Prognozowanie i symulacje
Projekt ekonometryczny

Autor/Autorzy	Maciej Odziemczyk, Karolina Stępień
Tytuł	*Trade-off* między wnioskowaniem a predykcją – problemy identyfikacyjne

Informacje o artykule będącym inspiracją dla badania

Tytuł	CAUSALITY Models, Reasoning, and Inference
Autor/Autorzy	Judea Pearl
Journal/Miejsce publikacji	Cambridge University Press
Rok	2000 (I edycja), 2009 (II edycja)
Zakres stron	cała książka
Tematyka, problemy i cele badawcze	Książka poświęcona jest m. in. obciążeniom wywołanym złą identyfikacją modelu. Zaprezentowane zostały grafy przyczynowe obrazujące procesy generujące dane.
Główne wnioski	Grafy przyczynowe ułatwiają zrozumienie problemów identyfikacyjnych. Okazuje się że w modelowaniu występuje wiele pułapek.
Metodyka badawcza	Książka przedstawia teorie, rozważania, wyprowadzenia, grafy, modele strukturalne i dowody.
Dane	Praca nie ma charakteru empirycznego.
Dlaczego wybrano właśnie ten artykuł?	Bez zrozumienia problemów identyfikacyjnych modelowanie ekonometryczne w sensie wnioskowania statystycznego przypomina loterię - albo wnioski są poprawne, albo nie. Warto poszerzyć swoją wiedzę poza problemy estymacyjne. Ponadto ciekawym wydaje się temat predykcji w świetle omawianych problemów.

Podstawowe informacje o badaniu

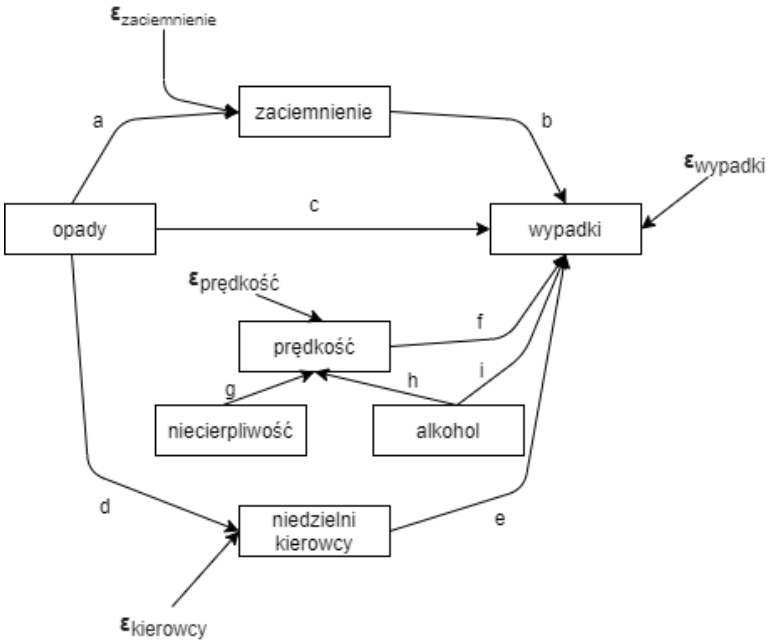
Problem badawczy	Badanie poświęcone jest wpływowi mediatorów, *confounderów* oraz instrumentów na wyniki estymacji modeli liniowych na przykładzie wymyślonego procesu generującego dane.
Metodyka badawcza	Na podstawie wymyślonego grafu przedstawiającego proces generujący dane, który zawierał zmienną instrumentalną, pominiętą (nieobserwowalną) oraz mediatora, wylosowano milion obserwacji (z czego 100 000 to próba *out of sample*), początkowo z zastosowaniem stałych parametrów strukturalnych. Na ich podstawie zbudowano modele regresji liniowej, przeprowadzono diagnostykę i prognozę. Na kolejnym etapie, w celu uogólnienia eksperymentu, przeprowadzono symulacje dla losowych parametrów strukturalnych w zdefiniowanym zakresie i ponownie przeprowadzono prognozy.
Dane	Milion obserwacji z procesu generującego dane (błędy losowe z rozkładu N(0,1)), początkowo ze stałymi parametrami strukturalnymi,

	a następnie z losowanymi z określonych zbiorów.
Główne wnioski	Wnioskiem jest możliwe występowanie *trade-offu* między poprawnym wnioskowaniem statystycznym a predykcją. Osoba budująca model na podstawie dopasowanego R^2 oraz błędów prognozy wybierze do predykcji model z obciążonymi parametrami, ponieważ okazuje się on być do tego lepszy niż model prawidłowy.

Opis badania

W celu zbadania problemów identyfikacyjnych możliwych do powstania w przypadku obecności mediatorów, zmiennych pominiętych i instrumentalnych na wyniki estymacji modeli regresji liniowej wygenerowano dane według procesu przedstawionego na rysunku 1.

Rys. 1. Graf przedstawiający proces generujący dane



Źródło: Opracowanie własne.

Aby lepiej rozumieć zależności między danymi i ułatwić interpretację modeli, przypisano przykładowe znaczenie każdej ze zmiennych (patrz: tabela 1). Każda ze zmiennych opisuje stan świata na danym obszarze w określonym czasie.

Tab. 1. Zmienne i stałe w modelu

zmienna objaśniana	**wypadki**	liczba wypadków i kolizji drogowych
zmienne objaśniające	**opady**	opady deszczu (l/m^2)
	zaciemnienie (mediator)	średnia widoczność na drodze (max - ciemność, min -maksymalna widoczność)
	niedzielni_kierowcy (mediator)	stosunek liczby słabych/mało doświadczonych kierowców do liczby kierowców na drodze
	predkosc	liczba kierowców, którzy przekroczyli prędkość, w stosunku do liczby kierowców ogółem
	alkohol (zmienna pominięta)	stosunek liczby kierowców, którzy jechali po wpływem alkoholu, do liczby kierowców ogółem
	niecierpliwosc (zmienna instrumentalna)	liczba niecierpliwych kierowców w stosunku do liczby kierowców na drodze
	ε...	błędy losowe z rozkładu N(0,1)
	a-f	parametry strukturalne procesu generującego dane

Źródło: Opracowanie własne.

Na potrzeby analizy wygenerowano milion obserwacji z uwzględnieniem zależności przedstawionych na powyższym grafie, a następnie podzielono zbiór na próbę treningową (900 000 obserwacji) i testową (100 000 obserwacji). W pierwszym etapie przyjęto, że parametry *a-c* oraz *e-i* mają wartość 1 i parametr *d* jest równy -1, a następnie oszacowano modele regresji liniowej. Ze względu na obciążoność parametru przy zmiennej *predkosc* w jednym z modeli (*model6*) wykorzystano zmienną instrumentalną *niecierpliwosc* za pomocą dwustopniowej MNK. Wyniki estymacji modeli na próbie treningowej przedstawia tabela 2.

Tab. 2. Wyniki estymacji modeli ze zmienną objaśnianą *wypadki*

	model1	model2	model3	model4	model5	model6
Intercept	-0.002	-0.003	-0.002	-0.003	0.001	-0.001
	(0.003)	(0.003)	(0.003)	(0.003)	(0.001)	(0.003)
fitted						0.999***
						(0.003)
niedzielni_kierowcy			1.006***	1.003***	1.001***	1.003***
			(0.003)	(0.003)	(0.001)	(0.003)
opady	1.004***	-0.003	2.009***	1.001***	1.001***	1.002***
	(0.003)	(0.004)	(0.004)	(0.005)	(0.002)	(0.004)
predkosc					1.333***	
					(0.001)	
zaciemnienie		1.007***		1.003***	0.998***	1.002***
		(0.003)		(0.003)	(0.001)	(0.003)
Observations	900,000	900,000	900,000	900,000	900,000	900,000
R2	0.100	0.201	0.201	0.301	0.834	0.401
Adjusted R2	0.100	0.201	0.201	0.301	0.834	0.401
Residual Std. Error	3.005 (df=899998)	2.832 (df=899997)	2.832 (df=899997)	2.649 (df=899996)	1.293 (df=899995)	2.452 (df=899995)
F Statistic	100361.835*** (df=1; 899998)	113388.019*** (df=2; 899997)	113106.571*** (df=2; 899997)	129255.800*** (df=3; 899996)	1127146.565*** (df=4; 899995)	150744.414*** (df=4; 899995)

Note: *p<0.1; **p<0.05; ***p<0.01

Źródło: Opracowanie własne.

Dopasowanie modelu z uwzględnieniem jedynie zmiennej objaśniającej *opady* jest niskie (0,100). Po dodaniu zmiennej *zaciemnienie*, mediator (którym dla *opady* jest *zaciemnienie*) ściąga wpływ opadów do zera, przez co *opady* stają się zmienną nieistotną, co nie jest zgodne z prawdą.

Model3 uwzględniający zmienną *niedzielni_kierowcy* zamiast *zaciemnienie* pokazuje, że *opady* obciążone są w drugą stronę – efekt jest przeszacowany. Zatem w przypadku, w którym zmienna posiada dwa mediatory i w modelu uwzględniony jest tylko jeden, obciążenie mediowanej zmiennej zależy od znaku, z jakim przenoszony jest efekt przez pominiętego mediatora. Uwzględnienie obu mediatorów i zmiennej mediowanej (*model4*) skutkuje wartością estymatorów zgodną z parametrami strukturalnymi procesu generującego dane. W przypadku tego modelu dopasowanie do danych jest niskie (R^2=0,301), co jest wynikiem pominięcia istotnej zmiennej (*predkosc*).

Zakładając, że zmienna *alkohol* jest nieobserwowalna, uwzględnienie *predkosc* w modelu stworzy problem równoczesności, tj. zmienna *predkosc* staje się zmienną endogeniczną i nie jest możliwe oszacowanie prawdziwego parametru (*model5*). *Model5* pomimo najlepszego dopasowania do danych oraz najniższego RMSE nie jest modelem, na podstawie którego można przeprowadzić poprawne wnioskowanie statystyczne z powodu obciążenia parametru zmiennej *predkosc*, co wynika z problemu równoczesności będącego skutkiem pominięcia istotnej, nieobserwowalnej zmiennej skorelowanej z *predkosc* i błędem z modelu.

Problem równoczesności może zostać rozwiązany za pomocą zmiennych instrumentalnych. W przypadku rozważanego procesu generującego dane instrumentem zmiennej *predkosc* może być zmienna *niecierpliwosc*. Niecierpliwość jest poprawnym instrumentem, ponieważ nie jest skorelowana z błędem losowym (pominięcie jej nie wpływa na zwiększenie reszt modelu, ponieważ nie ma ona bezpośredniego wpływu na zmienną objaśnianą), wpływa ona na endogeniczną zmienną objaśniającą *predkosc* i wpływ ten jest jednostronny (brak sprzężenia zwrotnego) oraz nie jest ona w żaden sposób powiązana z innymi zmiennymi objaśniającymi umieszczonymi w modelu. Estymator MZI wykorzystany do zmiennej instrumentalnej *niecierpliwosc* nie jest jednak estymatorem efektywnym (większe błędy standardowe oszacowań w *model6*), ale uzyskane estymatory są zgodne - odzwierciedlają parametry strukturalne procesu generującego dane. Okazuje się jednak, że model poprawny w kontekście wnioskowania statystycznego nie jest najlepszy pod kątem predykcji.

Wnioskuje się zatem o występowaniu sytuacji patowej i konieczności doboru modelu w zależności od potrzeb. W celu poprawnego wnioskowania statystycznego należy wybrać *model6* oszacowany przy wykorzystaniu Metody Zmiennych Instrumentalnych. W celu predykcji warto rozważyć *model5* oszacowany za pomocą MNK, którego oszacowania są obciążone (oszacowanie zmiennej *predkosc*).

Dla wszystkich modeli przeprowadzono testy diagnostyczne, które pozwoliły odpowiedzieć na pytanie, czy standardowa diagnostyka modelu jest w stanie pomóc wybrać model poprawny z punktu wnioskowania statystycznego. Przeprowadzono także prognozy dla próby *out of sample* i obliczono błędy średniokwadratowe (RMSE). Wyniki przedstawia tabela 3.

Tab. 3. Diagnostyka modeli

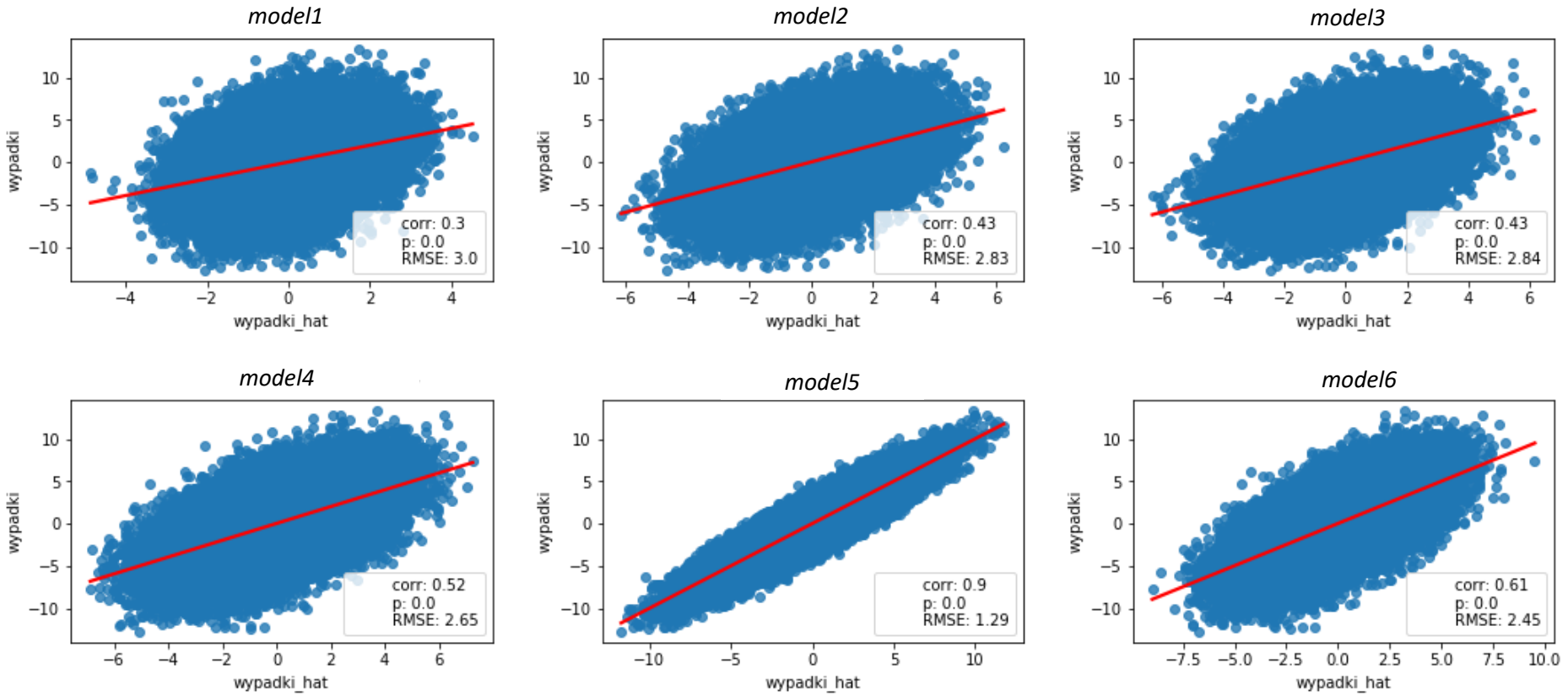
	model1	***model2***		***model3***		***model4***		***model5***		***model6***	
test RESET	p=0,2685	p=0,4378		p=0,2661		p=0,9323		p=0,3357		p=0,1988	
RMSE	3,0000	2,8307		2,8354		2,6534		1,2886		2,4525	
AIC	4534815	4427718		4428168		4307525		3015954		4168442	
BIC	4534839	4427718		4428203		4307572		3016012		4168500	
test Jarque-Bera	p=0,7784	p=0,6360		p=0,3493		p=0,8708		p=0,8750		p=0,9313	
test Breuscha-Pagana	p=0,9655	p=0,8655		p=0,6412		p=0,8758		p=0,6980		p=0,5095	
test Breuscha-Godfreya	p=0,6168	p=0,6088		p=0,8063		p=0,5920		p=0,5697		p=0,9862	
VIF		*opady*	2,0	*opady*	2,0	*opady*	3,0	*opady*	3,0	*opady*	3,0
						zaciemnienie	2,0	*zaciemnienie*	2,0	*zaciemnienie*	2,0
		zaciemnienie	2,0	*niedzielni_kierowcy*	2,0			*predkosc*	1,0	*niedzielni_kierowcy*	2,0
						niedzielni_kierowcy	2,0	*niedzielni_kierowcy*	2,0	fitted	1,0

Źródło: Opracowanie własne.

Statystyki VIF (<5) wskazują, że problem niedokładnej współliniowości nie występuje w żadnym modelu. Testy Jarque-Bera, Breuscha-Pagana i Breuscha-Godfreya nie pozwalają na odrzucenie hipotezy zerowej na powszechnie stosowanych progach istotności, zatem wnioskuje się o braku heteroskedastyczności, braku autokorelacji i o normalności reszt dla każdego modelu. Testy RESET nie pozwalają na odrzucenie hipotezy zerowej o poprawności (liniowości) formy funkcyjnej na żadnym sensownym poziomie istotności, zatem wnioskuje się o poprawnej (liniowej) formie funkcyjnej wszystkich modeli. Test RESET nie daje podstaw do odrzucenia H_0 o liniowości formy funkcyjnej, co nie dziwi, bowiem zdefiniowany proces generujący dane jest procesem liniowym. Proces generujący dane nie jest jednak czymś, co jest znane osobie modelującej, zatem na podstawie testu RESET w takim przypadku w żaden sposób nie jest możliwe zejście z błędnej ścieżki. Kryteria informacyjne AIC oraz BIC dają jednomyślny werdykt – za najlepszy wskazują model ze wszystkimi istotnymi zmiennymi, ale z obciążonym parametrem zmiennej *predkosc* (*model5*), a drugim najlepszym modelem jest model zgodny z procesem generującym dane (*model6*). RMSE dla próby *out of sample* jest najniższe dla modelu z obciążonym parametrem (bez zmiennej instrumentalnej), następny w kolejności jest *model6*. Podsumowując, *model5* i *model6* są dwoma najlepszymi modelami. Ponadto przeprowadzone testy i obliczone statystyki wyraźnie wskazują, że model bez zmiennej instrumentalnej jest wyraźnie lepszy od pozostałych. Oznacza to, że bez znajomości procesu generującego dane nie jest możliwe znalezienie prawdziwego modelu przy wykorzystaniu podstawowej diagnostyki, a budowanie modelu, znając proces generujący dane, jak wiadomo, nie ma sensu.

W celu pogłębienia analizy prognoz z perspektywy osoby wybierającej model, wygenerowano wykresy porównujące prognozowane wartości z faktycznymi realizacjami (rysunek 2).

Rys. 2. Wykresy prognoz



Źródło: Opracowanie własne.

Wnioski z wykresów są analogiczne do tych wyciągniętych na podstawie wyników regresji i testów diagnostycznych. Najlepszym pod kątem prognoz jest *model5*, korelacja prognoz z wartościami obserwowanymi jest największa (0,9) i istotna (p=0,0). Drugim najlepszym modelem jest *model6*, istotna korelacja wynosi 0,61. Z punktu widzenia prognoz nie ma znaczenia znak przenoszenia efektu przez pominiętego mediatora - drugi oraz trzeci model mają niemalże takie same wyniki. Najgorszy jest model z tylko jedną zmienną, nieuwzględniający mediatorów i fragmentu grafu ze zmienną pominiętą, co nie dziwi, bo wszystkie pominięte zmienne istotne trafiają do reszt modelu, zatem większa jest różnica wartości prognozowanych od obserwowanych i tym samym z definicji większe jest RMSE.

Aby zobaczyć, czy wyciągnięte wnioski utrzymują się na generalizacji grafu z rysunku 1, postanowiono uogólnić proces generujący dane. Symulacja polegała na losowaniu ze zwracaniem (1000 losowań) parametrów strukturalnych procesu generującego dane. Dla ułatwienia rozważanymi parametrami były jedynie liczby całkowite (uwzględnienie liczb rzeczywistych nie powinno zmienić wniosków, a jest dużo bardziej kosztowne obliczeniowo). Ponieważ rozważane są dodatnie i ujemne przeniesienia efektów przyczynowych przez mediatory, co skutkuje niedoszacowaniem lub przeszacowaniem zmiennej mediowanej (w skrajnym przypadku mamy nieistotność statystyczną zmiennej, która faktycznie jest istotna, jeżeli popełnimy problemy identyfikacyjne), oraz przeszacowanie efektu endogenicznej zmiennej objaśniającej wynikające z wystąpienia problemu zmiennej pominiętej, na symulację narzucone zostały pewne ograniczenia:

$$\{a,b,c,e,f,g,h,i\} \in Z \cap \{1,2,\ldots,10\}$$

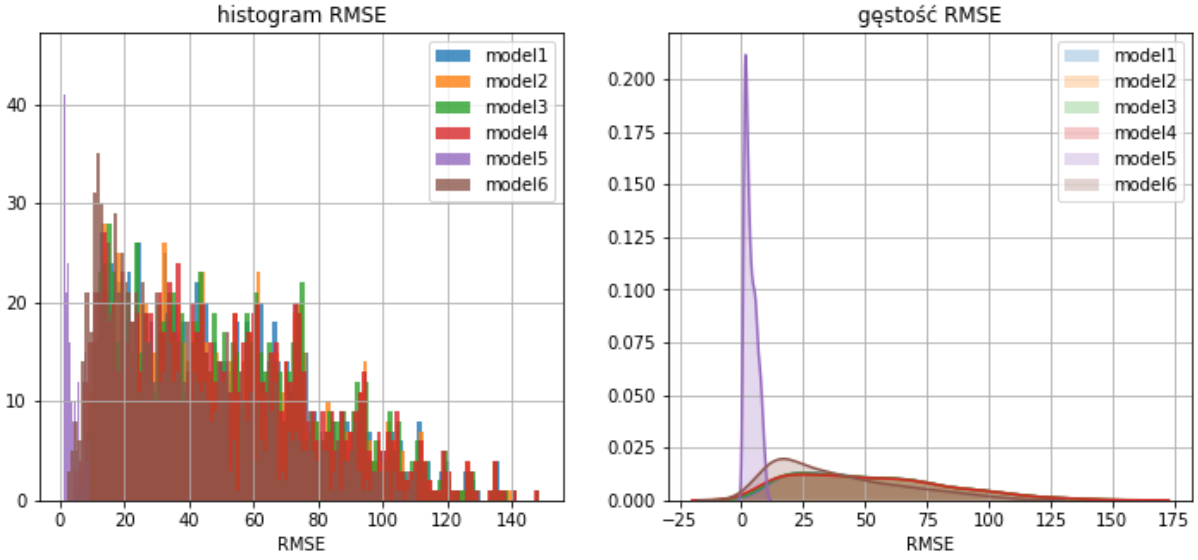
$$d \in Z \cap \{-10,-9,\ldots,-1\}$$

Na podstawie wyników poprzednich analiz postawiono hipotezę:

$$\text{Mediana}(RMSE_{model5}) < \text{Mediana}(RMSE_{model6})$$

Dla wszystkich modeli wygenerowano histogramy i oszacowano jądrową funkcję gęstości dla RMSE, co przedstawia rysunek 3.

Rys. 3. Histogram i gęstość RMSE

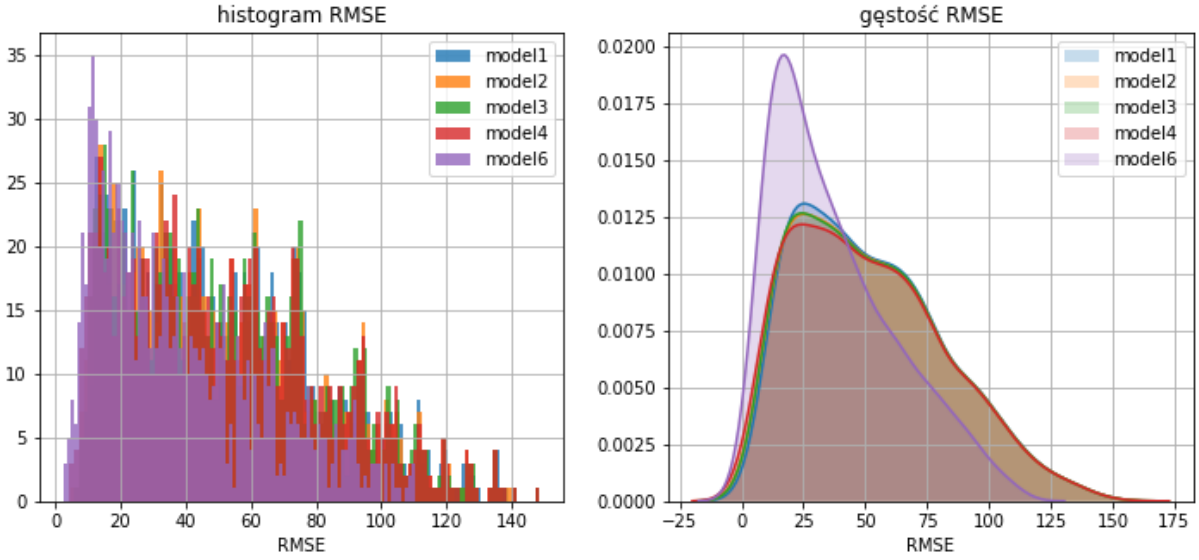


Źródło: Opracowanie własne.

Ewidentnie najbardziej skupiony wokół zera jest rozkład $RMSE_{model5}$. Reszta rozkładów jest podobna, jedynie rozkład $RMSE_{model6}$ charakteryzuje się dominantą widocznie bliższą zera

od pozostałych. Poniżej histogramy i gęstości rozkładów RMSE z pominięciem modelu piątego.

Rys. 4. Histogram i gęstość RMSE (bez *model5*)



Źródło: Opracowanie własne.

Analiza graficzna wskazuje, że mediany RMSE dla wszystkich modeli poza modelem piątym i szóstym powinny być do siebie zbliżone. Gęstość RMSE dla modelu ze zmienną instrumentalną wyraźnie różni się od pozostałych.

Do formalnego sprawdzenia równości median modelu piątego i szóstego posłużył test Wilcoxona, który jest testem nieparametrycznym niewymagającym założeń o równej wariancji porównywanych prób. Na podstawie wyników testu stwierdzono, że zależności między medianami analizowanych rozkładów RMSE dla poszczególnych modeli prezentują się następująco:

$$model1 > model2 > model4 > model6 > model5$$

oraz

$$model2 = model3$$

Podsumowując, na etapie budowania modeli zauważono, że jeżeli pomijany jest mediator przenoszący ujemny efekt, to efekt mediowanej zmiennej jest niedoszacowany, a w skrajnym przypadku równy zero i tym samym nieistotny, zatem zastosowanie procedury od ogółu do szczegółu może skutkować usunięciem z modelu zmiennej, która w rzeczywistości jest zmienną istotną. Jeżeli pomijany jest mediator przenoszący dodatni efekt, to efekt mediowanej zmiennej jest przeszacowany. W kontekście zmiennej *alkohol* zaobserwowano, że jeżeli wpływ zmiennej pominiętej na endogeniczną zmienną objaśniającą (*predkosc* w modelu piątym) jest dodatni, to parametr jest przeszacowany, a jeżeli ujemny, to jest niedoszacowany, przy założeniu, że endogeniczna zmienna objaśniająca ma dodatni wpływ na zmienną objaśnianą.

Wnioskiem z powyższych analiz jest występowanie *trade-offu* między predykcją a wnioskowaniem statystycznym. Osoba budująca model skłoni się przy okazji predykcji do modelu 5, mimo że jego oszacowania są obciążone, dlatego tak ważna jest analiza teoretyczna problemu i zastanowienie się nad sensem samego modelu. W przypadku złej identy-

fikacji, pomimo że model może być dobrze dopasowany do danych, poprawne wnioskowanie statystyczne niekoniecznie jest możliwe. Ponadto uogólnienie procesu generującego dane polegające na losowaniu parametrów *a-i* pokazało, że model bez zmiennej instrumentalnej (*model5*) jest lepszy w kontekście predykcji od prawidłowego modelu.