Akademia Górniczo-Hutnicza im. Stanisława Staszica

Wydział Zarządzania

Kierunek: Informatyka i Ekonometria

Przedmiot: Ekonometria, ćwiczenia laboratoryjne

Analiza czynników wpływających na liczbę zameldowań w powiatach w 2019 roku.

Wiktoria Mróz, 402108

Kraków, 14.06.2021r.

Spis treści

1. Wstęp

Temat poddany rozważeniu w tym projekcie będzie dotyczył migracji międzypowiatowej. Zjawisko to towarzyszy ludziom od początku historii człowieka i wynika z wielu różnych czynników. Ogólna definicja migracji mówi, że termin ten dotyczy przemieszczania się ludności związanego z zmianą miejsca zamieszkania (na okres stały lub tymczasowy), połączonego z przekroczeniem granicy administracyjnej podstawowej jednostki terytorialnej, w tym projekcie będą to powiaty. [1]
Powody, dla których ludzie decydują się opuścić miejsce zamieszkania są różne. Można podzielić je na czynniki społeczno-polityczne i związane z tym prześladowania na tle etnicznym, religijnym, rasowym, politycznym lub kulturowym. Czynniki demograficzne i ekonomiczne, które odwołują się do poprawienia standardu pracy i życia, możliwości edukacyjnych, rozwojowi osobistemu, bezrobocia jak i zmiany ze względu na charakter populacji. Często na migracje mają też wpływ czynniki środowiskowe, wszelkie huragany, powodzie, które zmuszają do opuszczenia miejsca zamieszkania.[2]

2. Cel projektu

Celem tego projektu będzie zbadanie co ma wpływ na migracje międzypowiatową w Polsce i jak duże ma to znaczenie. Do określenia tego posłuży budowa modelu ekonometrycznego, zbadane zostaną jego własności i przetestowane właściwości tak by otrzymać ostateczną postać. Model określi czynniki, które najmocniej motywują Polaków do zmiany miejsca zameldowania.

3. Hipotezy badawcze

W tym projekcie weryfikowane będą hipotezy w jaki sposób wybrane czynniki wpływają na migracje międzypowiatową, za miernik której przyjęta została liczba nowych zameldowań w powiatach.

1. Średnia cena mieszkań istotnie wpływa na liczbę zameldowań. Im wyższa średnia tym mniej zameldowań.
2. Liczba szkół ponadpodstawowych może mieć wpływ na zmienną objaśnianą. Im więcej szkół tym więcej zameldowań.
3. Przeciętne miesięczne wynagrodzenie brutto może mieć istotny wpływ na liczbę zameldowań. Im wyższe wynagrodzenie tym więcej ludzi decyduje się na zmianę miejsca zamieszkania w poszukiwaniu lepszych warunków życia.
4. Liczba nowych mieszkań to główny czynnik wpływający na liczbę zameldowań.

[1] https://stat.gov.pl/metainformacje/slownik-pojec/pojecia-stosowane-w-statystyce-publicznej/845,pojecie.html, dostęp: 14.06.2021r.
[2] „Przyczyny migracji: dlaczego ludzie migrują?” https://www.europarl.europa.eu/pdfs/news/expert/2020/7/story/20200624STO81906/20200624STO81906_pl.pdf , dostęp: 14.06.2021r.

4. Opis danych

Dane do projektu pochodzą z Głównego Urzędu Statystycznego, z Banku Danych Lokalnych.

Na zmienną objaśnianą wybrana została liczba nowych zameldowań w powiatach. Jako zmienne objaśniające przyjęte zostały (X1) średnia cena za mieszkania, (X2) liczba szkół ponadpodstawowych, (X3) przeciętne miesięczne wynagrodzenie brutto oraz (X4) liczba nowych mieszkań. Wszystkie ze zmiennych odnoszą się do 2019 roku.

Poniżej przedstawione są ścieżki dostępu do danych ze strony Głównego Urzędu Statystycznego.

Y - Bank danych lokalnych → Dane według dziedzin → Ludność → Migracje wewnętrzne i zagraniczne → Migracje na pobyt stały międzypowiatowe wg typu, kierunku i płci migrantów. Brałam pod uwagę zameldowania ogółem, bez wykazanego podziału.

X1 - Bank danych lokalnych → Dane według dziedzin → Rynek nieruchomości → Rynkowa sprzedaż lokali mieszkalnych → Średnia cena lokali mieszkalnych sprzedanych w ramach transakcji rynkowych.
Transakcje rynkowe ogólnie i powierzchnia użytkowa lokali mieszkalnych ogólnie są brane pod uwagę. Wartości obliczane są jako iloraz wartości i liczby lokali mieszkalnych zaliczonych do danego grupowania.

X2 - Bank danych lokalnych → Dane według dziedzin → Szkolnictwo → Szkoły ponadgimnazjalne i ponadpodstawowe oraz policealne → Szkoły ponadgimnazjalne i ponadpodstawowe oraz policealne ogółem.
Pod uwagę brana jest liczba placówek w powiatach ogółem.

X3 - Bank danych lokalnych → Dane według dziedzin → Wynagrodzenia i świadczenia społeczne → Wynagrodzenia → Przeciętne miesięczne wynagrodzenia brutto
Pod uwagę brana jest ogółem przeciętna miesięczna wysokość wynagrodzeń.

X4 - Bank danych lokalnych → Dane według dziedzin → Przemysł i budownictwo → Mieszkania oddane do użytku
Forma budownictwa: ogółem, zakres przedmiotowy: mieszkania.

W żadnej ze zmiennych nie występują brakujące dane.

5. Statystyki opisowe

Podstawowe statystyki opisowe zmiennych z projektu zapisane są w poniższej tabeli (tab. 1).

Tabela 1: Statystyki opisowe

Zmienna	Średnia	Mediana	Minimalna	Maksymalna
Y	905,100	451,500	100,000	24684
X1	182500	169920	0,000	699600
X2	24,445	17,000	2,000	391,000
X3	4442,900	4305,500	3537,600	8443,300
X4	545,860	230,500	23,000	21599
Zmienna	**Odch.stand.**	**Wsp. zmienności**	**Skośność**	**Kurtoza**
Y	1770,400	1,956	8,309	93,407
X1	69090	0,379	2,405	10,908
X2	30,830	1,261	6,377	59,072
X3	586,550	0,132	2,550	11,045
X4	1542,100	2,825	9,486	108,300
Zmienna	**Percentyl 5%**	**Percentyl 95%**	**Zakres Q3-Q1**	**Brakujące obs.**
Y	183,200	2837,900	521,750	0
X1	107870	316260	62632	0
X2	6,000	63,900	16,000	0
X3	3844,600	5631,700	543,470	0
X4	53,000	1433,800	322,250	0

Dla liczby zameldowań w 2019 roku (Y) ich średnia wartość to 905. Pozostałe obserwacje różnią się od średniej o 1770. Współczynnik zmienności jest wysoki, skośność wykazuje prawostronność rozkładu, a kurtoza wskazuje na rozkład leptokurtyczny, gdzie spora część danych gromadzi się wokół średniej.

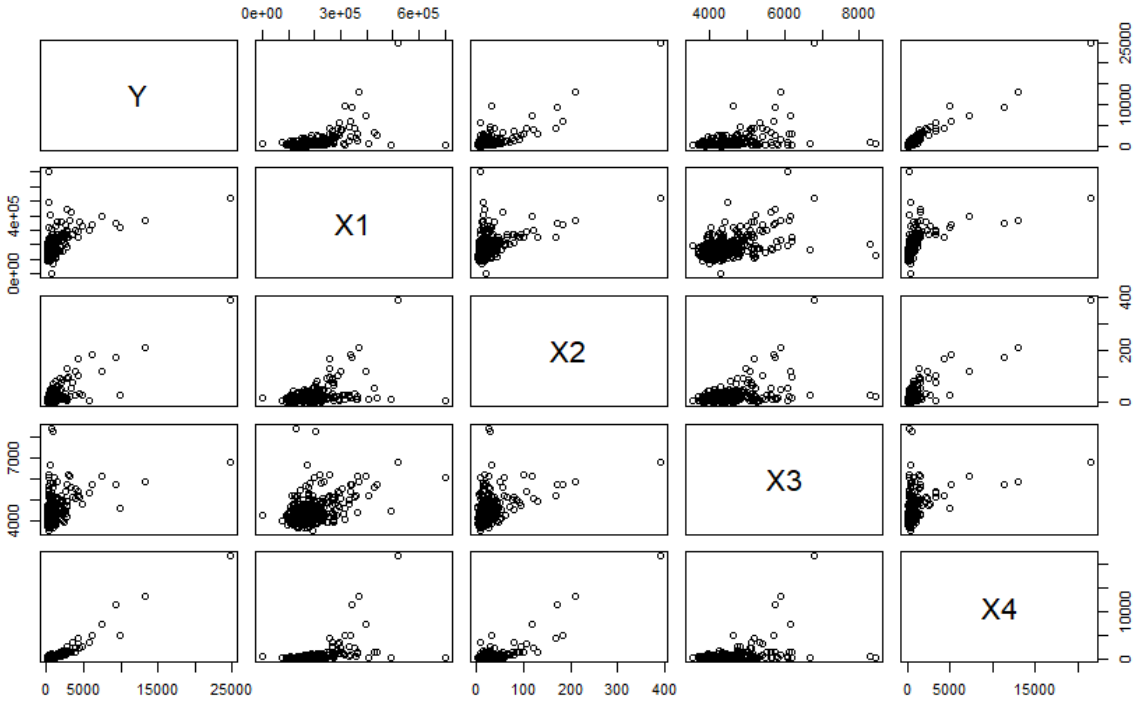
Dla średnich cen za mieszkania w 2019 roku (X1) ich średnia wartość to 182 500 zł. Pozostałe obserwacje różnią się od średniej o 69 090 zł. Współczynnik zmienności 38% wskazuje na przeciętne zróżnicowanie. Skośność i kurtozę można interpretować tak jak w przypadku Y.

Dla liczby szkół ponadpodstawowych zarejestrowanych w 2019 roku (X2) ich średnia wartość to 24. Pozostałe obserwacje różnią się od średniej o 31. Współczynnik zmienności jest wysoki, skośność i kurtozę można interpretować tak jak w przypadku Y.

Dla przeciętnego miesięcznego wynagrodzenia brutto w 2019 roku (X3) średnia wartość to 4442,90 zł. Pozostałe obserwacje różnią się od średniej o 586,55 zł. Współczynnik zmienności jest większy od 10% co oznacza przeciętne zróżnicowanie, skośność i kurtozę można interpretować tak jak w przypadku Y.

Dla liczby nowych mieszkań w 2019 roku (X4) ich średnia wartość to 546. Pozostałe obserwacje różnią się od średniej o 1543. Zmienna ta wykazuje największą współczynnik zmienności. Skośność i kurtozę można interpretować tak jak w przypadku Y.

Na poniższych wykresach (rys. 1.) przedstawione są zależności pomiędzy zmiennymi.



Rysunek 1: Wykresy zależności pomiędzy zmiennymi

Analizując wykresy zależności zmiennych (rys.1), liniowa zależność pomiędzy Y a zmienną X4 jest widoczna i można prognozować, że będą one mieć wysoką korelację. Podobnie pomiędzy zmienną X2 a Y. Reszta danych jednak gromadzi się w skupiskach i choć widać zależności to nie są one bardzo wysokie.

Zmienna X4 wydaję się być mocno zależna od X2, co mogłoby spowodować problemy przy budowie modelu z użyciem Klasycznej Metody Najmniejszych Kwadratów do estymacji.

6. Macierz korelacji

Macierz korelacji pomiędzy zmiennymi przedstawiona jest poniżej (tab. 2).

Tabela 2: Macierz korelacji zmiennych

Y	X1	X2	X3	X4	Zmienne
1,000	0,551	0,815	0,428	0,960	**Y**
	1,000	0,456	0,450	0,507	**X1**
		1,000	0,438	0,850	**X2**
			1,000	0,392	**X3**
				1,000	**X4**

Zmienne X2 i X4 są mocno skorelowane z Y. Reszta wykazuje umiarkowaną istotność korelacji. Pomiędzy zmiennymi X2 a X4 występuje niepokojąca, wysoka zależność, która jest niepożądanym zjawiskiem.

7. Model liniowy

Do wyestymowania parametrów modelu ściśle liniowego użyta zostanie klasyczna metoda najmniejszych kwadratów. Idea metody sprowadza się do takie doboru oszacowań $\hat{\alpha}, \hat{\beta}$ parametrów α, β, aby wartość wyrażenia: $Q = \sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2$ była jak najmniejsza.[3]

Założenia KMNK wymienione poniżej:

1. Model ma postać liniową lub sprowadzalną do liniowej.
2. Zmienne objaśniające są nielosowe i tym samym nie są skorelowane ze składnikiem losowym. Zmienne ekonomiczne są zmiennymi losowymi, zatem konieczne jest przyjęcie założenia, że są one nieskorelowane ze składnikiem losowym, czyli $E(X_k, \varepsilon) = 0$.
3. Pomiędzy żadną z podgrup zbioru zmiennych objaśniających nie występuje dokładna zależność liniowa, czyli $rz(X) = k + 1$ (rząd macierzy X jest równy liczbie parametrów strukturalnych modelu).
4. Wielkość próby jest większa od liczby szacowanych parametrów $n > k + 1$.
5. Wartość oczekiwana składnika losowego jest równa zero, tj. $E(\varepsilon) = 0$.
6. Wariancja składnika losowego jest skończona i stała ($\sigma_\varepsilon^2 < \infty$), natomiast kowariancje są równe zero (nie występuje autokorelacja składnika losowego).[4]

Estymatory uzyskane KMNK są najefektywniejszymi (w klasie liniowych i nieobciążonych estymatorów wektora parametrów modelu ma najmniejszą wariancję spośród nich), liniowymi (można przedstawić jako kombinację liniową zaobserwowanych wartości zmiennej zależnej Y), zgodnymi (prawdopodobieństwo, że jego wartość będzie zbliżona do wartości szacowanego parametru wzrasta wraz z podniesieniem liczebności próby, tj. liczby obserwacji) estymatorami nieobciążonymi (wartość oczekiwana estymatora jest równa wartości estymowanego parametru). [5]

Znając ideę Klasycznej Metody Najmniejszych Kwadratów można przejść do tworzenia wstępnego modelu.

Tabela 3: Model 1 - Estymacja modelu KMNK

	współczynnik	błąd std.	t-Studenta	wartość p	istotność
const	-541,315	198,652	-2,725	0,007	***
X1	0,002	0,002	4,433	0,000	***
X2	-1,592	1,528	-1,041	0,298	
X3	0,125	0,048	2,581	0,010	**
X4	1,068	0,031	34,600	0,000	***

[3] G.S. Maddala, „Ekonometria”, Wydawnictwo Naukowe PWN, Warszawa 2013, str. 104
[4] M.Osińska, M.Kośko, J.Stempińska, „Ekonometria współczesna” , Dom Organizatora, Toruń 2007, str. 47-48
[5] G.S. Maddala, „Ekonometria”, Wydawnictwo Naukowe PWN, Warszawa 2013, str. 104

Średn.aryt.zm.zależnej	905,103	**Odch.stand.zm.zależnej**	1770,425
Suma kwadratów reszt	83425788	**Błąd standardowy reszt**	471,666
Wsp. Determ. R-kwadrat	0,930	**Skorygowany R-kwadrat**	0,929
F(4, 375)	1241,201	**Wartość p dla testu F**	0,000
Logarytm wiarygodności	-2876,063	**Kryt. inform. Akaike'a**	5762,126
Kryt. Bayes. Schwarza	5781,927	**Kryt. Hannana-Quinna**	5769,943

Wzór pierwszego modelu:

Y = -541,315 + 0,002X1 – 1,592X2 + 0,125X3 + 1,068X4

Dokonując wstępnej analizy modelu (tab. 3.) można zauważyć, że 93% zmienności w liczbie zameldowań jest objaśniane przez model czyli zmienne takie jak średnia cena mieszkań, liczba szkół ponadpodstawowych, przeciętne miesięczne wynagrodzenie i ilość nowych mieszkań. Skorygowany R-kwadrat niewiele różni się od współczynnika R-kwadrat.

Istotność zmiennych wykazuje, że zmienna X2 nie jest istotna i należy sprawdzić czy bez niej model byłby lepszy. Usuwając ją również jest szansa na pozbycie się współliniowości, którą wcześniej wykazała macierz korelacji. Wynik testu F wskazuje, że przynajmniej jedna ze zmiennych w modelu jest istotna.

Reszty z modelu nie mają rozkładu normalnego, co wszystkie 4 testy (Test Doornika-Hansena, Test Shapiro-Wilka, Test Lillieforsa, Test Jarque'a-Bera) wykazują. Jednakże z uwagi na to, że w analizie jest używane 380 obserwacji, można skorzystać z Centralnego Twierdzenia Granicznego, które mówi, że dla n>30 rozkład normalny można odgórnie założyć. Zatem normalność rozkładu reszt zostaje założona.

8. Poprawa modelu – metoda Hellwiga

Ideą metody Hellwiga doboru zmiennych jest wybranie takiej kombinacji zmiennych objaśniających z listy możliwych podzbiorów by posiadała ona największą integralną pojemność informacyjną.

Przyjęte jest, że dany zbiór X={ X_1, X_1, X_2, X_3,... X_k } to „kandydatki" na zmienne objaśniające w jednorównaniowym modelu ekonometrycznym opisującym kształtowanie się wartości zmiennej objaśnianej Y oraz znane są współczynniki korelacji liniowej Pearsona między X_i i X_j oraz X_j i Y. Każda „kandydatka" posiada w sobie źródło wiedzy o zmiennej Y. Poszukiwana jest najlepsza kombinacja zmiennych dla modelu.

Na początku liczona jest indywidualna pojemność informacyjna nośnika X_j wchodzącego w skład s-tej kombinacji (wzór 8.1).

$$h_{sj} = \frac{r_j^2}{\sum_{i \in s} |r_{ij}|} \qquad 8.1$$

Następnie obliczana jest integralna pojemność informacyjna s-tej kombinacji (wzór 8.2).

$$H = \sum_{i\epsilon s} h_{sj} \qquad 8.2$$

Za optymalną kombinację nośników uznajemy ten podzbiór zmiennych dla którego integralna pojemność informacyjna jest największa.[6]

W rozważanym w tej pracy modelu występują cztery zmienne objaśniające, które tworzą łącznie 15 kombinacji. Dla każdej z nich liczona jest integralna pojemność informacyjna i zapisana w tabeli poniżej:

Tabela 4: Integralne pojemności informacyjne dla kombinacji

X1	0,304	X1,X2	0,665	X2,X4	0,858	X1,X3,X4	0,740
X2	0,665	X1,X3	0,336	X3,X4	0,784	X2,X3,X4	0,802
X3	0,183	X1,X4	0,814	X1,X2,X3	0,607	X1,X2,X3,X4	0,784
X4	**0,923**	X2,X3	0,590	X1,X2,X4	0,835		

Kombinacja składająca się ze zmiennej X_4 w przypadku budowanego modelu jest optymalna wg. Metody Hellwiga.

Nowy model z użyciem Klasycznej Metody Najmniejszych Kwadratów przedstawiony został poniżej (tab.5).

Tabela 5: Model 2 - Model KMNK po poprawie metodą Hellwiga

	współczynnik	**błąd std.**	**t-Studenta**	**wartość p**	**Istotność**
const	303,091	26,807	11,31	0,000	***
X4	1,103	0,016	67,220	0,000	***
Średn.aryt.zm.zależnej		905,103	**Odch.stand.zm.zależnej**		1770,425
Suma kwadratów reszt		91698235	**Błąd standardowy reszt**		492,532
Wsp. Determ. R-kwadrat		0,923	**Skorygowany R-kwadrat**		0,923
F(4, 378)		4518,943	**Wartość p dla testu F**		0,000
Logarytm wiarygodności		-2894,027	**Kryt. inform. Akaike'a**		5792,053
Kryt. Bayes. Schwarza		5799,934	**Kryt. Hannana-Quinna**		5795,180

Wzór analityczny modelu 2. po redukcji:

Y = 303,091 + 1,103X1

9. Poprawa modelu – metoda krokowa wstecz

Idea tej metody sprowadza się do stopniowego usuwanie zmiennych z modelu. W każdym kroku sprawdzane są kryteria takie jak wielkość współczynnika determinacji, wartość p-value przy zmiennych, decydująca o ich istotności dla modelu oraz wartość p dla testu F, mówiącego

[6] M. Gruszczyński, M. Podgórska, „Ekonometria", Szkoła Główna Handlowa w Warszawie, Warszawa 2004 str. 14-15

o wspólnej istotności zmiennych dla modelu. W ten sposób można ograniczyć model tylko do tych X-ów, które mają znaczenie.[7]

W podstawowym modelu (tab. 3.) zmienna X2 oznaczająca ilość szkół ponadpodstawowych w powiecie nie jest istotna. Z tego względu zostaje odrzucona z modelu.

Nowy model (tab. 6.) niewiele zmienia w kwestii współczynnik determinacji R-kwadrat. Tym razem wszystkie zmienne są istotne, p-value dla testu F pozwala odrzucić hipotezę o braku istotności zmiennych.

Tabela 6:Model 3 - Model KMNK po poprawie metodą krokową wstecz

	współczynnik	błąd std.	t-Studenta	wartość p	Istotność
const	-520,756	197,691	-2,634	0,008	***
X1	0,002	0,000	4,446	0,000	***
X3	0,114	0,048	2,417	0,016	**
X4	1,043	0,019	55,89	0,000	***
Średn.aryt.zm.zależnej		905,103	**Odch.stand.zm.zależnej**		1770,425
Suma kwadratów reszt		83667045	**Błąd standardowy reszt**		471,719
Wsp. Determ. R-kwadrat		0,930	**Skorygowany R-kwadrat**		0,929
F(4, 376)		1654,201	**Wartość p dla testu F**		0,000
Logarytm wiarygodności		-2876,612	**Kryt. inform. Akaike'a**		5762,223
Kryt. Bayes. Schwarza		5776,984	**Kryt. Hannana-Quinna**		5767,477

Wzór analityczny modelu 3. po redukcji:

Y = -520,756 + 0,002X1 + 0,114X3 + 1,043X4

10. Wybór modelu

Ze względu na zbliżone wartość współczynnika determinacji (92,3% a 93%), wybór oparty zostanie o kryteria informacyjne (tab.7), które pod uwagę biorą sumę kwadratów reszt, liczbę obserwacji i liczbę parametrów.[8]

Tabela 7: Porównanie kryteriów informacyjnych modeli

Kryteria informacyjne	Model – metoda Hellwiga	Model – metoda krokowa wstecz
Kryt. Inform. Akaike'a	5792,053	5762,223
Kryt. Bayes. Schwarza	5799,934	5776,984
Kryt. Hannana-Quinna	5795,180	5767,477

Model stworzony dzięki metodzie krokowej wstecz ma niższe wartości kryteriów informacyjnych, dlatego to on zostanie wybrany do dalszej analizy.

[7] B. R. Clarke, „Linear Models The Theory and Application of Analysis of Variance", A John Wiley & Sons INC. Publication New Jersey 2008, str. 65-66

[8] G.S. Maddala, „Ekonometria", Wydawnictwo Naukowe PWN, Warszawa 2013, str. 591

Model, który zostanie poddany dalszej analizie ma postać:

Y = -520,756 + 0,002X1 + 0,114X3 + 1,043X4

11. Analiza modelu

11.1 Współczynnik determinacji

Współczynnik determinacji to miara dopasowania modelu. Zawiera informacje ile procent zmienności Y jest objaśniane przez model. Dla modelu 3. (tab.6.) wynosi 93%, to znaczy że w tylu procentach zmienność liczby zameldowań jest wyjaśniona przez średnią cenę za mieszkania, przeciętne miesięczne wynagrodzenie brutto oraz liczbę nowych mieszkań.

11.2 Efekt katalizy

Efekt katalizy, czyli występowanie zmiennych nazywanych katalizatorami, może fałszować informacje jakie niesie ze sobą współczynnik determinacji. Takie zjawisko nazywane jest efektem katalizy. Żeby być pewnym wyniku R-kwadrat należy przebadać model w poszukiwaniu katalizatorów, tj zmiennych X_i z pary $\{X_i,X_j\}$, i < j, które spełniają nierówności: $r_{ij} < 0$ lub $r_{ij} > \frac{r_i}{r_j}$.

Skrypt przeznaczony do szukania katalizatorów nie znalazł żadnego, dlatego można powiedzieć, że w modelu 3. nie występuje efekt katalizy.

11.3 Normalność rozkładu składnika losowego

Tabela 8: Testy normalności rozkładu reszt

Nazwa testu	Wartość p	Czy hipoteza zerowa, mówiąca o normalności rozkładu zostaje odrzucona?
Test Doornika-Hansena	0,000	Tak
Test Shapiro-Wilka	0,000	Tak
Test Lillieforsa	0,000	Tak
Test Jarque'a-Bera	0,000	Tak

Tak jak już wcześniej zostało wspomniane, nie występuje w modelu normalność rozkładu składnika losowego(tab.8.). Jednakże z uwagi na to, że do analizy zostaje użytych 380 obserwacji, można skorzystać z Centralnego Twierdzenia Granicznego, które mówi, że dla n>30 rozkład normalny można odgórnie założyć. Zatem normalność rozkładu reszt zostaje założona.

11.4 Istotność zmiennych

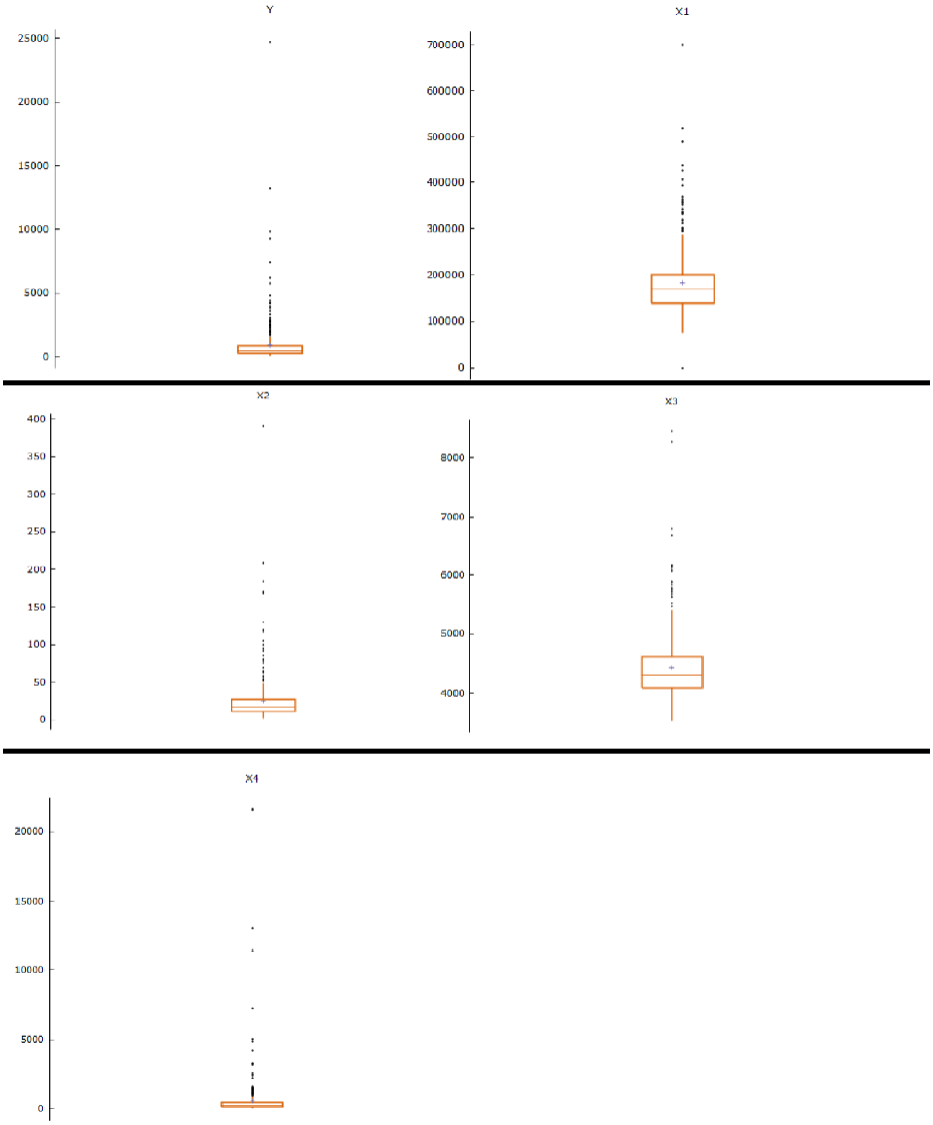
Model wykazuje, że wszystkie zmienne są istotne (tab.6.), test t-Studenta dla każdej zmiennej, który dla 5% poziomu istotności odrzuca hipotezę zerową mówiącą o braku istotności. W teście F Walda również zostaje odrzucona hipoteza zerowa, przynajmniej jedna ze zmiennych jest istotna dla modelu 3.

11.5 Testy pominiętych zmiennych

W przypadku modelu 3., testy zostały przeprowadzone przy metodzie krokowej wstecz i potwierdziły, że usunięcie zmiennej X2 poprawia 3 z 3 kryteriów informacyjnych. (patrz punkt 9.)

11.6 Obserwacje odstające

Do znalezienia obserwacji odstających posłużą wykresy pudełkowe z wąsem, które zawierają informacje o rozkładzie zmiennej.



Rysunek 2: Wykresy pudełkowe z wąsem dla zmiennych

Jak widać na wykresach kilka obserwacji odstaje od reszty danych, pochodzą one z powiatów miast takich jak Warszawa, Sopot oraz powiatów Jastrzębi-Zdrój i Lubin. Usunięcie takich obserwacji pogarsza model: zmniejsza współczynnik determinacji, powoduje zaburzenia w wynikach testów (np. liczby serii) i zmniejsza istotności zmiennych. Nie poprawia również wyników w testach normalności rozkładu reszt i teście badającym występowanie heteroskedastyczności.

Ważnym zjawiskiem w tematyce migracji jest zmiana miejsca zamieszkania z powiatów wiejskich na miejskie, dlatego ze względu na temat pracy usunięcie takich obserwacji z ogólnej analizy może mieć wpływ na niedokładne wyniki.

11.7 Test liczby serii

Podstawą testu liczy serii jest ciąg reszt, w przypadku modelu 3. uporządkowanych względem zmiennej objaśniającej X1. Hipoteza zerowa testu mówi, że oszacowany model ekonometryczny jest liniowy.[9] Dla poziomu istotności 5% nie ma podstaw do odrzucenia hipotezy zerowej (p-value=0,050).

11.8 Test RESET

Test ten weryfikuje hipotezę o stabilności postaci analitycznej modelu, co sprowadza się do sprawdzenia czy wybrana postać modelu faktycznie dobrze sprawdza się dla zagadnienia. W teście badane jest czy w szacowanym modelu ekonometrycznym nie pominięto zmiennych będących drugimi i trzecimi potęgami zmiennych objaśniających.

Tabela 9: Wartości testu RESET

Test RESET	Wartość p-value	Czy hipoteza zerowa zostaje odrzucona?
Kwadraty i sześciany zmiennych	0,000	Tak
Kwadraty zmiennych	0,146	Nie
Sześciany zmiennych	0,580	Nie

Dla samych kwadratów lub dla samych sześcianów nie ma podstaw do odrzucenia hipotezy zerowej. Wybór postaci analitycznej modelu jest prawidłowy. Kiedy brane pod uwagę są obie opcje to hipoteza zerowa zostaje odrzucona, postać analityczna nie jest prawidłowa.[10]

11.9 Testowanie heteroskedastyczności

Heteroskedastyczność jest niepożądanym zjawiskiem. Oznacza to, że składniki losowe nie mają wspólnej wariancji, czyli założenia MNK nie są spełnione.[11]

[9] M. Gruszczyński, M. Podgórska, „Ekonometria", Szkoła Główna Handlowa w Warszawie, Warszawa 2004 str. 52
[10] M. Gruszczyński, M. Podgórska, „Ekonometria", Szkoła Główna Handlowa w Warszawie, Warszawa 2004 str. 101
[11] G.S. Maddala, „Ekonometria", Wydawnictwo Naukowe PWN, Warszawa 2013, str. 241

Testując występowanie heteroskedastyczności w teście Breuscha-Pagana jak i w teście White'a otrzymana wartość p-value równą jest 0, co zmusza do odrzucenia hipotezy o braku tego zjawiska w modelu. Konsekwencją tego jest również potrzeba zmiany postaci modelu.

11.10 Korekta heteroskedastyczności

Poprawa modelu sprowadza się do zmiany sposobu estymacji – z Klasycznej Metody Najmniejszych Kwadratów na Uogólnioną Metodę Najmniejszych Kwadratów.

Uogólniona Metoda Najmniejszych Kwadratów od KMNK różni się brakiem własności efektywności estymatorów, bowiem uwzględnia ona zmiany w wariancji i kowariancji składnika losowego. [12] Macierz wariancji i kowariancji wektora składników losowych przyjmuje postać wskazaną poniżej (wzór 11.10.1).

$$D^2(\varepsilon) = \sigma^2 V, \sigma^2 < +\infty \qquad 11.10.1$$

V jest macierzą symetryczną, dodatnio określoną i znaną lub oszacowaną. Z jej użyciem określa się wektor parametrów modelu.

Poniżej znajduje się tabela zawierająca informacje o nowym modelu:

Tabela 10: Model 4 – Korekta heteroskedastyczności

	współczynnik	błąd std.	t-Studenta	wartość p	istotność
const	-223,619	130,357	-1,715	0,087	*
X1	-0,0009	0,000	-3,225	0,001	***
X3	0,129	0,031	4,175	0,000	***
X4	1,280	0,038	33,500	0,000	***
Średn.aryt.zm.zależnej		1262,656	**Błąd standardowy reszt**		1,833
Wsp. Determ. R-kwadrat		0,770	**Skorygowany R-kwadrat**		0,768
F(3, 376)		418,871	**Wartość p dla testu F**		0,000
Logarytm wiarygodności		-767,349	**Kryt. inform. Akaike'a**		1542,698
Kryt. Bayes. Schwarza		1558,459	**Kryt. Hannana-Quinna**		1548,952

Wzór analityczny modelu 4.: $Y = -223{,}619 - 0{,}0009X_1 + 0{,}129X_3 + 1{,}280X_4$

[12] M.Osińska, M.Kośko, J.Stempińska, „Ekonometria współczesna" , Dom Organizatora, Toruń 2007, str. 161-163

12. Analiza poprawionego modelu

12.1 Współczynnik determinacji

W modelu 4. współczynnik determinacji jest znacznie niższy od tego w modelu 3. jednak nadal ma wysoką wartość. W 78% zmienność liczby zameldowań jest wyjaśniona przez średnią cenę za mieszkania, przeciętne miesięczne wynagrodzenie brutto oraz liczbę nowych mieszkań.

12.2 Normalność rozkładu składnika losowego

Nie występuje w modelu normalność rozkładu składnika losowego. Jednakże z uwagi na to, że w analizie używane jest 380 obserwacji, można skorzystać z Centralnego Twierdzenia Granicznego, które mówi, że dla n>30 rozkład normalny można odgórnie założyć. Zatem normalność rozkładu reszt zostaje założona.

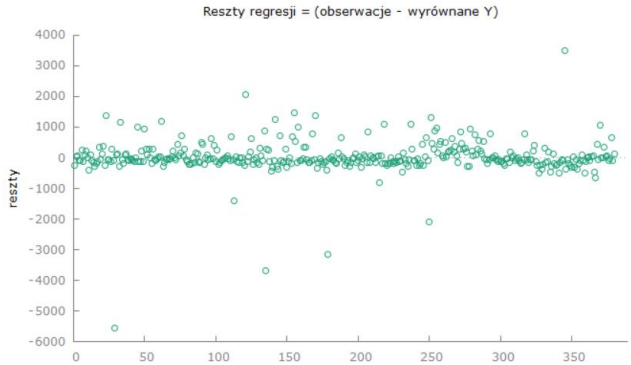
12.3 Istotność zmiennych

Model 4. wykazuje, że wszystkie zmienne są istotne (tab.9.), test t-Studenta dla każdej zmiennej, dla 5% poziomu istotności odrzuca hipotezę zerową mówiącą o braku istotności. W teście F Walda również zostaje odrzucona hipoteza zerowa, tzn. że przynajmniej jedna ze zmiennych jest istotna dla modelu.

12.4 Test Chowa

Test Chowa weryfikuje stabilność parametrów modelu ekonometrycznego. Przeprowadzenie testu sprowadza się do stworzenia modeli i wyciągnięcia z nich resztowej sumy kwadratów. Na początku model jest stworzony dla wszystkich zmiennych, później jest on dzielony.

Podział modelu oparty jest na sposobie przedstawionym w teście Harrisona-McCabe'a. Dla reszt modelu nie występuje monotonia, nie wykazują one też tendencji rosnących lub malejących (rys. 3.), dlatego wybrane zostanie takie m by spełniało nierówność $m > k + 1\ oraz\ n - m > k + 1$.[13]



Rysunek 3: Wykres reszt modelu uzależniony od obserwacji

[13] M. Gruszczyński, M. Podgórska, „Ekonometria", Szkoła Główna Handlowa w Warszawie, Warszawa 2004, str.76-77

Dla $m = 100$, wartość p wynosi 0,05 i na poziomie istotności 5% nie ma powodów do odrzucenia hipotezy mówiącej o stabilności parametrów modelu.

12.5 Współliniowość

Współliniowość uniemożliwia zastosowanie metody estymacji, która zakłada że rz(X) < k+1. Metoda VIF pozwala nam wyszukać zmienną odpowiadającą za współliniowość. Wysoka wartość wskazuje na przeszacowanie współczynnika w modelu. [14]

Tabela 11: Wartości VIF dla zmiennych w modelu 4.

Zmienne	Wartość VIF
X1	1,496
X3	1,315
X4	1,409

Żadna ze zmiennych nie ma wartości większej od 10, która wskazywałaby na problem ze współliniowością.

12.6 Koincydencja

Zjawisko koincydencji, w skrócie badanie merytorycznej sensowności ocen parametrów strukturalnych modelu, jest pożądaną cecha modelu. Sprawdzenie czy w modelu występuje koincydencja sprowadza się do porównania znaków przy współczynniku korelacji ze zmienną Y ze znakiem przy parametrze.[15]

Tabela 12: Koincydencja w modelu 4.

Zmienna	Znak przy współczynniki korelacji zmiennej z Y (Tab. 2)	Znak przy parametrze zmiennej (Tab. 9)
X1	+	-
X3	+	+
X4	+	+

Dla zmiennej X1 (tab.11.) mamy różne znaki, czyli nie można powiedzieć, że w modelu występuje koincydencja. Kierunek zależności Y od X1 nie zgadza się z zależnością wynikającą z danych empirycznych.

[14] A. H. Studenmund, „Using Econometrics: A practical Guide", Pearson, Boston 2016, str. 232-233
[15] M. Gruszczyński, M. Podgórska, „Ekonometria", Szkoła Główna Handlowa w Warszawie, Warszawa 2004, str. 42

12.7 Interpretacja parametrów

Wzrost zmiennej X1 – średnia cena za mieszkanie o jednostkę, ceteris paribus, powoduje zmniejszenie Y – liczby zameldowań o 0,0009 jednostek.

Wzrost zmiennej X3 – przeciętne miesięczne wynagrodzenie brutto o jednostkę, ceteris paribus, powoduje zwiększenie Y – liczby zameldowań, o 0,129 jednostek.

Wzrost zmiennej X4 – liczba nowych mieszkań o jednostkę, ceteris paribus, powoduje wzrost Y – liczby zameldowań, o 1,280 jednostek.

12.8 Prognoza

W tym punkcie będzie wyznaczana prognoza punktowa dla wartości średnich. Wektor wartości zmiennych egzogenicznych modelu dla okresu przyjmuje następujący wygląd $\tau : x_\tau^T = [1 \ X_{1\tau} \ X_{3\tau} \ X_{4\tau}]$. Prognozę wyznaczona zostanie ze wzoru $y_\tau^P = x_\tau^T \, a$, gdzie a to macierz parametrów modelu. [16]

Prognoza punktowa wynosi 884,513. Oznacza to, że prognozowana liczba zameldowań szacowana jest na 885.

Wariancja prognozy jest równa 356,24.

Błąd prognozy wynosi 18,874.

Przedział ufności dla prognozy znajduje się pomiędzy 847,401 a 921,626.

13. Wnioski końcowe

Zbudowanie modelu, który pozytywnie przejdzie wszystkie testy jest trudne do osiągnięcia, ponieważ wpływa na to wiele czynników.

Odwołując się do hipotez badawczych, model faktycznie wskazuje, że im wyższa cena mieszkania tym mniejsza liczba zameldowań. Zmienna X2 okazuje się nie być istotna dla modelu i ku uniknięciu problemu współliniowości została z niego usunięta. Przeciętne miesięczne wynagrodzenie nie ma spodziewanego dużego wpływu na migracje. Liczba nowych mieszkań tak jak zakładano, głownie wpływa na liczbę zameldowań. W tym projekcie zjawisko migracji międzypowiatowej wyrażonej w liczbie zameldowań udało się wytłumaczyć za pomocą trzech zmiennych, objaśniają one model w 77% i jest to zadawalający wynik.

14. Bibliografia

Clarke B. R., „Linear Models The Theory and Application of Analysis of Variance”, A John Wiley & Sons INC. Publication New Jersey 2008

[16] M. Gruszczyński, M. Podgórska, „Ekonometria”, Szkoła Główna Handlowa w Warszawie, Warszawa 2004, str. 106-107

Główny Urząd Statystyczny, https://stat.gov.pl/metainformacje/slownik-pojec/pojecia-stosowane-w-statystyce-publicznej/845,pojecie.html

Gruszczyński M., Podgórska M., „Ekonometria", Szkoła Główna Handlowa w Warszawie, Warszawa 2004

Maddala G.S., „Ekonometria", Wydawnictwo Naukowe PWN, Warszawa 2013

Osińska M., Kośko M., Stempińska J., „Ekonometria współczesna", Dom Organizatora, Toruń 2007

„Przyczyny migracji: dlaczego ludzie migrują?", https://www.europarl.europa.eu/pdfs/news/expert/2020/7/story/20200624STO81906/20200624STO81906_pl.pdf

Studenmund A. H., „Using Econometrics: A practical Guide", Pearson, Boston 2016

15. Spis tabel

16. Spis rysunków