Projekt 3

Ekonometria finansowa i dynamiczna

Informatyka i Ekonometria rok V semestr I

Grzegorz Bylina

Kamila Kucharska

Andrzej Miczek

Spis treści

Wprowadzenie

Celem projektu jest za pomocą odpowiednich symulacji zbadać własności estymatorów parametrów modelu ARMA w przypadku błędnej specyfikacji oraz wpływu tego zjawiska na dokładność prognoz.

Model ARMA jest podstawowym narzędziem analizy szeregów czasowych, używanym do opisu dynamiki danych i prognozowania. W praktyce istotnym problemem jest wpływ błędnej specyfikacji modelu na estymację parametrów oraz dokładność prognoz. Celem projektu jest zbadanie, jak niedopasowanie rzędu modelu wpływa na własności estymatorów parametrów oraz prognozy.

Przeanalizujemy dwie sytuacje: gdy dane pochodzą z modelu AR(1), a estymacja odbywa się przy użyciu modelu AR(3), oraz odwrotnie – dane z AR(3), a dopasowany model to AR(1). Procedura obejmie: generowanie danych na podstawie ustalonych parametrów modelu AR(1) dla długości szeregu N (50, 100), estymację parametrów modelu AR(3) oraz analizę autokorelacji reszt. Kroki te zostaną powtórzone 1000 razy, co pozwoli na ocenę rozkładu estymatorów, ich obciążenia oraz wyników testów istotności. Następnie zostanie przeprowadzona analiza odwrotna (dane z AR(3), estymacja dla AR(1)), aby porównać wyniki. Dodatkowo każda symulacja obejmie prognozowanie kolejnej wartości szeregu czasowego (N+1), co umożliwi ocenę wpływu błędnej specyfikacji na dokładność prognoz.

Postać modelu

W ramach projektu wykorzystane zostaną modele autoregresyjne AR(1) i AR(3), których specyfikacja matematyczna jest następująca:

Model AR(1):

$$Xt = \alpha + \phi_1 * Xt\text{-}_1 + \varepsilon t$$

gdzie

α - wyraz wolny,

ϕ_1 - parametr modelu,

εt - składnik losowy o rozkładzie normalnym $N(0,\sigma^2)$.

Model AR(3):

$$Xt = \alpha + \phi_1 * Xt\text{-}_1 + \phi_2 * Xt\text{-}_2 + \phi_3 * Xt\text{-}_3 + \varepsilon t$$

gdzie

α - wyraz wolny,

ϕ_1, ϕ_2, ϕ_3 - parametry modelu,

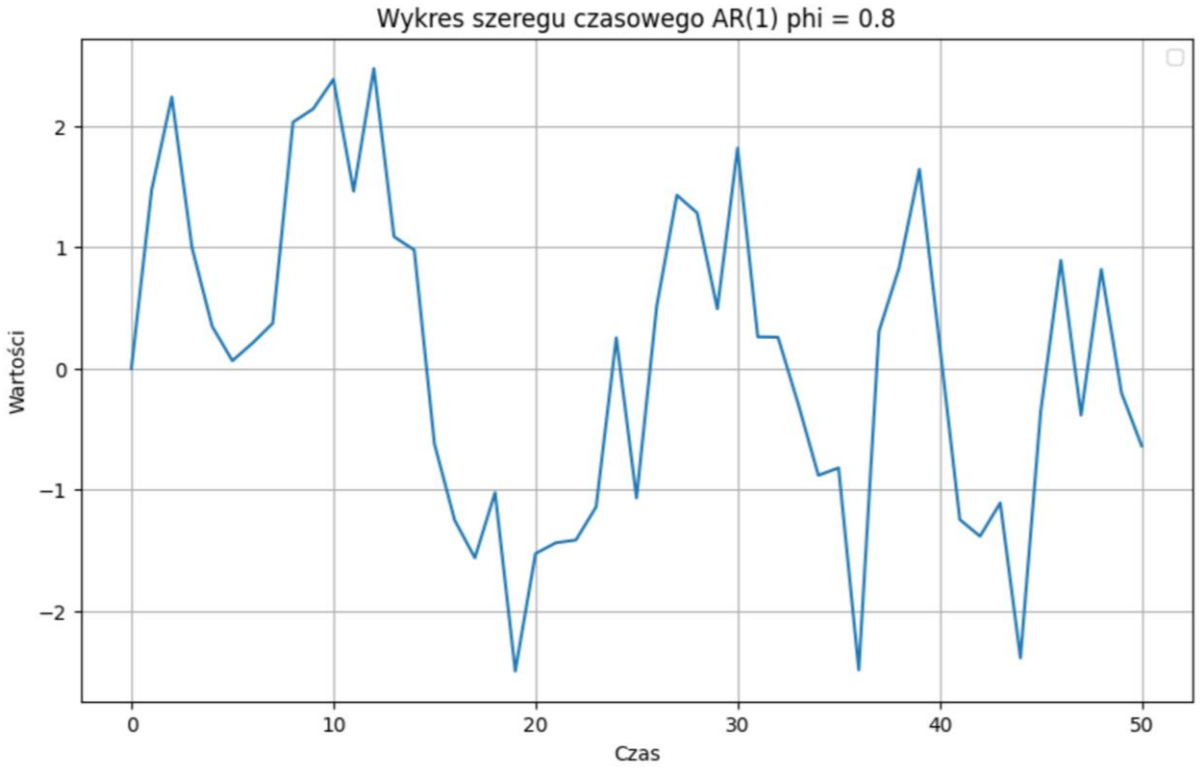
εt - składnik losowy o rozkładzie normalnym $N(0,\sigma^2)$.

Ilustracja 1 Początek przykładowej otrzymanej tabeli z 1000 szeregów czasowych

Lag1	Lag1_SE	LB_Stat_Lag1	LB_PValue_Lag1	Lag1_Bias	Normality_PValue	Lag1_pvalue	Stationarity_PValue	Actual_value	Next_Value_Estimate
0.607282	0.025120	85.533439	2.278163e-20	-0.192718	0.901641	4.071120e-129	6.150873e-22	-1.577262	-0.984157
0.593768	0.025420	73.424404	1.045652e-17	-0.206232	0.598668	1.123422e-120	0.000000e+00	-0.062010	-0.015064
0.593456	0.025428	82.230267	1.211208e-19	-0.206544	0.082730	1.792953e-120	0.000000e+00	-0.438952	-0.288516
0.606799	0.025119	89.499738	3.066791e-21	-0.193201	0.716786	6.257460e-129	5.072710e-22	0.988829	0.610626
0.597373	0.025335	83.261551	7.188425e-20	-0.202627	0.249695	6.380239e-123	0.000000e+00	-0.116397	-0.039345

Dla każdego szeregu wyliczono następujące statystyki, które później zostaną poddane analizie.

Wykres 1 Przykładowy wygenerowany szereg czasowy AR(1) z parametrami $\phi = 0.8$ $N = 50$.



Oszacowanie parametrów AR(3) na podstawie wartości AR(1)

	phi	N=50	N=100	N=1000
MAPE (%)	0,8	211,79	66,69	30,05
	-0,8	206,66	213,82	182,93
	0,1	214,98	131,77	117,47
	-0,1	171,37	158,75	137,21
RMSE	0,8	0,35	0,30	0,27
	-0,8	2,43	2,35	2,39
	0,1	0,78	0,73	0,70
	-0,1	0,94	0,88	0,84
Odsetek obciążonych estymatorów dla I opóźnienia	0,8	73%	62%	11%
	-0,8	73%	60%	11%
	0,1	74%	61%	11%
	-0,1	74%	61%	11%
Odsetek obciążonych estymatorów dla II opóźnienia	0,8	75%	67%	21%
	-0,8	79%	69%	23%
	0,1	73%	62%	13%
	-0,1	74%	62%	13%
Odsetek obciążonych estymatorów dla III opóźnienia	0,8	75%	62%	12%
	-0,8	72%	59%	11%
	0,1	72%	62%	11%
	-0,1	72%	61%	11%
Odsetek istotnych parametrów dla II opóźnienia	0,8	4%	4%	5%
	-0,8	5%	5%	5%
	0,1	5%	5%	6%
	-0,1	5%	5%	5%
Odsetek istotnych parametrów dla III opóźnienia	0,8	6%	5%	5%
	-0,8	5%	5%	4%
	0,1	5%	5%	5%
	-0,1	5%	5%	6%
Odsetek rozkładów normalnych	0,8	94%	95%	95%
	-0,8	93%	95%	95%
	0,1	94%	95%	95%
	-0,1	94%	95%	95%
Odsetek reszt, w których występuje autokorelacja	0,8	77%	89%	100%
	-0,8	78%	89%	100%
	0,1	79%	89%	100%
	-0,1	80%	90%	100%
Różnica wartości parametrów przy pierwszym opóźnieniu	0,8	-0,052	-0,018	-0,002
	-0,8	-0,013	-0,002	0,000
	0,1	-0,032	-0,010	-0,001
	-0,1	-0,028	-0,008	-0,001

W tym przypadku dane generowano z modelu AR(1), a następnie estymowano parametry przy użyciu modelu AR(3). Analizowano wartości MAPE, RMSE, autokorelację reszt, procent estymatorów obciążonych oraz wyniki testów istotności parametrów w zależności od długości szeregu czasowego N (50, 100, 1000) oraz różnych wartości ϕ (0,8 , -0,8 , 0,1 , -0,1).

W pierwszym kroku zdecydowano się przeanalizować jakość prognozy badanych modeli, obliczając dla każdej kombinacji długości szeregów czasowych i parametrów ϕ, przy użyciu średniego bezwzględnego błędu procentowego. Uzyskane wartości MAPE są wysokie, szczególnie przy krótszych szeregach czasowych (N=50), co wskazuje na duży błąd prognozy wynikający z niedopasowania modelu.

Dla małej liczby obserwacji (N=50), błędy prognoz są znacząco wyższe, niezależnie od wartości ϕ, co wskazuje na trudności w modelowaniu krótkich szeregów czasowych. Dla większych próbek (N=100, N=1000), MAPE znacząco maleje, co pozwala stwierdzić występowanie poprawy jakości prognoz wraz z zwiększeniem ilości danych. Przy ϕ=0.8 obserwuje się większe błędy (wyższe MAPE), co może wynikać z silnej zależności między kolejnymi wartościami, co oznacza większą autokorelację, którą trudniej dopasowuje się w przypadku błędnej specyfikacji modelu.

Wzrost N powoduje obniżenie RMSE w małym stopniu, jednakże nadal wskazuje to na poprawę dopasowania modelu przy większej liczbie obserwacji. RMSE zmienia się różnie, wraz ze zmianą $|\phi|$, co sugeruje, że modele mają różną dokładność dla różnej autokorelacji danych.

Dla pierwszego opóźnienia odsetek obciążonych estymatorów jest zawsze wysoki (100%), co odzwierciedla niedopasowanie modelu. Dla drugiego i trzeciego opóźnienia odsetek maleje wraz z N, co sugeruje, że model lepiej rozpoznaje brak tych opóźnień przy dłuższych szeregach czasowych. Wartości są wyższe przy mniejszych próbkach (N=50, N=100), co wskazuje na większe problemy z obciążeniem parametrów w krótszych szeregach. Dla większych próbek (N=1000) odsetek obciążonych estymatorów znacznie maleje, co potwierdza, że większe próbki redukują błędy estymacji.

Wartości odsetek istotnych parametrów dla II i III opóźnienia są bardzo niskie (4-6%), co wskazuje, że w modelach błędnie dopasowanych jak w badanym przypadku, błędnego dopasowania do modelu AR(3) parametry są rzadko istotne statystycznie, co jest logicznym wnioskiem wynikającym, z ów błędu.

Niezależnie od liczby obserwacji, wyniki wskazują na trudności z poprawnym wykrywaniem istotności parametrów w przypadku błędnej specyfikacji modelu.

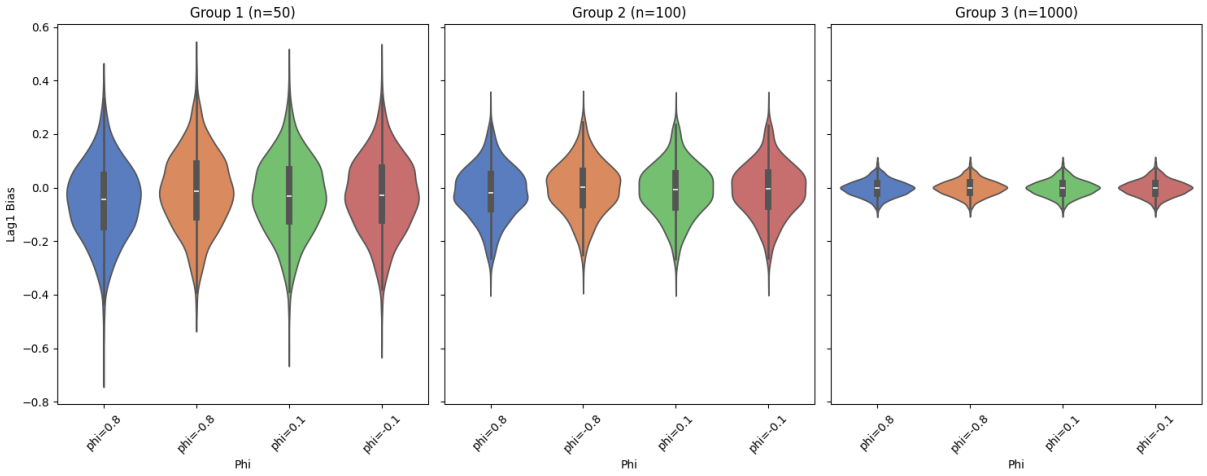
Wysoki procent odsetek rozkładów normalnych (93-95%) wskazuje, że nawet przy błędnej specyfikacji modele generują reszty zbliżone do rozkładu normalnego.

Wyniki są stabilne niezależnie od wartości ϕ i liczby obserwacji.

Wartości autokorelacji reszt wzrastają z długością szeregu, co wskazuje na problem z niewystarczającym uwzględnieniem zależności czasowych w modelu.

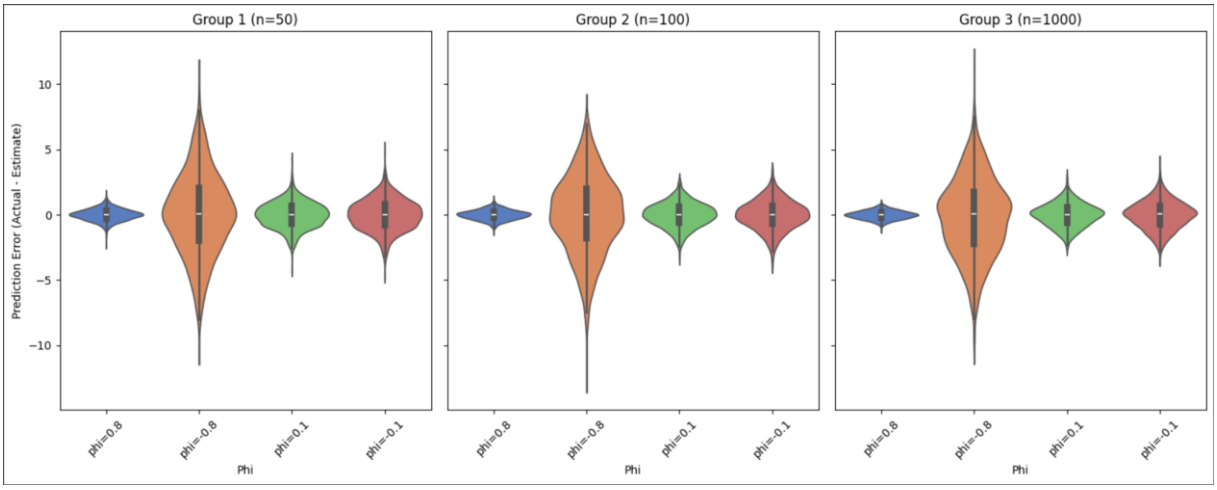
Wraz ze wzrostem próby od N=50 do N=1000, odsetek reszt wykazujących autokorelację wzrasta do 100%, niezależnie od wartości ϕ. To oznacza, że przy większych próbach autokorelacja staje się bardziej oczywista i trudniejsza do przeoczenia. Dla małych próbek różnice są większe, co wskazuje na większe ryzyko błędnych wniosków w analizie krótkich szeregów czasowych.

Wykres 2 Wykres skrzypcowy rozkładów wartości Lag1_Bias dla modeli AR(3).



Wykres przedstawia rozkład różnic wartości parametru oryginalnego phi AR(1), a wartością przy pierwszym opóźnieniu modelu AR(3). Jak widać wpływ wartości samego phi jest znikomy. Mediana wszystkich rozkładów oscyluje wokół wartości 0. Kluczowe znaczenie ma natomiast długość badanego szeregu czasowego. Zmiana N=50 na 1000 znacząco redukuje wariancję rozkładu. Wyniki pokrywają się z początkowymi przypuszczeniami, im więcej danych tym model się może lepiej dopasować.

Wykres 3 Wykres skrzypcowy rozkładów wartości różnicy wartości rzeczywistej AR(1), a wartości wyestymowanej AR(3).



Najlepsze estymacje zrobił model ϕ = 0.8, a najgorsze z ϕ= -0,8, które również ma największą wariancję ze wszystkich. Można zauważyć trend wraz ze wzrostem N, stożki rozkładów stają się nieco wyższe i węższe - więcej wartości oscyluje bliżej zera. Czyli wzrost długości szeregu czasowego może wpłynąć na poprawę estymacji.

Oszacowanie parametrów AR(1) na podstawie wartości AR(2)

	phi1	N=50	N=100	N=1000
MAPE (%)	0,8	62,8	54,98	41,8
	-0,8	211,86	182,68	169,97
	0,1	160,14	115,82	93,9
	-0,1	130,05	117,52	106,55
RMSE	0,8	0,447	0,462	0,457
	-0,8	1,889	1,829	1,815
	0,1	0,754	0,755	0,727
	-0,1	0,877	0,867	0,827
Odsetek obciążonych estymatorów dla I opóźnienia	0,8	100%	100%	100%
	-0,8	100%	100%	100%
	0,1	78%	84%	100%
	-0,1	67%	55%	24%
Odsetek rozkładów normalnych	0,8	95%	96%	95%
	-0,8	94%	95%	95%
	0,1	94%	95%	96%
	-0,1	95%	95%	96%
Odsetek reszt, w których występuje autokorelacja	0,8	0%	0%	0%
	-0,8	11%	5%	0%
	0,1	47%	28%	0%
	-0,1	69%	67%	0%
Różnica wartości parametrów przy pierwszym opóźnieniu	0,8	-0,212	-0,201	-0,194
	-0,8	0,138	0,136	0,127
	0,1	0,006	0,017	0,024
	-0,1	0,05	0,061	0,067

W tym przypadku dane generowano z modelu AR(3), a do estymacji zastosowano model AR(1). W tym przypadku przeanalizowano wskaźniki, takie jak MAPE, RMSE, obciążenia estymatorów oraz autokorelację reszt. W każdym przypadku ϕ_2 = -0,3 oraz ϕ_3 = -0,2.

Dla ϕ_1 =0,8 średni procentowy błąd prognozowania (MAPE) maleje wraz ze wzrostem N, co oznacza, że większa ilość danych pozwala na dokładniejsze prognozy. Jednak nawet przy N=1000 MAPE wynosi około 41.8%, co świadczy o znaczącej niedokładności prognoz wynikającej z niedopasowania modelu. Dla ϕ_1 =−0,8 błąd prognozy jest znacznie większy (nawet powyżej 200% dla N=50), co wskazuje, że niedopasowanie modelu szczególnie mocno wpływa na prognozowanie w przypadku silnie ujemnej autokorelacji. Dla ϕ_1=0.1 i ϕ_1=−0.1 prognozy są relatywnie dokładniejsze, co sugeruje, że błędna specyfikacja modelu jest mniej szkodliwa przy słabszej autokorelacji w danych.

RMSE również maleje, choć w bardzo niewielkim stopniu wraz ze wzrostem N, co odzwierciedla większą stabilność estymacji przy większej ilości danych. Niemniej ϕ_1=0.8 daje najniższe RMSE, co oznacza lepsze dopasowanie modelu AR(1) do danych z dodatnią autokorelacją. ϕ_1=−0.8 charakteryzuje się największym RMSE, wskazując na problemy z odwzorowaniem silnej ujemnej autokorelacji przez model AR(1) w przypadku błędnej specyfikacji modelu..

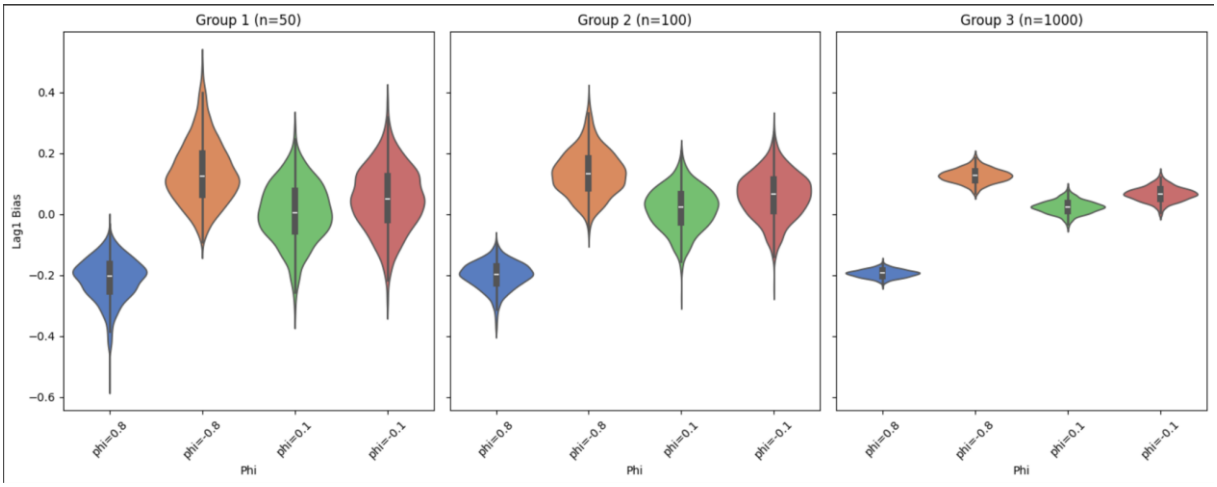
Dla wszystkich wartości ϕ_1, odsetek obciążonych estymatorów wynosi 100% przy N=50, co pokazuje, że dla krótkich szeregów czasowych błędna specyfikacja modelu prowadzi do silnego obciążenia estymatorów. Przy N=1000, obciążenie jest widoczne dla większych wartości ϕ_1=0,8 i ϕ_1=-0,8.

Większość estymatorów ma rozkład zbliżony do normalnego (odsetek ponad 90%) dla wszystkich przypadków, niezależnie od wartości ϕ_1 i N.

W przypadku autokorelacji można wysunąć wniosek wynikający również z poprzedniej sytuacji, że występowanie autokorelacji w przypadku błędnej specyfikacji modelu, może zależeć właśnie od ów błędu, a nie liczby danych i wartości ϕ_1.

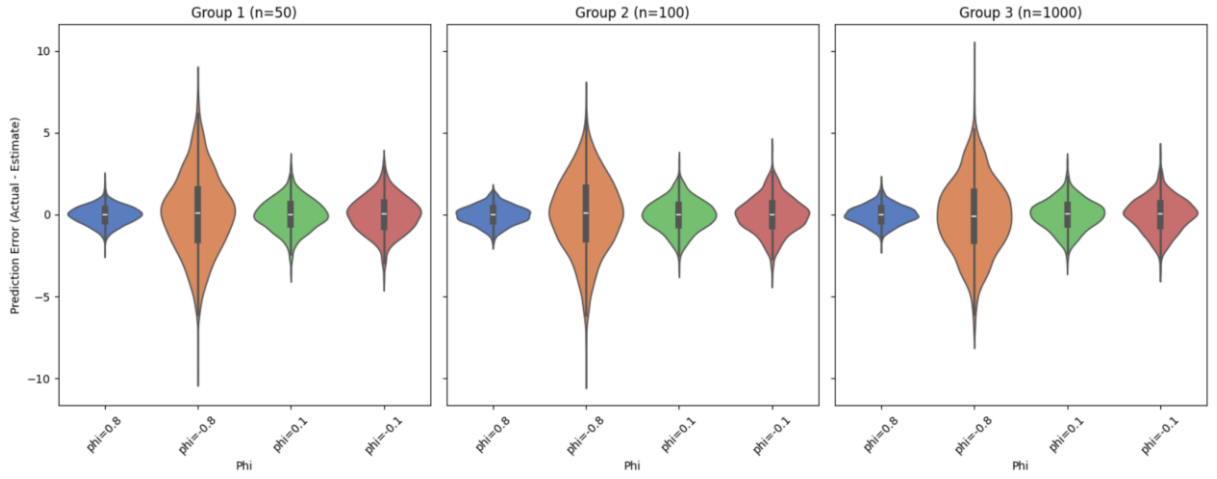
Wartości różnic w estymatorach są większe dla silniejszych autokorelacji (ϕ_1=0,8 i ϕ_1=−0,8), przy słabszej autokorelacji (ϕ_1=0,1 i ϕ_1=−0,1), różnice są mniejsze, co wskazuje na mniejsze problemy z estymacją.

Wykres 4 Wykres skrzypcowy rozkładów wartości Lag1_Bias dla modeli AR(1).



Ciekawym spostrzeżeniem w od niesieniu do wykresów AR(3) jest to, że mediany nie oscylują wokół 0 i jest widoczny pewien bias. Również tutaj zwiększanie N znacząco wpływa na redukcję wariancji rozkładów.

Wykres 5 Wykres skrzypcowy rozkładów wartości różnicy wartości rzeczywistej AR(3), a wartości wyestymowanej AR(1).



Tak samo jak poprzednio najlepsze estymacje zrobił model ϕ_1 = 0,8, a najgorsze z ϕ_1 = -0,8. Ciężko na podstawie wykresu ocenić wpływ zmiany N.

Podsumowanie

W przeprowadzonej analizie porównano wyniki estymacji parametrów modeli AR(1) i AR(3) przy różnych długościach szeregów czasowych (N=50, N=100, N=1000) oraz różnych wartościach ϕ (0,8, -0,8, 0,1, -0,1). Celem było zbadanie jakości prognoz oraz właściwości estymatorów w zależności od w założeniu błędnej specyfikacji modelu oraz liczby danych.

Wyniki wskazują na znaczący wpływ długości szeregu czasowego na jakość prognoz. Dla krótkich szeregów (N=50) błędy prognoz, sugerują trudności w modelowaniu przy małej liczbie obserwacji. W szczególności, przy małych próbkach, błędy prognoz są wyraźnie wyższe, niezależnie od wartości ϕ, co potwierdza, że krótkie szeregi czasowe utrudniają dokładną estymację parametrów i prognozowanie. Z kolei dla dłuższych szeregów (N=100, N=1000) MAPE i RMSE znacznie się poprawiają, co pokazuje, że większa liczba danych pozwala na lepsze dopasowanie modelu i dokładniejsze prognozy. Wzrost długości próby prowadzi do stabilizacji wyników i zmniejszenia błędów estymacji.

Wartości ϕ mają również istotny wpływ na wyniki. Dla większych wartości autokorelacji (ϕ = 0,8) błędy prognoz są wyższe, co wskazuje na trudności w dopasowaniu modelu do danych o silnej autokorelacji. W przypadku silnej ujemnej autokorelacji (ϕ = -0,8) prognozy są szczególnie niedokładne, zwłaszcza przy krótkich próbkach, gdzie błąd prognozy może przekraczać 200%. Wartości ϕ = 0,1 i ϕ = -0,1 generują dokładniejsze prognozy, co sugeruje, że błędna specyfikacja modelu ma mniejszy wpływ przy słabszej autokorelacji w danych.

W miarę wzrostu liczby danych, obciążenie estymatorów maleje, co sugeruje, że większe próbki ułatwiają prawidłowe dopasowanie parametrów.

Autokorelacja reszt zmienia się w zależności od błędnej specyfikacji modelu, na co wskazują prawie odwrotnie proporcjonalnie wyniki dla obu badanych sytuacji, na co może mieć wpływ problem z wykrywaniem autokorelacji.

Podsumowując, analiza pokazuje, że błędna specyfikacja modelu, zarówno przy estymacji AR(1) na danych AR(3), jak i odwrotnie, prowadzi do istotnych trudności w prognozowaniu, szczególnie w przypadku krótkich szeregów czasowych. Zwiększenie liczby danych poprawia jakość prognoz i zmniejsza obciążenie estymatorów, ale nie eliminuje całkowicie problemów związanych z niewłaściwą specyfikacją modelu, co utwierdza w przekonaniu o konieczności prawidłowej budowy modelu.