyjny. Zastosowanie właściwej do danego problemu metody uogólnionej GLS rozwiąże problem błędnego szacowania błędów standardowych estymatorów.

3.2. AUTOKORELACJA BŁĘDÓW

Rozważony zostanie następujący model [3]

$$\begin{aligned} y_i &= \beta_1 + \beta_2 x_i + e_i \\ e_i &= \rho e_{i-1} + v_i \quad , \\ x_i &= cx_{i-1} + u_i, \quad i = 1 \dots n \end{aligned} \quad (11)$$

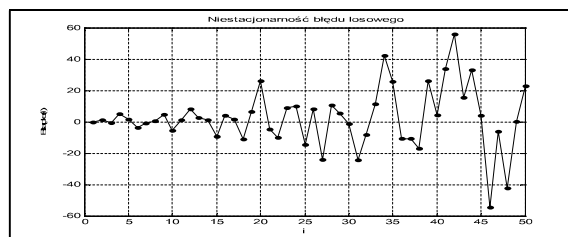
Składowy v_i błędu losowego¹¹ spełnia założenia (2), (3). W modelu zarówno e jak i x podlegają autokorelacji. Ponadto zakładamy, że zmienna niezależna x rośnie wraz z czasem. Składowy losowy v i u pochodzą z rozkładu normalnego o średniej równej odpowiednio -0.1 i 0.85 oraz wariancji równej 1. Na podstawie specyfikacji (11) wygenerowane przykładowe wartości e oraz x a następnie obliczono y . W modelu mamy $n = 50, x_0 = 1, \rho = 0.8, c = 1, \beta_1 = 0.5, \beta_2 = 2$. Po zastosowaniu metody OLS zademonstrowany zostanie negatywny wpływ skorelowania błędów losowych na wynikowy model regresyjny. Kontrprzykładem do (11) będzie analogiczny zbiór danych, w którym błąd e_i nie będzie podlegał autokorelacji. Dla zachowania „skali” będzie miał on ponadto dokładnie taką samą wariancję. Zastosowanie odpowiedniej do problemu metody uogólnionej GLS rozwiąże problem błędnego szacowania błędów standardowych estymatorów.

4. WYNIKI SYMULACJI KOMPUTEROWYCH

4.1. NIESTACJONARNOŚĆ WARIANCJI BŁĘDÓW

Na rysunku 1 pokazano przebieg zmienności błędu losowego użytego w przykładzie (patrz rozdział 3.1). Widać na nim wyraźnie wzrost jego wartości w funkcji kolejnych próbek. W tabeli 1 zamieszczono wyniki regresji z użyciem metody OLS, która w żaden sposób nie

uwzględnia faktu naruszenia założenia (2). Pogrubioną czcionką zaznaczono wartości wyestymowanych parametrów (Coeff.), ich błędy standardowe (S.E.) oraz 95% przedziały ufności dla parametrów. W tabeli 2 zamieszczono wyniki regresji z użyciem metody GLS uwzględniającej niestacjonarność błędów. Użyto tu tzw. estymatora ważonego najmniejszych kwadratów (ang. Weighted Least Squares, WLS). Odpowiada to wykonaniu zwykłej regresji OLS dla odpowiednio przetransformowanych zmiennych. Transformacja ta [2] ma na celu usunięcie z modelu niestacjonarności składnika błędu losowego. Za wagę przyjęto wartość $1/\sqrt{x_i}$. Odpowiada ona założonej w (10) postaci błędu¹².



Rys. 1. Wykres zmienności błędów losowych modelu

Tabela 1.

Regresja OLS dla danych z błędami niestacjonarnymi

	Coeff.	S.E.	t-ratio	p-value
const	0.755	5.278	0.143	0.887
xi	2.104	0.212	9.933	3.15e-013
95% CONFIDENCE INTERVAL				
const	0.755	-9.857	11.367	
xi	2.104	1.679	2.530	

Tabela 2.

Regresja GLS dla danych z błędami niestacjonarnymi

	Coeff.	S.E.	t-ratio	p-value
const	0.770	1.967	0.3914	0.6972
xi	2.104	0.140	14.96	1.06e-019
95% CONFIDENCE INTERVAL				
const	0.770	-3.184	4.724	
xi	2.104	1.821	2.385	

Porównując wyniki obu regresji widać, że estymator OLS jest nieobciążony, gdyż wartości parametrów są (prawie) takie same jak te uzyskane z użyciem estymatora GLS. Wartości błędów standardowych istotnie jednak różnią się, co rzutuje na wartości przedziałów ufności. Pamiętajmy, że poprawne wyznaczenie tych przedziałów jest bardzo ważne z praktycznego punktu widzenia, gdyż umożliwia rzeczywistą ocenę jakości modelu.

4.2. AUTOKORELACJA BŁĘDÓW

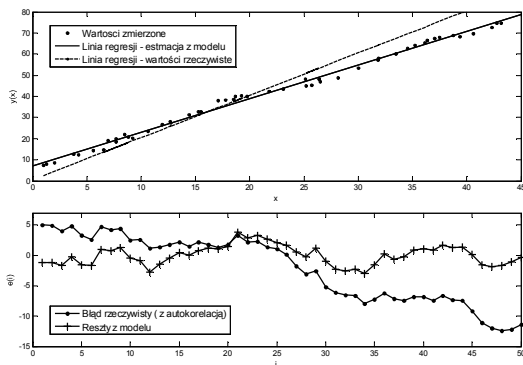
Na rysunku 2a pokazano przebieg prawdziwej linii regresji (zdefiniowanej przez wartości $\beta_1 = 0.5, \beta_2 = 2$, patrz rozdział 3.2) oraz linii regresji otrzymanej z modelu. Na rysunku 2b pokazano rzeczywistą postać składnika losowego e_i oraz jego przebieg w postaci otrzymanych reszt z modelu (residuum). Różnice w przebiegach (otrzymane residua nie „oddają” rzeczywistego charakteru

¹⁰ Fakt monotonicznego wzrostu x (w tym i następnym przykładzie) nie ma tu znaczenia merytorycznego. Ułatwia jednak zaobserwowanie opisywanego zjawiska.

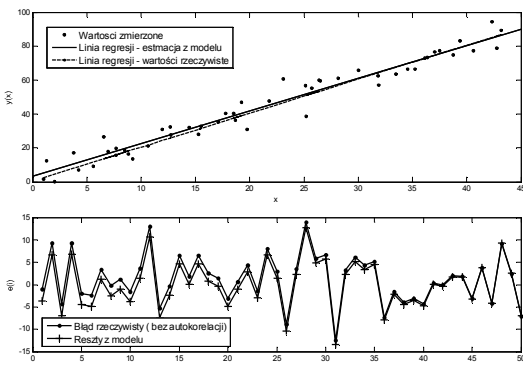
¹¹ W przypadku błędu losowego o charakterze AR(1), w literaturze przyjęło się rozróżniać dwa jego elementy. Składowy e to błąd (ang. *error term*), składowy v to zakłócenie (ang. *disturbance*) lub zmienna losowa (ang. *random variable*).

¹² W wielu praktycznych zastosowaniach model błędu, w którym jego rozrzut (wariancja) rośnie wraz z mierzona wartością, będzie właściwy.

błędów) są powodem stosunkowo dużej rozbieżności obu linii regresji. W przypadku braku autokorelacji składnika losowego (rysunek 3a), różnice pomiędzy rzeczywistym błędem a resztami z modelu są bardzo niewielkie¹³. Skutkuje to dużo lepszą zgodnością modelu (rysunek 3b). W tabelach 3 oraz 4, analogicznie jak w pierwszym przykładzie, pokazano różnice pomiędzy wartościami błędów standardowych otrzymanych metodą OLS (niewłaściwą) oraz GLS (uwzględniającą autokorelację). Inaczej niż w pierwszym przykładzie poprawne wartości są tu większe. Stosując więc niewłaściwą do problemu metodę, uzyskujemy wyniki zbyt optymistyczne!



Rys. 2. Regresja dla przypadku błędów skorelowanych. (a) linie regresji, (b) błędy rzeczywiste i zaobserwowane z modelu



Rys. 3 Regresja dla przypadku błędów nieskorelowanych. (a) linie regresji, (b) błędy rzeczywiste i zaobserwowane z modelu

Tabela 3.

Regresja OLS dla danych z błędami skorelowanymi

	Coeff.	S.E.	t-ratio	p-value
const	7.098	0.453	15.66	1.70e-020
xi	1.589	0.018	87.31	1.49e-054
Durbin-Watson	0.421			
	95% CONFIDENCE INTERVAL			
const	7.098	6.186	8.009	
xi	1.588	1.552	1.625	

Tabela 4.

Regresja GLS dla danych z błędami skorelowanymi

	Coeff.	S.E.	t-ratio	p-value
const	6.085	1.240	4.909	1.10e-05
xi	1.630	0.047	34.69	1.20e-035

Durbin-Watson	1.888		
	95% CONFIDENCE INTERVAL		
const	6.085	3.592	8.577
xi	1.630	1.536	1.724

5. PODSUMOWANIE

W przypadku występowania niestacjonarności błędów losowych zastosowanie niewłaściwej metody estymacji parametrów modelu regresji prowadzi do zawyżenia błędów standardowych estymatorów. W przypadku występowania autokorelacji błędów losowych, zastosowanie niewłaściwej metody prowadzi do niedoszacowania wartości tych błędów. Ponadto otrzymane reszty nie oddają charakteru rzeczywistego zakłócenia. W obu przypadkach skutkuje to błędnymi oszacowaniami przedziałów istotności wyliczonych parametrów równań regresji.

5.1. DALSZE PRACE

W pracy zakłada się, że dostępne dane opisują obiekt w jego statycznym stanie pracy. Podobne zjawiska mogą wystąpić również w obiektach dynamicznych. W modelach takich obiektów występować będą odpowiednio opóźnione zmienne objaśniane i/lub objaśniające oraz mogą również mieć miejsce opisywane w niniejszej pracy skorelowane lub o zmiennej wariancji błędy losowe. W takich przypadkach konieczne będzie podanie stosownych do sytuacji metod estymacji parametrów modeli.

W pracy nie przeanalizowano przypadków jednoczesnego występowania w modelu obu rodzajów zakłóceń. Nie przeanalizowano przypadku, w którym zmienne objaśniane powinny być traktowane jako zmienne losowe (uwzględnienie błędów ich pomiaru) oraz przypadków skorelowania zmiennych objaśniających z błędami losowymi.

LITERATURA

- [1] Weisberg S., *Applied Linear Regression*. Wiley-Interscience, 2005.
- [2] Sheather S. J., *A Modern Approach to Regression with R*. Springer, 2009.
- [3] Wonnacott R., Wonnacott T., *Econometrics*. John Wiley & Sons, 1989.



dr inż. Jarosław Gramacki
Uniwersytet Zielonogórski
Wydział Elektrotechniki, Informatyki
i Telekomunikacji
Instytut Informatyki i Elektroniki

ul. Podgórna 50
65-246 Zielona Góra
e-mail: a.gramacki@iie.uz.zgora.pl



dr inż. Artur Gramacki
Uniwersytet Zielonogórski
Wydział Elektrotechniki, Informatyki
i Telekomunikacji
Instytut Informatyki i Elektroniki

ul. Podgórna 50
65-246 Zielona Góra
e-mail: a.gramacki@iie.uz.zgora.pl

¹³ Podkreślimy również, że użyte w przykładach wariancje błędów z i bez autokorelacji są takie same (mają wartość ok. 30).