

Piotr Peternek

Uniwersytet Ekonomiczny
we Wrocławiu

Marek Kośny

Uniwersytet Ekonomiczny
we Wrocławiu

Kilka uwag o testowaniu istotności współczynnika korelacji

Streszczenie. W artykule przedstawiono problem interpretacji wyników testu współczynnika korelacji Pearsona w powiązaniu z budowanym modelem regresji. Zwrócono przy tym uwagę na znaczną wrażliwość testu istotności na wielkość próby. W tym kontekście przedstawiono pewien alternatywny sposób testowania oraz rekomendowano znany, lecz stosunkowo rzadko wykorzystywany test, umożliwiający poprawną interpretację wyników testu istotności współczynnika korelacji.

Słowa kluczowe: współczynnik korelacji Pearsona, testy istotności, model regresji

Wstęp

Problematyka badania zależności zmiennych zarówno o charakterze ilościowym, jak i jakościowym jest jednym z najczęściej wykorzystywanych zagadnień statystyki w badaniach empirycznych. Szerokie możliwości zastosowania umożliwiają przedstawienie wielu technik i narzędzi pomiarowych. Mimo znacznej liczby metod w zakresie analizy zmiennych o charakterze mierzalnym, na pierwsze miejsce w częstotliwości wykorzystania wysuwa się doskonale znany współczynnik korelacji Pearsona. Jego zastosowanie daje dodatkową możliwość, jaką jest

eksploatacja modelu regresji. Jednak konstrukcja każdego modelu matematycznego, w tym także modelu regresji, powinna być poprzedzona przeprowadzeniem adekwatnego testu statystycznego, czyli weryfikacją modelu. Wykorzystanie testów istotności wiąże się jednak z pewnymi ograniczeniami, których badacze często nie są w pełni świadomi. Do takich ograniczeń należy problem wielkości próby (problem zarówno zbyt małej, jak i bardzo licznej próby) powiązany z odpowiednią interpretacją. Czy zasadne jest bowiem konstruowanie modelu regresji zawsze w przypadku, gdy przeprowadzony test istotności współczynnika korelacji odrzuca hipotezę zerową (czy zasadne jest budowanie regresji dla istotnie różnego od zera współczynnika korelacji wynoszącego np. $r = 0,1$)?

W pracy przedstawiony zostanie problem zależności wyników najczęściej wykorzystywanego testu istotności współczynnika korelacji od wielkości próby. Zostanie też podana pewna propozycja mająca na celu ułatwienie przeprowadzenia takiego testu oraz przypomniany i rekomendowany będzie alternatywny sposób testowania współczynnika korelacji mający na celu poprawne zbudowanie modelu regresji.

1. Test istotności współczynnika korelacji a wielkość próby

Rozpatrywany jest liniowy model ekonometryczny postaci

$$y = \alpha X + \beta + \varepsilon,$$

gdzie $E(\varepsilon) = 0$, $E(\varepsilon\varepsilon^T) = \sigma^2 I$.

Wybór zmiennych do takiego modelu bazuje zwykle na analizie istotności współczynnika korelacji Pearsona¹. W tej pracy ograniczono się wyłącznie do modelu jednej zmiennej, co nie zmienia ogólności rozważań.

Wykorzystanie w badaniach empirycznych współczynnika korelacji Pearsona sprowadza się po jego obliczeniu do interpretacji (zob. tab. 1) oraz przeprowadzenia wspomnianego wcześniej testu istotności. Test ten najczęściej przybiera postać, w której formułowane są następujące hipotezy:

$$H_0: \rho = 0, \quad H_1: \rho \neq 0,$$

gdzie ρ oznacza prawdziwą wartość współczynnika korelacji w populacji, a hipotezę zerową interpretuje się jako brak zależności liniowej pomiędzy zmiennymi. Tu przypomnijmy jeszcze, że odrzucenie hipotezy zerowej oznacza, że wartość współczynnika korelacji jest istotnie różna od zera.

Weryfikacja prawdziwości hipotezy H_0 dokonywana jest na podstawie statystyki:

¹ Zob. np. *Ekonometria. Metody, zadania, przykłady*, red. J. Dziechciarz, Wydawnictwo AE we Wrocławiu, Wrocław 2002.

$$T_{n-2} = \frac{r}{\sqrt{1-r^2}} \sqrt{n-2}, \quad (1)$$

która (przy założeniu prawdziwości H_0) ma rozkład t -Studenta o $n - 2$ stopniach swobody². Jak wynika ze wzoru (1), wartość statystyki testowej zależy wyłącznie od wartości współczynnika korelacji w analizowanej próbie oraz od liczebności próby. Zamiast więc weryfikacji, czy uzyskany wynik jest istotny statystycznie, można sformułować pytanie, jak duża powinna być próba, by dla określonej wartości współczynnika korelacji oraz na zadanym poziomie istotności należało odrzucić hipotezę zerową na rzecz alternatywnej.

Tabela 1. Wartości współczynnika korelacji i ich interpretacje

Wartość współczynnika korelacji	Interpretacja
Poniżej 0,2	Praktycznie brak związku liniowego między badanymi cechami
0,2 do 0,4	Zależność liniowa wyraźna, lecz niska
0,4 do 0,7	Zależność umiarkowana
0,7 do 0,9	Zależność znacząca
Powyżej 0,9	Zależność bardzo silna

Źródło: opracowanie własne na podstawie: S. Ostasiewicz, Z. Rusnak, U. Siedlecka, *Statystyka. Elementy teorii i zadania*, Wydawnictwo Naukowe PWN, Warszawa 2000, s. 311.

Przeprowadzono zatem symulację, w której dla założonego poziomu istotności $\alpha = 0,05$ (tradycyjnie najczęściej przyjmowany poziom istotności) i przy wykorzystaniu metod numerycznych wygenerowano liczbę obserwacji niezbędną do uzyskania istotności współczynnika korelacji o dowolnej wartości. Wyniki dla wybranych wartości współczynnika korelacji przedstawiono w tabeli 2. Wartości te uzyskano wykorzystując moduł optymalizacyjny Solver zawarty w programie MS Excel. Przyjęto wybrane wartości współczynnika korelacji, a następnie metodami numerycznymi poszukiwano takich wartości liczebności próby, które gwarantowałyby istotność współczynnika korelacji. Zatem np. dla współczynnika korelacji $r = 0,8$ próba, która jest niezbędna do uznania takiego współczynnika za istotny, wynosi $n = 7$, gdyż wartość statystyki testowej (1) wynosi wówczas $T_5 = \frac{0,8}{\sqrt{1-0,8^2}} \sqrt{7-2} \approx 2,98$, natomiast wartość krytyczna uzyskana z tablic rozkładu Studenta dla 5 stopni swobody i przyjętego poziomu istotności wynosi wówczas $t = 2,571$. Warto zwrócić uwagę na możliwość zapewnienia istotności współczynnika korelacji nawet bardzo niskiej wartości przy stosunkowo

² Por. W. Krysiński, J. Bartos, W. Dyczka, K. Królikowska, M. Wasilewski, *Rachunek prawdopodobieństwa i statystyka matematyczna w zadaniach*, Wydawnictwo Naukowe PWN, Warszawa 2000.

niewielkiej próbie. Zwróćmy bowiem uwagę, że np. dla współczynnika korelacji rzędu $r = 0,1$ wystarczy próba o liczebności 385, by współczynnik taki uznać za istotnie różny od zera. Podobnie ma się sytuacja ze współczynnikiem korelacji $r = 0,01$, tu (ku zaskoczeniu autorów pracy) wystarczy próba, w której znajdzie się 38 416 elementów, by współczynnik o tak niskiej wartości uznać za istotnie różny od zera.

Tabela 2. Liczebności próby gwarantujące istotność współczynnika korelacji

r	N	r	N	r	N	r	N
0,01	38416	0,26	58	0,51	16	0,76	7
0,02	9605	0,27	54	0,52	15	0,77	7
0,03	4269	0,28	50	0,53	15	0,78	7
0,04	2402	0,29	47	0,54	14	0,79	7
0,05	1538	0,30	44	0,55	14	0,80	7
0,06	1068	0,31	41	0,56	13	0,81	7
0,07	785	0,32	39	0,57	13	0,82	6
0,08	601	0,33	36	0,58	12	0,83	6
0,09	475	0,34	34	0,59	12	0,84	6
0,10	385	0,35	32	0,60	12	0,85	6
0,11	319	0,36	31	0,61	11	0,86	6
0,12	268	0,37	29	0,62	11	0,87	6
0,13	228	0,38	28	0,63	11	0,88	5
0,14	197	0,39	26	0,64	10	0,89	5
0,15	172	0,40	25	0,65	10	0,90	5
0,16	151	0,41	24	0,66	10	0,91	5
0,17	134	0,42	23	0,67	9	0,92	5
0,18	120	0,43	22	0,68	9	0,93	5
0,19	107	0,44	21	0,69	9	0,94	5
0,20	97	0,45	20	0,70	9	0,95	4
0,21	88	0,46	19	0,71	8	0,96	4
0,22	80	0,47	18	0,72	8	0,97	4
0,23	74	0,48	18	0,73	8	0,98	4
0,24	68	0,49	17	0,74	8	0,99	4
0,25	63	0,50	16	0,75	8		

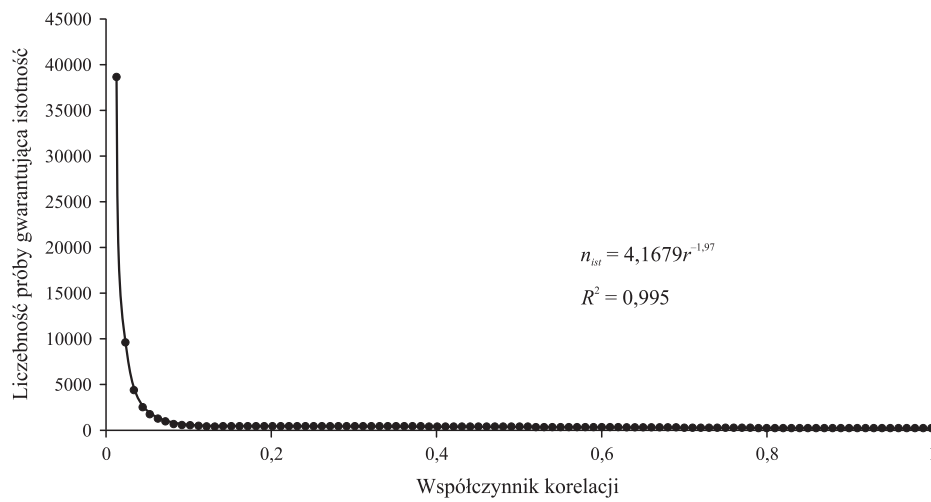
Źródło: opracowanie własne.

Znalezienie liczby obserwacji gwarantującej istotność dla dowolnego współczynnika korelacji umożliwiło przedstawienie uzyskanych rezultatów na wykresie, a w konsekwencji dopasowania do nich adekwatnej funkcji (zob. rys. 1).

Funkcją najdokładniej opisującą prezentowane dane ($R^2 = 0,995$) była funkcja potęgowa postaci

$$n_{ist} = 4,1679r^{-1,79},$$

gdzie n_{ist} oznacza liczebność próby gwarantującą istotność współczynnika korelacji r .



Rys. 1. Dopasowanie funkcji potęgowej do zależności współczynnika korelacji – wielkość próby

Źródło: opracowanie własne.

Tak dobre dopasowanie funkcji ($R^2 = 0,995$) umożliwia przedstawienie alternatywnego i jak się wydaje łatwiejszego sposobu testowania istotności współczynnika korelacji. W tabeli 3 przedstawiono wartość współczynnika korelacji z odpowiadającą mu minimalną liczebnością próby gwarantującą istotność oraz wartość liczebności próby uzyskaną z dopasowanej wcześniej funkcji potęgowej, a także różnicę między prezentowanymi liczebnościami. Zauważmy (por. tab. 3), że dla każdego współczynnika korelacji $|r| > 0,6$ można uznać go za istotnie różny od zera, jeżeli liczba obserwacji n , na podstawie której liczony był współczynnik r , jest większa od $n_{ist} = 4,1679r^{-1,79}$. Zamiast zatem tradycyjnego testowania istotności współczynnika korelacji polegającego na obliczeniu odpowiedniej statystyki testowej, a następnie odczytaniu wartości krytycznych z tablic i ich wzajemnym porównaniu, można przeprowadzić test istotności z wykorzystaniem dopasowanej funkcji. Jeżeli bowiem liczebność próby, na podstawie której liczony był współczynnik korelacji, jest większa od $n_{ist} = 4,1679r^{-1,79}$, to współczynnik taki należałoby uznać za istotnie różny od zera. Przypomnijmy tu, że ten sposób postępowania jest odpowiedni dla sytuacji, w której przyjęto błąd pierwszego rodzaju równy 5% i jednocześnie $|r| > 0,6$.

Tabela 3. Współczynniki korelacji wraz z odpowiadającymi im liczebnościami próby gwarantującymi istotność oraz ich różnice

<i>r</i>	<i>n</i>	<i>n_{ist}</i>	<i>n_{ist} - n</i>	<i>r</i>	<i>n</i>	<i>n_{ist}</i>	<i>n_{ist} - n</i>	<i>r</i>	<i>n</i>	<i>n_{ist}</i>	<i>n_{ist} - n</i>	<i>r</i>	<i>n</i>	<i>n_{ist}</i>	<i>n_{ist} - n</i>
0,01	38416	36300,89	-115,11	0,26	58	59,21	1,21	0,51	16	15,70	-0,30	0,76	7	7,16	0,16
0,02	9605	9265,91	-339,09	0,27	54	54,97	0,97	0,52	15	15,11	0,11	0,77	7	6,97	-0,03
0,03	4269	4168,58	-100,42	0,28	50	51,17	1,17	0,53	15	14,56	-0,44	0,78	7	6,80	-0,20
0,04	2402	2365,15	-36,85	0,29	47	47,75	0,75	0,54	14	14,03	0,03	0,79	7	6,63	-0,37
0,05	1538	1523,86	-14,14	0,30	44	44,67	0,67	0,55	14	13,53	-0,47	0,80	7	6,47	-0,53
0,06	1068	1064,04	-3,96	0,31	41	41,87	0,87	0,56	13	13,06	0,06	0,81	7	6,31	-0,69
0,07	785	785,37	0,37	0,32	39	39,33	0,33	0,57	13	12,61	-0,39	0,82	6	6,16	0,16
0,08	601	603,71	2,71	0,33	36	37,02	1,02	0,58	12	12,19	0,19	0,83	6	6,02	0,02
0,09	475	478,70	3,70	0,34	34	34,91	0,91	0,59	12	11,79	-0,21	0,84	6	5,88	-0,12
0,10	385	388,97	3,97	0,35	32	32,97	0,97	0,60	12	11,40	-0,60	0,85	6	5,74	-0,26
0,11	319	322,38	3,38	0,36	31	31,19	0,19	0,61	11	11,04	0,04	0,86	6	5,61	-0,39
0,12	268	271,60	3,60	0,37	29	29,55	0,55	0,62	11	10,69	-0,31	0,87	6	5,48	-0,52
0,13	228	231,98	3,98	0,38	28	28,04	0,04	0,63	11	10,36	-0,64	0,88	5	5,36	0,36
0,14	197	200,47	3,47	0,39	26	26,64	0,64	0,64	10	10,04	0,04	0,89	5	5,24	0,24
0,15	172	174,99	2,99	0,40	25	25,34	0,34	0,65	10	9,74	-0,26	0,90	5	5,13	0,13
0,16	151	154,10	3,10	0,41	24	24,14	0,14	0,66	10	9,45	-0,55	0,91	5	5,02	0,02
0,17	134	136,75	2,75	0,42	23	23,02	0,02	0,67	9	9,17	0,17	0,92	5	4,91	-0,09
0,18	120	122,19	2,19	0,43	22	21,98	-0,02	0,68	9	8,91	-0,09	0,93	5	4,81	-0,19
0,19	107	109,84	2,84	0,44	21	21,00	0,00	0,69	9	8,66	-0,34	0,94	5	4,71	-0,29
0,20	97	99,29	2,29	0,45	20	20,10	0,10	0,70	9	8,42	-0,58	0,95	4	4,61	0,61
0,21	88	90,19	2,19	0,46	19	19,24	0,24	0,71	8	8,18	0,18	0,96	4	4,52	0,52
0,22	80	82,29	2,29	0,47	18	18,45	0,45	0,72	8	7,96	-0,04	0,97	4	4,43	0,43
0,23	74	75,39	1,39	0,48	18	17,70	-0,30	0,73	8	7,75	-0,25	0,98	4	4,34	0,34
0,24	68	69,33	1,33	0,49	17	16,99	-0,01	0,74	8	7,54	-0,46	0,99	4	4,25	0,25
0,25	63	63,97	0,97	0,50	16	16,33	0,33	0,75	8	7,35	-0,65				

Źródło: opracowanie własne.

[34]

2. Interpretacja wyników testowania istotności współczynnika korelacji

Przeprowadzenie testu istotności współczynnika korelacji musi zostać zakończone przedstawieniem wniosku wynikającego z badania. Przypomnijmy tu raz jeszcze, że brak odrzucenia hipotezy zerowej interpretuje się jako brak zależności liniowej, a bardziej formalnie jako „współczynnik korelacji jest nieistotnie różny od zera”. Natomiast odrzucenie hipotezy zerowej oznacza „współczynnik korelacji jest istotnie różny od zera”, co często interpretuje się jako występowanie zależności liniowej. Jeżeli zatem ta „zależność liniowa” występuje często, dąży się do konstrukcji modelu regresji liniowej. Zwróćmy tu jednak uwagę, że odrzucenie hipotezy zerowej może nastąpić dla każdej wartości współczynnika korelacji (zob. tab. 2). Zatem czy zasadna jest konstrukcja modelu regresji dla istotnego współczynnika korelacji $r = 0,01$? Wydaje się, że zupełnie bezzasadne byłoby budowanie takiego modelu, wartość współczynnika determinacji dyskwalifikowałaby bowiem taką zależność. Istotność współczynnika korelacji wynikałaby tu nie z wartości współczynnika korelacji, lecz z wielkości próby.

Rozpatrzmy teraz przykład, w którym zaprezentowano szkic zadania 4.18 z pracy Krysickiego i in.³

1. Zadano próbę 25-elementową pochodzącą z dwuwymiarowego rozkładu normalnego.
2. Obliczono współczynnik korelacji $r = -0,1$ oraz przeprowadzono test istotności na poziomie istotności $\alpha = 0,05$.

Tabela 4. Dane z zadania 4.18

	5	10	15	20	25
5			1		
6		1	3	2	
7	1	3	3	3	1
8		2	3	1	
9			1		

Źródło: opracowanie własne na podstawie: W. Krysicki i in., *Rachunek prawdopodobieństwa i statystyka matematyczna w zadaniach*, Wydawnictwo Naukowe PWN, Warszawa 2000, s. 174.

Wartość statystyki $t_e = -0,482$ nie znalazła się w zbiorze krytycznym $K = (-\infty; -2,069) \cup (2,069; \infty)$, co oznacza brak podstaw do odrzucenia hipotezy, że cechy X i Y są nieskorelowane.

Przeanalizujmy teraz modyfikację tego zadania polegającą na tym, że liczba wszystkich obserwacji zostanie 20-krotnie zreplikowana, przez co otrzymana została próba zaprezentowana w tabeli 5.

³ W. Krysicki i in., wyd. cyt., s. 174.

Tabela 5. Dane z zadania 4.18 po modyfikacji

	5	10	15	20	25
5			20		
6		20	60	40	
7	20	60	60	60	20
8		40	60	20	
9			20		

Źródło: opracowanie własne.

Dla tak skonstruowanej próby współczynnik korelacji wynosi w dalszym ciągu $-0,1$. Wartość statystyki testowej (ze względu na wielkość próby) wynosi już jednak $t_e = -2,243$, co oznacza, że znajduje się w zbiorze krytycznym $K = (-\infty; -1,965) \cup (1,965; \infty)$. W związku z tym hipotezę zerową, o tym że cechy X i Y są nieskorelowane, należy odrzucić. Powstaje pytanie o zasadność budowania dla drugiego przykładu modelu regresji, skoro dla pierwotnego przykładu nie jest to możliwe. Wnikliwy badacz nie będzie miał wątpliwości co do możliwości dalszych badań przy wykorzystaniu modelu regresji, pozostaje jednak wątpliwość, czy powinien być wykorzystywany test, który mimo że pozwala uzyskać pozytywny dla badacza rezultat nie może być w żaden sposób wykorzystany. Wydaje się zatem, że zamiast testować hipotezę o braku zależności liniowej, czyli hipotezę zerową postaci

$$H_0: \rho = 0,$$

należałoby testować hipotezę zerową postaci

$$H_0: |\rho| = \rho_0.$$

Kluczowym zagadnieniem staje się tu przyjęcie wartości ρ_0 . Według autorów wartość ta powinna wynosić około $0,7$, daje to bowiem dla modelu regresji wartość współczynnika determinacji około 50% .

Formalny zapis przedstawianego testu jest następujący⁴:

$$H_0: |\rho| = \rho_0, \quad H_1: |\rho| > \rho_0.$$

Zwróćmy tu uwagę, że hipoteza alternatywna może także być lewostronna, natomiast z charakteru testu wynika zasadność używania wyłącznie testu prawostronnego. Weryfikacja hipotezy zerowej dokonywana jest z wykorzystaniem statystyki

$$T = (Z - z_0)\sqrt{n-3},$$

⁴ Tamże.

gdzie

$$Z = \frac{1}{2} \ln\left(\frac{1+r}{1-r}\right) \quad \text{oraz} \quad z_0 = \frac{1}{2} \ln\left(\frac{1+\rho_0}{1-\rho_0}\right).$$

Statystyka T ma dla prób większych od 10 rozkład normalny $N(0,1)$.

Tabela 6. Współczynniki korelacji i wielkość próby gwarantująca odrzucenie hipotezy zerowej

r	n	r	n	r	n
0,71	6900	0,81	44	0,91	10
0,72	1676	0,82	36	0,92	9
0,73	725	0,83	30	0,93	8
0,74	397	0,84	25	0,94	7
0,75	247	0,85	22	0,95	6
0,76	167	0,86	18	0,96	6
0,77	120	0,87	16	0,97	5
0,78	89	0,88	14	0,98	5
0,79	69	0,89	12	0,99	4
0,80	54	0,90	11		

Źródło: opracowanie własne.

W tabeli 6 przedstawiono wielkości próby niezbędne do odrzucenie hipotezy zerowej dla różnych współczynników korelacji przy założeniu, że poziom istotności wynosi 0,05, a wartość $\rho_0 = 0,7$. Warto zauważyć, że nie ma takiej liczebności próby, która umożliwiłaby uzyskanie istotności współczynnika korelacji, dla którego $r \leq 0,7$.

Zakończenie

Tradycyjnie wykorzystywany test istotności współczynnika korelacji pozwala na odrzucenie hipotezy zerowej w wyniku pobrania odpowiednio dużej próby. To poprawne pod względem formalnym „działanie” testu istotności ma jednak dwójakiego rodzaju skutki. Z jednej strony umożliwia konstrukcję łatwiejszego jak się wydaje sposobu weryfikacji hipotezy zerowej, pozwala bowiem na porównywanie (przy określonych założeniach) wyłącznie wielkości próby, na podstawie której liczony był współczynnik korelacji z wartością uzyskaną jako wynik pewnej funkcji współczynnika korelacji. Z drugiej strony wprowadza problemy interpretacyjne, które powstają w sytuacji, gdy obliczenie i testowanie współczynnika korelacji jest tylko drogą do budowy modelu regresji. W takiej sytuacji autorzy

zalecają korzystanie z mniej rozpowszechnionego testu weryfikującego występowanie zależności liniowej pomiędzy zmiennymi.

Literatura

- Ekonometria. Metody, zadania, przykłady*, red. J. Dziechciarz, Wydawnictwo AE we Wrocławiu, Wrocław 2002.
- Hellwig Z., *On the Measurement of Stochastic Dependence*, „Zastosowania Matematyki” 1969, X.
- Krysicki W., Bartos J., Dyczka W., Królikowska K., Wasilewski M., *Rachunek prawdopodobieństwa i statystyka matematyczna w zadaniach*, Wydawnictwo Naukowe PWN, Warszawa 2000.
- Ostasiewicz S., Rusnak Z., Siedlecka U., *Statystyka. Elementy teorii i zadania*, Wydawnictwo AE we Wrocławiu, Wrocław 2001.
- Steczkowski J., Zeliaś A., *Metody statystyczne w badaniach zjawisk jakościowych*, Wydawnictwo AE, Kraków 1997.