2020

Analiza parametrów wpływających na ceny smartfonów

KAROL DOLIŃSKI
INFORMATYKA I EKONOMETRIA

WYDZIAŁ ZARZĄDZANIA
AKADEMIA GÓRNICZO-HUTNICZA W KRAKOWIE

Spis treści

1. Wprowadzenie do tematu 3
1.1. Cel projektu 3
1.2. Hipotezy badawcze 3
1.3. Opis danych 4
1.4. Przyjęte założenia i użyte oprogramowanie 4
1.5. Statystyki opisowe 5
1.6. Wykresy zależności 8
1.7. Korelacja pomiędzy zmiennymi 9
2. Model ściśle liniowy 10
2.1. Klasyczna metoda najmniejszych kwadratów 10
2.2. Estymacja parametrów modelu ściśle liniowego 11
2.3. Wady modelu ściśle liniowego 12
2.4. Metoda Hellwiga 13
2.5. Metoda krokowa wsteczna 14
2.6. Wybór podzbioru zmiennych objaśniających 14
3. Ostateczna postać modelu 15
3.1. Estymacja finalnego modelu 16
4. Opis i testowanie własności modelu 17
4.1. Współczynnik determinacji 17
4.2. Efekt katalizy 17
4.3. Normalność rozkładu składnika losowego 18
4.4. Istotność zmiennych 19
4.4.1. Istotność pojedynczych zmiennych 19
4.4.2. Istotność całego zbioru zmiennych 20
4.5. Testy dodanych zmiennych 20
4.6. Testy pominiętych zmiennych 21

4.7. Obserwacje odstające 22
4.8. Test liczby serii 23
4.9. Test RESET 24
4.10. Testowanie heteroskedastyczności 24
4.11. Test Chowa 25
4.12. Współliniowość 26
4.13. Koincydencja 26
4.14. Interpretacja parametrów modelu 27
4.15. Predykcja 27
5. Podsumowanie 28
Bibliografia 29
Spis tabel 30
Spis rysunków 30

1. Wprowadzenie do tematu

1.1. Cel projektu

Rynek smartfonów jest jednym z bardziej dynamicznych rynków zarówno na świecie jak i w Polsce. Kilkunastu producentów prześciga się w dostosowywaniu swoich produktów do bieżących oczekiwań i potrzeb klientów. Natomiast tylko kilku z nich zdominowało rynek smartfonów. W Polsce najwięcej sprzedanych modeli można przypisać firmom: Apple, Samsung, Huawei oraz Xiaomi, mniejszymi dostawcami są OnePlus, Motorola oraz Nokia. Przez ostatnie 3 lata liczba kupionych przez Polaków urządzeń oscylowała wokół 10 mln rocznie, a odsetek społeczeństwa korzystającego ze smartfonów przekroczył 67%[1].

Ceny smartfonów na rynku są bardzo zróżnicowane, wahają się od kilkuset do nawet 8 tysięcy złotych. Przyjęło się, że wpływ na to ma jakość aparatu, pamięć, przekątna ekranu czy pojemność baterii. Celem projektu jest zbadanie, które parametry mają istotny wpływ na cenę urządzenia oraz w jakim stopniu poszczególne elementy ją kształtują.

1.2. Hipotezy badawcze

Przed szczegółową analizą ekonometryczną problemu postawiono hipotezy badawcze napisane na podstawie utartych w społeczeństwie przekonań odnośnie tego, co kształtuje ceny smartfonów.

Na wysokość ceny smartfonów wpływa:

- ✓ pamięć telefonu;
- ✓ liczba pikseli w aparacie zlokalizowanym z tyłu telefonu;
- ✓ pojemność baterii;
- ✓ przekątna ekranu.

[1] Główny Urząd Statystyczny, *Mały Rocznik Statystyczny Polski,* Warszawa 2019, s. 170

1.3. Opis danych

Dane przedstawiają 79 obserwacji, z których każda zawiera 8 zmiennych:

- ✓ *cena* – cena danego telefonu (w PLN);
- ✓ *RAM* – ilość pamięci o tak zwanym dostępnie swobodnym określanej skrótem RAM (w GB);
- ✓ *pamiec* – ilość pamięci w danym telefonie (w GB);
- ✓ *przekatna* – długość przekątnej w danym telefonie (w calach);
- ✓ *pojemnosc_baterii* – pojemność baterii danego telefonu (w mAh);
- ✓ *aparat_tyl* – ilość px w tylnym aparacie;
- ✓ *PPI* – liczba pikseli przypadająca na cal długości, określa rozdzielczość;
- ✓ *lcd* – zmienna binarna, przyjmuje wartość 1 jeśli ekran jest ekranem typu LSD, 0 – jeśli inny.

Dane zostały zebrane w dniu 24 maja 2020 roku z dwóch stron internetowych:

- ✓ https://www.x-kom.pl/
  zmienne: cena, RAM, pamiec, przekatna, pojemnosc_baterii;
- ✓ https://phonesdata.com/pl/
  zmienne: aparat_tyl, PPI, lcd

Przy gromadzeniu danych odnośnie ceny smartfonów nie uwzględniono akcji promocyjnych sklepu dotyczących niektórych z badanych urządzeń.

1.4. Przyjęte założenia i użyte oprogramowanie

Wszystkie operacje i badania są wykonywane w programie Gretl autorstwa Allina Cottrella z Uniwersytetu Wake Forest w Północnej Karolinie w Stanach Zjednoczonych.

Przyjęty poziom istotności $\alpha = 0{,}05$.

1.5. Statystyki opisowe

Statystyki opisowe (zob. Tabela 1) zawierają najważniejsze właściwości zmiennych, dzięki którym już na początku badań można wyciągnąć wstępne wnioski i wskazywać potencjalne zmienne, które można wykorzystać przy późniejszej konstrukcji modelu – nie będą to m.in. takie, dla których współczynnik zmienności jest mniejszy niż 10%. Jest to spowodowane tym iż niska wartość tej miary zmienności świadczy o małym zróżnicowaniu cechy, a jest to niepożądana własność zmiennej w modelu ekonometrycznym.

Zmienna	Średnia	Mediana	Minimalna	Maksymalna
cena	2629,4	2499,0	449,00	7399,0
RAM	5,6709	4,0000	2,0000	12,000
pamiec	140,96	128,00	16,000	512,00
przekatna	6,3192	6,3900	5,4500	6,9000
pojemnosc_baterii	3857,2	4000,0	2658,0	5260,0
aparat_tyl	51,392	40,000	12,000	167,00
PPI	402,37	402,00	268,00	566,00
lcd	0,45570	0,0000	0,0000	1,0000
Zmienna	**Odchylenie standardowe**	**Wsp. zmienności**	**Skośność**	**Kurtoza**
cena	1703,8	0,64798	0,68517	-0,19559
RAM	2,5905	0,45681	0,91855	0,34813
pamiec	110,65	0,78494	1,9354	3,9385
przekatna	0,31362	0,049629	-0,71098	0,18352
pojemnosc_baterii	721,40	0,18703	0,14404	-0,81507
aparat_tyl	35,537	0,69148	1,1118	0,98613
PPI	70,226	0,17453	-0,18469	-0,19733
lcd	0,50122	1,0999	0,17791	-1,9683
Zmienna	**Percentyl 5%**	**Percentyl 95%**	**Zakres Q3-Q1**	**Brakujące obserwacje**
cena	499,00	5999,0	2750,0	0
RAM	2,0000	12,000	4,0000	0
pamiec	32,000	512,00	64,000	0
przekatna	5,7100	6,7800	0,43000	0
pojemnosc_baterii	2658,0	5260,0	1310,0	0
aparat_tyl	12,000	133,00	48,000	0
PPI	271,00	515,00	72,000	0
lcd	0,0000	1,0000	1,0000	0

Tabela 1. Statystyki opisowe badanych zmiennych

Charakterystyka statystyk opisowych omawianych zmiennych:

A. *cena*

Średnia cena smartfonu to ponad 2 600 zł, z dużym odchyleniem standardowym o wartości 1 700 zł. Mediana jest zbliżona do średniej, co świadczy o umiarkowanej symetryczności badanej cechy. Najtańszy telefon kosztuje 449 PLN, a najdroższy 7399 PLN. Skośność jest prawostronna, co oznacza, że cena większości badanych smartfonów jest niższa od średniej. Ujemna kurtoza wskazuje na słabą koncentrację cen wokół średniej. Występuje silna zmienność cechy.

B. *RAM*

Średnia ilość pamięci RAM w telefonie to prawie 6 GB z odchyleniem standardowym o wartości bliskiej 3 GB. Telefon z najmniejszą pamięcią RAM ma jej 2 GB, a ten z największą – 12 GB. Skośność jest prawostronna, co oznacza, że liczba GB w większości telefonów jest niższa od średniej. Dodatnia kurtoza wskazuje na silną koncentrację ilości pamięci RAM wokół średniej. Występuje silna zmienność cechy.

C. *pamiec*

Średnia pamięć w telefonie to ponad 140 GB, a wyniki odchylają się od niej o około 110 GB. Telefon z najmniejszą pamięcią ma jej 12 GB, a ten z największą – 512 GB. Skośność jest prawostronna, co oznacza, że liczba GB pamięci w większości telefonów jest niższa od średniej. Dodatnia i duża kurtoza wskazuje na bardzo silną koncentrację ilości pamięci wokół średniej. Występuje silna zmienność cechy.

D. *przekatna*

Średnia przekątna ma długość 6,32 cala, a wyniki odchylają się o około 0,31 cala. Najmniejsza przekątna ekranu wynosi 5,45 cala, a największa 6,9 cala. Skośność jest lewostronna, co oznacza, że większość urządzeń ma ekran o przekątnej większej niż średnia. Dodatnia kurtoza wskazuje na silną koncentrację wartości cechy wokół średniej. Występuje silna zmienność cechy.

E. *pojemność_baterii*

Średnia pojemność baterii w smartfonie wynosi 3857,2 mAh z odchyleniem standardowym o wartości 721,4 mAh. Najmniejsza pojemność baterii wynosi 2658 mAh, a największa 5260 mAh. Skośność jest prawostronna, co oznacza, że większość urządzeń ma pojemność baterii mniejszą niż średnia. Ujemna kurtoza wskazuje na słabą koncentrację cechy wokół średniej. Występuje umiarkowana zmienność cechy.

F. aparat_tyl

Średnia ilość px w aparacie zlokalizowanym z tyłu telefonu wynosi 51,392 px, a wartości te odchylają się od średniej o około 36 px. Największa liczba pikseli w aparacie to 167, a najmniejsza – 2. Skośność jest prawostronna, co oznacza, że aparat w większości urządzeń ma mniej px niż średnia. Dodatnia kurtoza wskazuje na silną koncentrację wartości cechy wokół średniej. Występuje silna zmienność cechy.

G. PPI

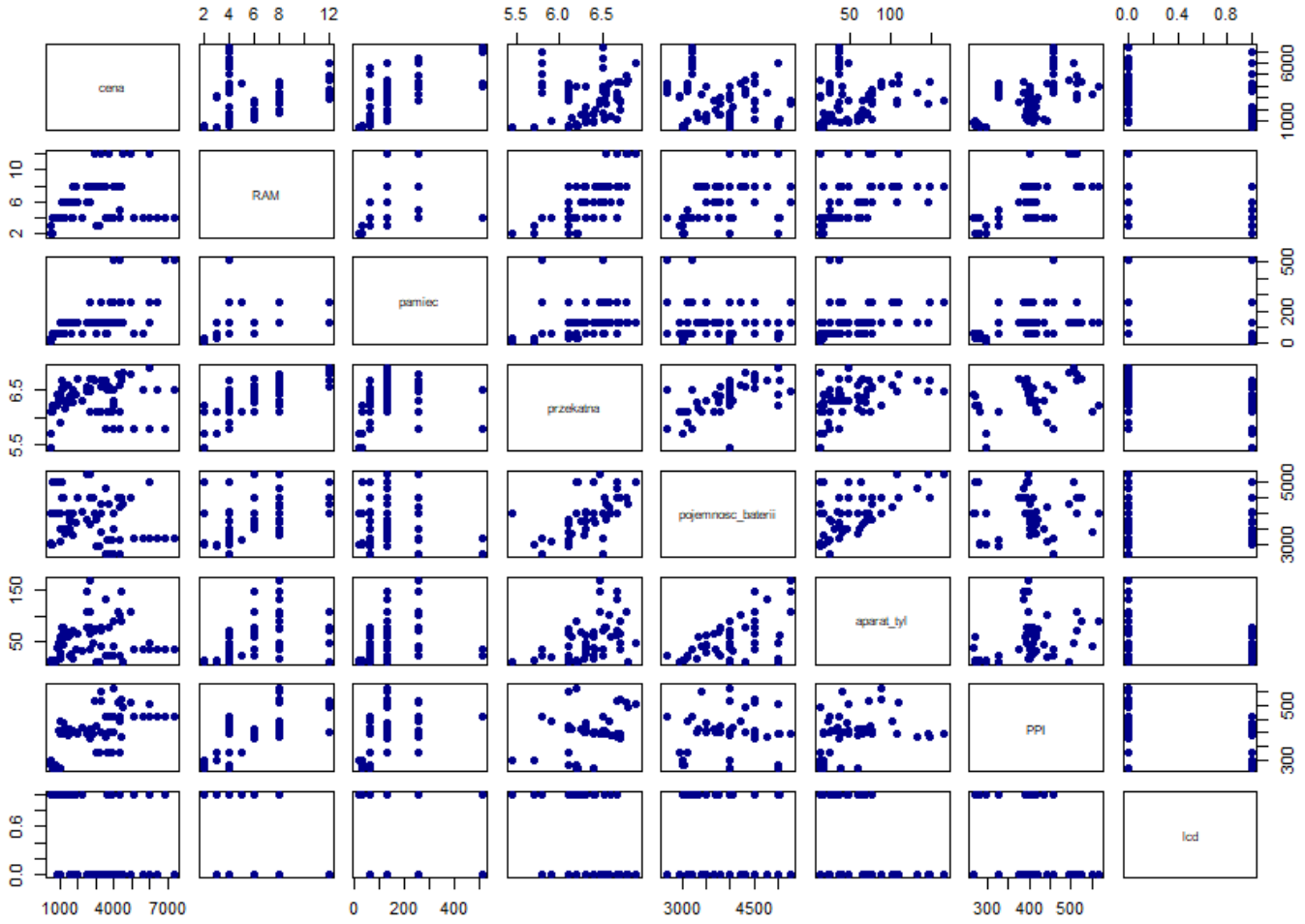
Średnia rozdzielczość w telefonie to około 402 PPI z odchyleniem standardowym równym 70,226 PPI. Najmniejsza wartość PPI w telefonie to 268, a największa to 566. Skośność jest lewostronna, co oznacza, że większość urządzeń ma rozdzielczość ekranu większą niż średnia. Ujemna kurtoza wskazuje na słabą koncentrację cechy wokół średniej. Występuje umiarkowana zmienność cechy.

H. lcd

Średnia wartość wynosi w przybliżeniu 0,46, co oznacza, że większość smartfonów ma wyświetlacz innego typu niż lcd.

1.6. Wykresy zależności

Wykresy zależności zmiennych (zob. Rysunek 1), podobnie jak statystyki opisowe pomagają w wyciąganiu wstępnych wniosków dotyczących badanych cech.



Rysunek 1. Wykresy zależności zmiennych

Na podstawie powyższych wykresów można wnioskować, że wzrost liczby px w tylnym aparacie telefonu nie przekłada się na wzrost ceny, podobnie jak ilość pamięci RAM. Wraz ze wzrostem przekątnej ekranu rośnie też pojemność baterii. Większość najdroższych telefonów ma ekran innego typu niż lcd. Rosnącej cenie towarzyszy wzrost wbudowanej w smartfonie ilości pamięci.

Na pozostałych wykresach nie widać innych istotnych relacji pomiędzy zmiennymi.

1.7. Korelacja pomiędzy zmiennymi

W celu prezentacji danych w czytelniejszej formie zastosowano następujące oznaczenia:

p_y – cena
p_1 – RAM
p_2 – pamiec
p_3 – przekatna
p_4 – pojemnosc_baterii
p_5 – aparat_tyl
p_6 – PPI
p_7 – lcd

p_y	p_1	p_2	p_3	p_4	p_5	p_6	p_7	
1	0,35	0,66	0,19	-0,19	0,16	0,60	-0,50	p_y
	1	0,20	0,64	0,42	0,55	0,56	-0,50	p_1
		1	0,11	-0,18	0,17	0,41	-0,36	p_2
			1	0,55	0,53	0,35	-0,41	p_3
				1	0,61	-0,02	-0,09	p_4
					1	0,32	-0,39	p_5
						1	-0,48	p_6
							1	p_7

Tabela 2. Macierz korelacji (w zaokrągleniu do dwóch miejsc po przecinku)

Kolorem jasnoniebieskim zaznaczono największe, a kolorem jasnoszarym najmniejsze wartości współczynnika korelacji (zob. Tabela 2). W modelu ekonometrycznym powinna występować duża korelacja pomiędzy zmienną objaśnianą i zmiennymi objaśniającymi, a mała korelacja pomiędzy dwiema zmiennymi objaśniającymi. Największa korelacja ze zmienną objaśnianą (*cena*) występuje w przypadku trzech zmiennych: p_2 (*pamiec*), p_6 (*PPI*) oraz p_7 (*lcd*), a najmniejsza ze zmienną p_5 (*aparat_tyl*). Na tej podstawie można przypuszczać, że to właśnie zmienne p_2, p_6 oraz p_7 zostaną wykorzystane w finalnej postaci modelu, w przeciwieństwie do zmiennej p_5. Ponadto silna korelacja występuje pomiędzy: zmienną *RAM* i każdą ze zmiennych p_3, p_5, p_6 oraz parą zmiennych *przekatna*, *pojemność_baterii* i parą *aparat_tyl*, *pojemność_baterii* co prowadzi do wniosku, że te pary nie znajdą się razem w modelu.

2. Model ściśle liniowy

2.1. Klasyczna metoda najmniejszych kwadratów

Analiza ekonometryczna problemu zostanie zrealizowana przy pomocy estymacji parametrów modelu regresji wielorakiej postaci: $y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_k x_k + \epsilon$ gdzie:

- ✓ y – zmienna objaśniana,
- ✓ $x_1, x_2, \ldots, x_k$ – zmienne objaśniające,
- ✓ ϵ – składnik losowy,
- ✓ $\alpha_1, \alpha_2, \ldots, \alpha_k$ – parametry do estymacji.

Najczęściej stosowaną metodą estymacji parametrów strukturalnych α modelu $y = X\alpha + \varepsilon$ jest metoda najmniejszych kwadratów o następujących założeniach[2]:

- ✓ Z1: zmienne objaśniające X_j są nielosowe i nieskorelowane ze składnikiem losowym ε,
- ✓ Z2: $rz(X) = k + 1 \leq n$,
- ✓ Z3: $E(\varepsilon) = 0$,
- ✓ Z4: $D^2(\varepsilon) = E(\varepsilon\, \varepsilon^T) = \sigma^2 I$, przy czym $\sigma^2 < \infty$
- ✓ Z5: Składnik losowy w każdym z okresów t ma rozkład normalny o wartości oczekiwanej 0 i skończonej, stałej wariancji σ^2.

Estymator uzyskany KMNK jest liniowy, zgodny, nieobciążony i najefektywniejszy w klasie liniowych i nieobciążonych estymatorów wektora parametrów α modelu – twierdzenie Gaussa-Markowa.

[2] M. Gruszczyński, T. Kuszewski, M. Podgórska, *Ekonometria i badania operacyjne,* Warszawa 2009, s.32

2.2. Estymacja parametrów modelu ściśle liniowego

Model wyestymowano Klasyczną Metodą Najmniejszych Kwadratów (zob. Tabela 3) – uwzględniono wszystkie zmienne.

Zmienna	Współczynnik	Błąd standardowy	Statystyka t-Studenta	Wartość p
const	−167,373	3402,86	−0,04919	0,9609
RAM	54,7325	77,2421	0,7086	0,4809
pamiec	6,76274	1,33327	5,072	<0,0001
przekatna	38,4660	586,372	0,06560	0,9479
pojemnosc_baterii	−0,297564	0,274729	−1,083	0,2824
aparat_tyl	−2,66866	5,26982	−0,5064	0,6141
PPI	7,14187	2,50471	2,851	0,0057
lcd	−655,560	320,985	−2,042	0,0448

Średnia arytmetyczna zmiennej zależnej	2629,367	Odchylenie stand. zmiennej zależnej	1703,771
Suma kwadratów reszt	88360149	Błąd standardowy reszt	1115,576
Wsp. determinacji R^2	0,609753	Skorygowany wsp. determinacji R^2	0,571278
F(10, 68)	15,84803	Wartość p dla testu F	2,40e-12
Logarytm wiarygodności	−662,2318	Kryt. inform. Akaike'a	1340,464
Kryt. bayes. Schwarza	1359,419	Kryt. Hannana-Quinna	1348,058

Tabela 3. Specyfikacja modelu ściśle liniowego wyestymowanego KMNK

Postać modelu po estymacji (parametry zaokrąglone do dwóch miejsc po przecinku):

$$p_y = -167{,}37 + 54{,}73p_1 + 6{,}76p_2 + 38{,}47p_3 - 0{,}30p_4 - 2{,}69p_5 + 7{,}14p_6 - 655{,}56p_7$$

p_y – cena
p_1 – RAM
p_2 – pamiec
p_3 – przekatna
p_4 – pojemnosc_baterii
p_5 – aparat_tyl
p_6 – PPI
p_7 – lcd

2.3. Wady modelu ściśle liniowego

W poniższym podrozdziale zostaną przedstawione tylko wady modelu ściśle liniowego. Nie będą one szczegółowo wyjaśniane. Istota poniższych testów i cech, które musi spełniać model zostanie dokładnie omówiona w rozdziale czwartym.

W modelu liniowym zaledwie trzy na siedem zmiennych objaśniających są istotne (wartość p mniejsza niż przyjęty poziom istotności 5%). Są to: pamiec, PPI, lcd.

Test RESET wskazuje, że postać analityczna modelu jest niewłaściwie dobrana. Test serii punktuje również nielosowość próby. Występują katalizatory (zmienne, które zafałszowują już nie nazbyt satysfakcjonujący 60% współczynnik determinacji, który umożliwia zmierzenie w jakim stopniu model pozwala na objaśnienie zmienności zmiennej Y, czyli ceny), a natężenie efektu katalizy wynosi aż 17%. Katalizatory to zmienne, które trzeba z modelu usunąć.

Z uwagi na powyższe wady model zostanie poprawiony. Nie będzie badane zjawisko autokorelacji, gdyż wykorzystane w badaniach dane są przekrojowe.

2.4. Metoda Hellwiga

Metoda Hellwiga jest jedną z wielu metod stosowanych w ekonometrii do doboru zmiennych objaśniających.

Y – zmienna objaśniana

$X = \{X_1, X_2, \ldots, X_k\}$ – potencjalne zmienne objaśniające

n – liczba obserwacji

r_j – wartość współczynnika korelacji Pearsona pomiędzy Y a X_j

r_{ij} – wartość współczynnika korelacji Pearsona pomiędzy X_i a X_j

Każda z potencjalnych zmiennych jest źródłem informacji o zmiennej Y. Ponadto rozważa się wszystkie niepustych kombinacje zmiennych ze zbioru X (liczba kombinacji: $L = 2^k - 1$). Dla każdej kombinacji s ($s = \{1, 2, \ldots, 2^k-1\}$; C_s – zbiór numerów zmiennych tworzących s-tą kombinację) oblicza się indywidualną pojemność informacyjną określoną wzorem:

$$h_{sj} = \frac{r_j^2}{\sum_{i \in C_s} |r_{ij}|}$$

Następnie należy obliczyć integralną pojemność integracyjną s-tej kombinacji:

$$H_s = \sum_{j \in C_s} h_{sj}$$

Optymalnym podzbiorem zmiennych objaśniających w sensie Hellwiga jest ten podzbiór dla którego wartość integralnej pojemności informacyjnej jest największa.[3]

Dla omawianego problemu optymalnym podzbiorem zmiennych w sensie Hellwiga jest zbiór: pamiec, PPI, lcd.

[3] M. Podgórska, M. Gruszczyński, *Ekonometria,* Warszawa 2004, s.15

2.5. Metoda krokowa wsteczna

Metoda krokowa wsteczna polega na wyborze najlepszego podzbioru zmiennych objaśniających wg poniższego schematu:

A. Estymacja parametrów modelu.
B. Usunięcie zmiennej, której wartość p (statystyki t-studenta) jest największa.
C. Powtarzanie kroków A i B do momentu, aż wartość p dla wszystkich pozostawionych zmiennych nie będzie mniejsza niż przyjęty poziom istotności co jest równoważne istotności pozostawionych zmiennych.

KROK	Zmienna z największą wartością p statystyki t-studenta	Normalność reszt (test Doornika-Hansena)	Założenie normalności reszt na podstawie centralnego twierdzenia granicznego
1.	przekatna	NIE	TAK
2.	aparat_tyl	NIE	TAK
3.	RAM	NIE	TAK
4.	pojemnosc_baterii	NIE	TAK

Tabela 4. Metoda krokowa wsteczna

Normalność reszt zostanie założona na podstawie Centralnego Twierdzenia Granicznego. Optymalny podzbiór zmiennych uzyskany metodą krokową wsteczną (zob. Tabela 4) to: pamiec, PPI, lcd.

2.6. Wybór podzbioru zmiennych objaśniających

Metoda Hellwiga jak i metoda krokowa wsteczna wskazują ten sam optymalny podzbiór zmiennych objaśniających: pamiec, PPI, lcd.

3. Ostateczna postać modelu

Porównane zostaną rożne postaci modeli zawierające zmienne: pamiec, PPI, lcd. Wybór zostanie dokonany w oparciu o wartość skorygowanego współczynnika R^2, kryteria informacyjne.

Im większa jest wartość skorygowanego współczynnika determinacji R^2 (pojęcie zostanie wyjaśnione w rozdziale 4) i im mniejsza wartość kryteriów informacyjnych tym model jest lepszy.

Analizowano trzy modele (zob. Tabela 5):

(1) $y = a_0 + a_1x_1 + a_2x_2 + a_3x_3$

(2) $\ln(y) = a_0 + a_1\ln(x_1) + a_2\ln(x_2) + a_3x_3$

(3) $\ln(y) = a_0 + a_1x_1 + a_2x_2 + a_3x_3$

gdzie: y – cena, x_1 – pamiec, x_2 – PPI, x_3 – lcd

Numer modelu	Model (1)	Model (2)	Model (3)	Najlepszy model
Skorygowany współczynnik determinacji R^2	0,575188	0,692463	0,653815	(2)
Kryterium informacyjne Akaike'a	1336,070	90,13887	99,49080	(2)
Kryterium informacyjne Schwarza	1345,547	99,61666	108,9686	(2)
Kryterium informacyjne Hannana-Quinna	1339,867	93,93596	103,2879	(2)

Tabela 5. Wybór ostatecznej postaci modelu

Najmniejsze wartości kryteriów informacyjnych i największy skorygowany współczynnik determinacji cechuje model (2). Ze względu na powyższe ostateczna postać modelu to:

$$\ln(y) = a_0 + a_1\ln(x_1) + a_2\ln(x_2) + a_3x_3,$$

gdzie: y – cena, x_1 – pamiec, x_2 – PPI, x_3 – lcd

W dalszych podrozdziałach będą zastosowane następujące oznaczenia odnośnie postaci modelu:

$$y = a_0 + a_1x_1 + a_2x_2 + a_3x_3,$$

gdzie: y – ln(cena) dalej jako l_cena, x_1 – ln(pamiec) dalej jako l_pamiec,

x_2 – ln(PPI) dalej jako l_PPI, x_3 – lcd

3.1. Estymacja finalnego modelu

Model wyestymowano Klasyczną Metodą Najmniejszych Kwadratów – uwzględniono wszystkie zmienne.

Zmienna	Współczynnik	Błąd standardowy	Statystyka t-Studenta	Wartość p
const	−1,93834	1,83485	−1,056	0,2942
lcd	−0,363488	0,112014	−3,245	0,0018
l_pamiec	0,477080	0,0831196	5,740	<0,0001
l_PPI	1,25278	0,327216	3,829	0,0003

Średnia arytmetyczna zmiennej zależnej	7,628777	Odchylenie stand. zmiennej zależnej	0,753134
Suma kwadratów reszt	13,08286	Błąd standardowy reszt	0,417658
Wsp. determinacji R^2	0,704291	Skorygowany wsp. determinacji R^2	0,692463
F(10, 68)	59,54271	Wartość p dla testu F	8,46e-20
Logarytm wiarygodności	−41,06943	Kryt. inform. Akaike'a	90,13887
Kryt. bayes. Schwarza	99,61666	Kryt. Hannana-Quinna	93,93596

Tabela 6. Specyfikacja finalnego modelu wyestymowanego KMNK

Model po estymacji (zob. Tabela 6) przyjmuje postać:

$$y = -1,93834 + 0,47708x_1 + 1,25278x_2 - 0,363488x_3$$

gdzie:

- ✓ y – l_cena (ln(cena)),
- ✓ x_1 – l_pamiec (ln(pamiec)),
- ✓ x_2 – l_PPI (ln(PPI)),
- ✓ x_3 – lcd.

4. Opis i testowanie własności modelu

4.1. Współczynnik determinacji

Współczynnik determinacji R^2 jest miarą dokładności dopasowania modelu do danych. Umożliwia zmierzenie w jakim stopniu model umożliwia objaśnienie zmienności zmiennej Y.

W omawianym modelu współczynnik R^2 jest równy 70,4291%, co oznacza, że model umożliwia objaśnienie zmienności zmiennej Y w 70,4291%.

4.2. Efekt katalizy

Efekt katalizy to zjawisko, kiedy wysoka wartość współczynnika determinacji nie wynika z charakteru i siły powiązań zmiennych objaśniających i zmiennej objaśnianej. Dzieje się to na skutek występowania w modelu zmiennych zwanych katalizatorami.

Niech:
R_0 – macierz korelacji pomiędzy zmienną Y a zmiennymi X_i
R – macierz korelacji pomiędzy X_i a X_j

Zmienna X_i z pary zmiennych (X_i, X_j), $i < j$, jest katalizatorem jeżeli: $r_{ij} < 0$ lub $r_{ij} > \frac{r_i}{r_j}$. Badanie natężenie efekty katalizy wylicza się ze wzoru: $\eta = R^2 - H$, gdzie H jest integralną pojemnością integracyjną zestawu zmiennych objaśniających modelu.

W omawianym modelu nie ma katalizatorów. Natężenie efektu katalizy wynosi 0,93% (w modelu ściśle liniowym było to 17%) i jest stosunkowo niewielkie.

4.3. Normalność rozkładu składnika losowego

Pozytywna weryfikacja założenia normalności rozkładu składnika losowego (w modelu szacowanym KMNK) ma kluczowe znaczenie, gdyż tylko wtedy estymator uzyskany w ten sposób ma własności użyteczne w konstruowaniu testów statystycznych w celu sprawdzania różnych cech modelu ekonometrycznego.

Istnieje wiele testów statystycznych umożliwiających testowanie normalności, są to m.in.:

- ✓ Shapiro-Wilka,
- ✓ Jarque-Bera,
- ✓ Doornika-Hansena,
- ✓ Lillieforsa.

Do testowania normalności rozkładu składnika losowego stosuje się poniższy zestaw hipotez:

H_0: reszty mają rozkład normalny

H_1: reszty nie mają rozkładu normalnego

Test	Wartość p	Wniosek
Shapiro-Wilka	0,0009 < 0,05	Istnieją podstawy do odrzucenia H_0, reszty nie mają rozkładu normalnego
Jarque-Bera	0,0006 < 0,05	Istnieją podstawy do odrzucenia H_0, reszty nie mają rozkładu normalnego
Doornika-Hansena	0,0006 < 0,05	Istnieją podstawy do odrzucenia H_0, reszty nie mają rozkładu normalnego
Lillieforsa	0 < 0,05	Istnieją podstawy do odrzucenia H_0, reszty nie mają rozkładu normalnego

Tabela 7. Wyniki testów normalności rozkładu reszt

Żaden powyższy test (zob. Tabela 7) nie pozwala na stwierdzenie, że reszty w badanym modelu mają rozkład normalny. Z tego powodu ich normalność zostanie założona na podstawie Centralnego Twierdzenia Granicznego.

Centralne twierdzenie graniczne Lindeberga-Levy'ego:

Jeżeli $\{X_k\}$ jest ciągiem niezależnych zmiennych losowych o identycznych rozkładach i skończonej wariancji, to ciąg dystrybuant $\{F_n(t)\}$ zmiennych losowych T_n określonych wzorem: $T_n = \frac{Z_n - nE(X)}{D(X)\sqrt{n}}$, (gdzie $Z_n = \sum_{k=1}^{n} X_k$) spełnia: $\lim_{n\to\infty} F_n(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{\frac{-z^2}{2}} dz$ dla każdej wartości t. Oznacza to, że ciąg zmiennych losowych $\{T_n\}$ jest zbieżny do rozkładu normalnego N(0, 1)[4].

4.4. Istotność zmiennych

4.4.1. Istotność pojedynczych zmiennych

Badanie istotności wpływu zmian danej zmiennej egzogenicznej (objaśniającej) na zmiany wartości zmiennej endogenicznej (objaśnianej) odbywa się z wykorzystaniem testu istotności t-studenta[5].

$H_0: \alpha_j = 0$ zmienna x_j jest nieistotna

$H_1: \alpha_j \neq 0$ zmienna x_j jest istotna

Reszty mają rozkład normalny (zob. 4.3.). Jeśli prawdziwa jest hipoteza zerowa, to zmienna losowa $t = \frac{a_j}{S_{a_j}}$ ma rozkład t-studenta o n – k – 1 stopniach swobody.

Zmienna	Wartość p statystyki t-Studenta	Wniosek
L_pamiec	0,0018	Istnieją podstawy do odrzucenia H_0, zmienna jest istotna
L_PPI	0	Istnieją podstawy do odrzucenia H_0, zmienna jest istotna
lcd	0,0003	Istnieją podstawy do odrzucenia H_0, zmienna jest istotna

Tabela 8. Analiza istotności zmiennych w modelu

W każdym przypadku (zob. Tabela 8) istnieją podstawy do odrzucenia hipotezy zerowej. Wszystkie ze zmiennych l_pamiec, l_PPI, lcd są istotne.

[4] J. Jóźwiak, J. Podgórski, *Statystyka od podstaw,* Warszawa 2012, s. 163

[5] M. Podgórska, M. Gruszczyński, op. cit., s. 55-56

4.4.2. Istotność całego zbioru zmiennych

Badanie, czy dany podzbiór zmiennych objaśniających jest istotny odbywa się przy pomocy uogólnionego testu Walda.

H_0: $\alpha_1 = \alpha_2 = . \ . \ . = \alpha_j = 0$

H_1: co najmniej jeden z parametrów α_j, j = 1, 2, …, k jest różny od zera

Wartość p dla uogólnionego testu Walda wynosi 8,46e-20 i jest mniejsza niż przyjęty poziom istotności 5%. Istnieją podstawy do odrzucenia hipotezy zerowej. Zbiór zmiennych objaśniających wykorzystany w modelu jest istotny.

4.5. Testy dodanych zmiennych

Test dodanych zmiennych weryfikuje, czy wcześniejsze usunięcie zmiennej z modelu było właściwe.

H_0: parametr regresji jest równy zero dla x_i

H_1: parametr regresji nie jest równy zero dla x_i

Zmienna	Wartość p statystyki F	Wniosek
ln(RAM)	0,7261	Brak podstaw do odrzucenia H_0
ln(przekatna)	0,3650	Brak podstaw do odrzucenia H_0
ln(pojemność_baterii)	0,0462	Brak podstaw do odrzucenia H_0
ln(aparat_tyl)	0,1670	Brak podstaw do odrzucenia H_0

Tabela 9. Wyniki testu dodanych zmiennych

We wszystkich czterech przypadkach (zob. Tabela 9) brak jest podstaw do odrzucenia hipotezy zerowej. Parametry wszystkich badanych zmiennych są istotnie równe 0 w modelu. Wcześniejsze usunięcie tych zmiennych z modelu było właściwe.

4.6. Testy pominiętych zmiennych

Test pominiętych zmiennych weryfikuje, czy usunięcie zmiennej z modelu byłoby właściwe.

H_0: parametr regresji jest równy zero dla x_i

H_1: parametr regresji nie jest równy zero dla x_i

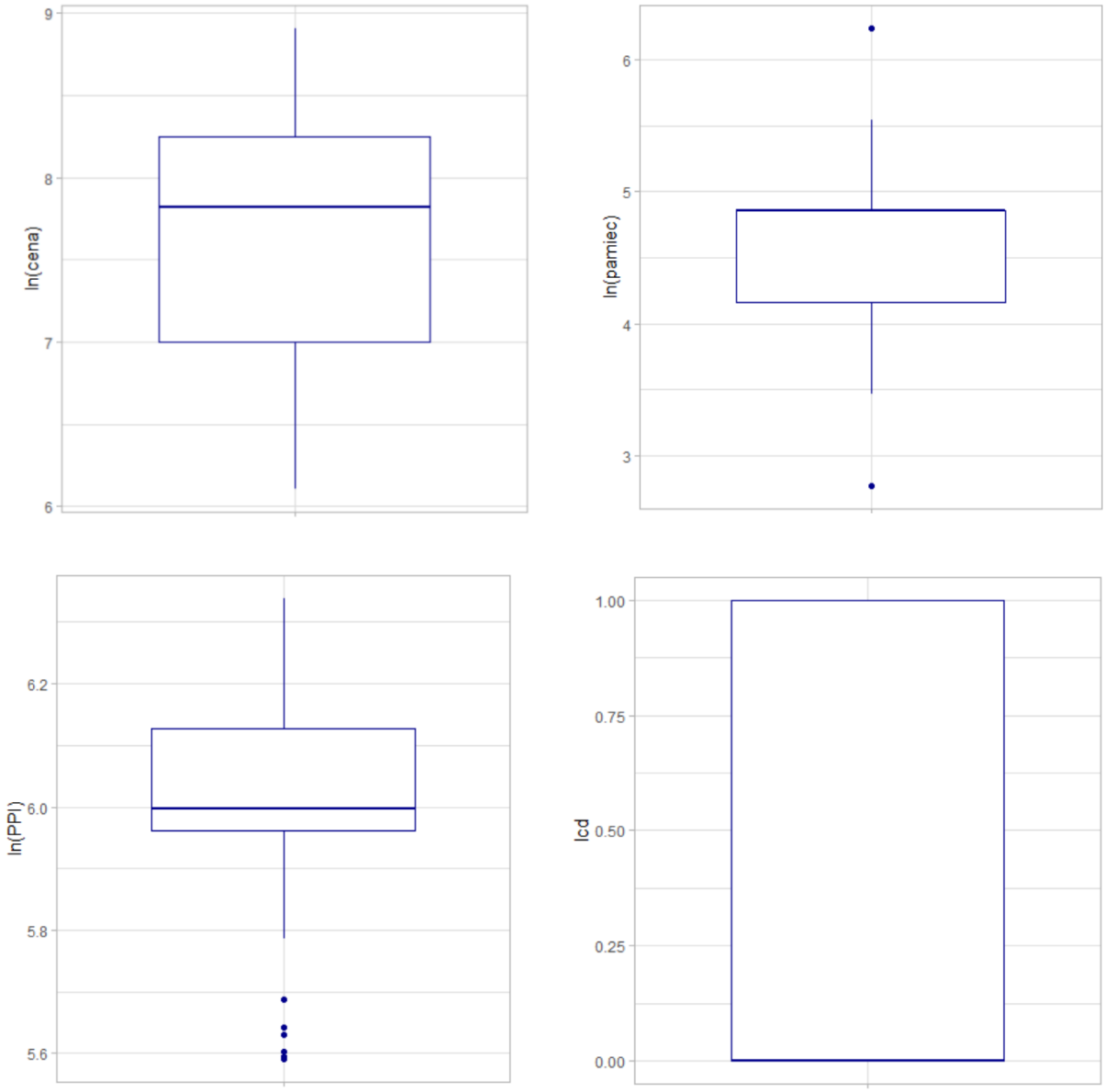
Zmienna	Wartość p statystyki F	Wniosek
L_pamiec	0,0001	Istnieją podstawy do odrzucenia H_0
L_PPI	0,0003	Istnieją podstawy do odrzucenia H_0
lcd	0,0018	Istnieją podstawy do odrzucenia H_0

Tabela 10. Wyniki testu pominiętych zmiennych

We wszystkich trzech przypadkach (zob. Tabela 10) istnieją podstawy do odrzucenia hipotezy zerowej. Parametry wszystkich badanych zmiennych są istotnie różne od 0 w modelu. Usunięcie tych zmiennych z modelu nie byłoby właściwe.

4.7. Obserwacje odstające

Obserwacje odstające to obserwacje relatywnie odległe od pozostałych elementów próby. Można je zaobserwować na wykresach pudełkowych zmiennych (zob. Rysunek 2).



Rysunek 2. Wykresy pudełkowe zmiennych

Obserwacje odstające występują w zmiennej PPI oraz pamiec. Z uwagi na to, że liczba obserwacji wynosi 79 podjęto decyzję o nieusuwaniu obserwacji odstających.

4.8. Test liczby serii

Test liczby serii jest wykorzystywany do weryfikacji założenia o liniowości zależności zmiennej objaśnianej od zmiennych objaśniających, co ma kluczowe znaczenie przy poprawnej interpretacji współczynnika determinacji R^2. Przy przeprowadzaniu testu istotną rzeczą jest posortowanie obserwacji według wybranej zmiennej objaśniającej.[6] W omawianym modelu obserwacje zostały posortowane wg zmiennej *pamiec*.

H_0: Oszacowany model ekonometryczny jest liniowy (postać modelu jest poprawnie dobrana)
H_1: Oszacowany model ekonometryczny nie jest liniowy

Przy dwustronnym obszarze krytycznym wartość p wynosi 0,141032. Brak jest podstaw do odrzucenia hipotezy zerowej. Oszacowany model ekonometryczny jest liniowy, postać modelu jest poprawnie dobrana.

Test liczby serii jest również wykorzystywany do weryfikacji czy badana próba została dobrana w sposób losowy.

H_0: Próba została dobrana w sposób losowy
H_1: Próba nie została dobrana w sposób losowy

Podobnie jak przy postaci hipotez dot. liniowości tak i w przypadku hipotez dot. losowości dobrania próby brak jest podstaw do odrzucenia hipotezy zerowej. Próba została dobrana w sposób losowy.

[6] Ibidem, s. 52

4.9. Test RESET

Test RESET (test Ramseya) jest testem służącym do weryfikacji stabilności postaci analitycznej modelu. Badana jest poprawność liniowej postaci za pomocą sprawdzania czy nie pominięto drugich i trzecich potęg zmiennych objaśniających.[7]

H_0: Wybór postaci analitycznej modelu ekonometrycznego jest prawidłowy

H_1: Wybór postaci analitycznej modelu ekonometrycznego nie jest prawidłowy

1) Test RESET na specyfikację (kwadrat i sześcian zmiennej)
   Statystyka testu: F = 0,530162, z wartością p = P(F(2,73) > 0,530162) = 0,591
   Brak podstaw do odrzucenia hipotezy zerowej.

2) Test RESET na specyfikację (tylko kwadrat zmiennej)
   Statystyka testu: F = 0,897427, z wartością p = P(F(1,74) > 0,897427) = 0,347
   Brak podstaw do odrzucenia hipotezy zerowej.

3) Test RESET na specyfikację (tylko sześcian zmiennej)
   Statystyka testu: F = 0,919246, z wartością p = P(F(1,74) > 0,919246) = 0,341
   Brak podstaw do odrzucenia hipotezy zerowej.

W każdym z trzech przypadków brak jest podstaw do odrzucenia hipotezy zerowej. Wybór postaci analitycznej modelu ekonometrycznego jest prawidłowy.

4.10. Testowanie heteroskedastyczności

Heteroskedastyczność składnika losowego oznacza, że składniki losowe są wzajemnie nieskorelowane, ale mają różne wariancje. W związku z tym estymator a wektora parametrów α modelu nie jest najefektywniejszym estymatorem w klasie estymatorów liniowych i nieobciążonych.[8] Heteroskedastyczność jest zjawiskiem niepożądanym.

[7] Ibidem, s. 101
[8] Ibidem, s. 76

Istnieje kilka testów statystycznych umożliwiających testowanie heteroskedastyczności, są to m.in.:

- ✓ Breuscha-Pagana,
- ✓ White'a,
- ✓ Koenkera.

Do testowania heteroskedastyczności stosuje się poniższy zestaw hipotez:

H_0: Składnik losowy jest homoskedastyczny

H_1: Składnik losowy jest heteroskedastyczny

Test	Wartość p	Wniosek
Breuscha-Pagana	0,2899 > 0,05	Brak podstaw do odrzucenia H_0
White'a	0,0905 > 0,05	Brak podstaw do odrzucenia H_0
Koenkera	0,4566 > 0,05	Brak podstaw do odrzucenia H_0

Tabela 11. Wyniki testów na heteroskedastyczność modelu

We wszystkich przypadkach (zob. Tabela 11) brak jest podstaw do odrzucenia hipotezy zerowej. Składnik losowy jest homoskedastyczny.

4.11. Test Chowa

Test jest używany do badania stabilności parametrów modelu ekonometrycznego. Jest oparty na analizie wariancji i został opisany w 1960 roku przez G. C. Chowa[9].

H_0: Parametry modelu są stabilne

H_1: Parametry modelu nie są stabilne

Próbę podzielono w połowie (obserwacja nr 40). Wartość p testu Chowa jest równa 0,4924 (większa niż przyjęty poziom istotności 0,05). Z tego powodu brak jest podstaw do odrzucenia hipotezy zerowej. Parametry modelu są stabilne.

[9] R. Davidson, J. G. MacKinnon, *Econometric Theory and Methods,* New York 2004 s. 146

4.12. Współliniowość

Współliniowość dla danych przekrojowych polega na proporcjonalnym zmienianiu się wartości zmiennych objaśniających. Zjawisko to jest niepożądane, gdyż uniemożliwia poprawne szacowanie parametrów modelu KMNK. Ocenę współliniowości umożliwia VIF (ang. variance inflation factor) – czynnik inflacji wariancji. Jego wartość powyżej 10 jest oznaką współliniowości, która powoduje trwałe zakłócenie jakości modelu[10].

Test	VIF	Wniosek
l_pamiec	1,635	Brak oznak współliniowości
l_PPI	1,618	Brak oznak współliniowości
lcd	1,409	Brak oznak współliniowości

Tabela 12. Ocena współliniowości modelu

W badanym modelu nie występuje zjawisko współliniowości (zob. Tabela 12).

4.13. Koincydencja

Model jest koincydentny jeżeli zachodzi warunek: sgn r_i = sgn a_i, gdzie r_i to współczynnik korelacji pomiędzy zmienną Y a X_i, natomiast a_i to ocena parametru strukturalnego α_i dla i = 1, 2, …, k. Brak koincydencji może wskazywać m.in. na występowanie współliniowości.

Zmienna	a_i	sgn a_i	r_i	sgn r_i
l_pamiec	0,477080	1	0,7536	1
l_PPI	1,25278	1	0,6858	1
lcd	−0,363488	-1	-0,6092	-1

Tabela 13. Koincydencja modelu

W każdym z trzech przypadków (zob. Tabela 13) warunek: sgn r_i = sgn a_i jest spełniony. Model jest koincydentny.

[10] M. Podgórska, M. Gruszczyński, op. cit., s. 83

4.14. Interpretacja parametrów modelu

Postać modelu: $y = -1{,}93834 + 0{,}47708x_1 + 1{,}25278x_2 - 0{,}363488x_3$

gdzie:

- ✓ y – l_cena,
- ✓ x_1 – l_pamiec,
- ✓ x_2 – l_PPI,
- ✓ x_3 – lcd.

Interpretacja parametrów:

- ✓ Wzrost pamięci o 1% powoduje wzrost ceny o około 0,48%, ceteris paribus[11].
- ✓ Wzrost PPI o 1% powoduje wzrost ceny o około 1,25%, ceteris paribus.
- ✓ Cena telefonu z ekranem typu lcd jest niższa o około 0,36 jednostki, ceteris paribus.

4.15. Predykcja

Prognoza została zrealizowana dla median (zob. Tabela 14).

Zmienna	l_pamiec	l_PPI	lcd
Wartość mediany	4,85	6	0
PROGNOZA			
Prognoza punktowa	7,88873	Błąd prognozy	3,61764
Wariancja prognozy	13,0873	Przedział ufności (95%)	(0,682026; 15,0954)

Tabela 14. Prognoza wartości y na podstawie wartości median zmiennych X

Na 95% y będzie w przedziale (0,422687; 14,8349). Błąd prognozy jest stosunkowo duży.

Prognoza punktowa, po przekształceniu *y = ln(cena)* na *y = cena* wynosi w przybliżeniu: 2 667 PLN.

[11] Ceteris paribus (łac. ceteri - wszyscy inni, reszta; pariter - równie, w podobny sposób) – zwrot oznaczający "wszystko inne takie samo, bez zmian". Umożliwia badanie oddziaływania na siebie dwóch czynników, dzięki odrzuceniu wpływu pozostałych zmiennych.

5. Podsumowanie

Właściwe jest odniesienie się do hipotez postawionych w rozdziale 1.2 i ich zweryfikowanie. Dla przypomnienia wyglądają one następująco.

Na wzrost ceny smartfonów wpływa:

- ✓ pamięć telefonu;
- ✓ liczba pikseli w aparacie zlokalizowanym z tyłu telefonu;
- ✓ pojemność baterii;
- ✓ przekątna ekranu.

Na podstawie opracowanego modelu można stwierdzić, że z tych czterech wymienionych parametrów tylko pamięć telefonu istotnie wpływa na jego cenę.

Wysoki błąd prognozy potwierdza tezę, że model, który przeszedł pozytywną weryfikacje w przypadku wszystkich testów nie zawsze jest modelem odpowiednim do prognoz.

Parametrów w telefonie jest kilkadziesiąt, różnią się specyfikacją, dlatego problem prognozowania cen smartfonów jest trudny. Nie tylko one odgrywają istotną rolę w kształtowaniu ceny, a inne cechy są trudno mierzalne. Wiele osób używa sformułowania „płacenie za markę", czyli domniemane dodatkowe pieniądze, które trzeba dopłacić przy zakupie urządzenia topowego producenta – co jest trudne do przedstawienia wartościowo. Ceny w Polsce są również uzależnione od warunków celno-skarbowych. Wbrew pozorom istnieje szereg czynników, które wpływają na ceny smartfonów w Polsce i nie są to tylko parametry techniczne telefonu.

Bibliografia

1. Główny Urząd Statystyczny, *Mały Rocznik Statystyczny Polski,* Warszawa 2019
2. M. Gruszczyński, T. Kuszewski, M. Podgórska, *Ekonometria i badania operacyjne* Warszawa 2009
3. M. Podgórska, M. Gruszczyński, *Ekonometria,* Warszawa 2004
4. J. Jóźwiak, J. Podgórski, *Statystyka od podstaw,* Warszawa 2012
5. R. Davidson, J. G. MacKinnon, *Econometric Theory and Methods,* New York 2004

Spis tabel

Tabela 1. Statystyki opisowe badanych zmiennych *5*
Tabela 2. Macierz korelacji (w zaokrągleniu do dwóch miejsc po przecinku) *9*
Tabela 3. Specyfikacja modelu ściśle liniowego wyestymowanego KMNK *11*
Tabela 4. Metoda krokowa wsteczna *14*
Tabela 5. Wybór ostatecznej postaci modelu *15*
Tabela 6. Specyfikacja finalnego modelu wyestymowanego KMNK *16*
Tabela 7. Wyniki testów normalności rozkładu reszt *18*
Tabela 8. Analiza istotności zmiennych w modelu *19*
Tabela 9. Wyniki testu dodanych zmiennych *20*
Tabela 10. Wyniki testu pominiętych zmiennych *21*
Tabela 11. Wyniki testów na heteroskedastyczność modelu *25*
Tabela 12. Ocena współliniowości modelu *26*
Tabela 13. Koincydencja modelu *26*
Tabela 14. Prognoza wartości y na podstawie wartości median zmiennych X *27*

Spis rysunków

Rysunek 1. Wykresy zależności zmiennych *8*
Rysunek 2. Wykresy pudełkowe zmiennych *22*