



AGH

**AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA
W KRAKOWIE**

Badanie czynników wpływających na liczbę urodzeń w Polsce

Karol Kita

Nr albumu: 408914

Kierunek: Informatyka i Ekonometria

Rok: II

Przedmiot: Ekonometria

Kraków, 17.06.2020r.

Spis treści:

I. Część pierwsza – wprowadzenie do tematu projektu oraz opis zmiennych	3
1. Opis projektu	3
2. Opis oraz źródła danych	3
3. Statystyki opisowe	3
4. Wykresy zależności zmiennych	4
a) Wykres zależności zmiennej Y od zmiennych X1 – X5	4
a) Wykres zależności zmiennej X1 od zmiennych X2 – X5	5
a) Wykres zależności zmiennej X2 od zmiennych X3 – X5	5
a) Wykres zależności zmiennej X3 od zmiennych X4 – X5	6
a) Wykres zależności zmiennej X4 od zmiennych X5	6
5. Macierz korelacji	7
II. Część druga– model ściśle liniowy	8
1. Szacowanie parametrów regresji	8
2. Wstępna analiza modelu	8
3. Redukcja ilości zmiennych	9
a) Metoda Hellwiga	9
b) Metoda krokowa – wstecz	9
III. Część trzecia – wybór ostatecznej postaci modelu	10
IV. Część czwarta – opis oraz testowanie własności modelu	10
1. Współczynnik determinacji	10
2. Efekt katalizy	10
3. Normalność rozkładu składnika losowego	11
4. Istotność zmiennych	11
5. Testy dodanych zmiennych	11
6. Obserwacje odstające	12
7. Test RESET.....	12
8. Testowanie heteroskedastyczności	13
9. Test Chowa	13
10. Współliniowość	14
11. Koïncydencja	15
12. Interpretacja parametrów modelu	15
13. Predykcja	15
V. Część piąta – podsumowanie	15
VI. Bibliografia	16

I. Część pierwsza – wprowadzenie do tematu projektu oraz opis zmiennych

1. Opis projektu

Liczba urodzeń jest jednym z głównych czynników wpływających na sytuację demograficzną w Polsce, która z roku na rok staje się coraz to gorsza. W projekcie zostanie przebadany wpływ czynników takich jak: zarobki, liczba mieszkań, małżeństwa, wydatki gospodarstw oraz wiek kobiet na liczbę dzieci rodzących się w Polsce.

2. Opis oraz źródło danych

Do projektu zostało użytych 5 zmiennych objaśniających.

Opis oraz oznaczenia zmiennych:

- zmienna objaśniana:
Y: liczba urodzonych dzieci
- zmienne objaśniające:
X1: przeciętny miesięczny dochód rozporządzalny na 1 osobę [zł]
X2: liczba mieszkań oddanych do użytku
X3: liczba zawartych małżeństw
X4: przeciętne miesięczne wydatki na 1 osobę [zł]
X5: średni wiek kobiet podczas porodu pierwszego dziecka

Dane pochodzą ze stron internetowych Głównego Urzędu Statystycznego (<https://stat.gov.pl/>) oraz Eurostat (<https://ec.europa.eu/eurostat>).

3. Statystyki opisowe

	Y	X1	X2	X3	X4	X5
Średnia	380774,90	1118,04	142016,20	206701,70	917,42	26,23
Mediana	376754	1153,66	144151	195284	974,06	26,35
Minimum	351072	610,51	87789	180396	599,49	24,50
Maksimum	417589	1819,14	207425	257744	1251,73	27,60
Odchylenie standardowe	19483,75	371,71	31152,55	23345,12	216,45	0,91
Współczynnik zmienności	5%	33%	22%	11%	24%	3%
Skośność	0,46	0,22	0,12	1,06	-0,15	-0,32
Kurtoza	-0,59	-1,08	-0,59	-0,15	-1,42	-0,93

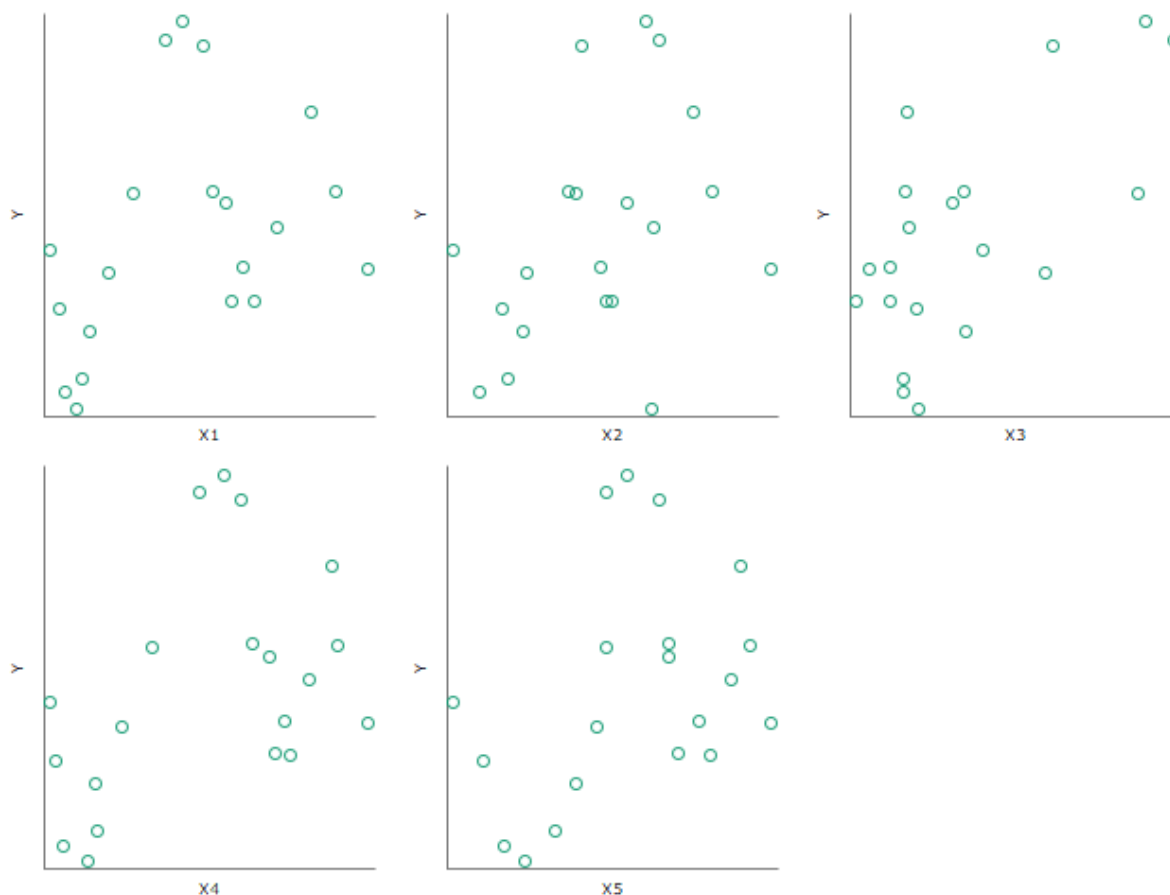
Tabela 1 - statystyki opisowe dla zmiennych

W latach 2000 – 2019 rodziło się średnio około 380 775 dzieci (Y) wartość ta jest większa od mediany. Najmniej dzieci urodziło się w 2003 roku, a najwięcej w 2009 roku. Zmienna Y posiada bardzo dużą wartość odchylenia standardowego co oznacza, że wielkości są szeroko rozrzucone wokół średniej. Z kolei współczynnik zmienności wynosi tylko 5%. Rozkład danych jest spłaszczony oraz prawoskośny.

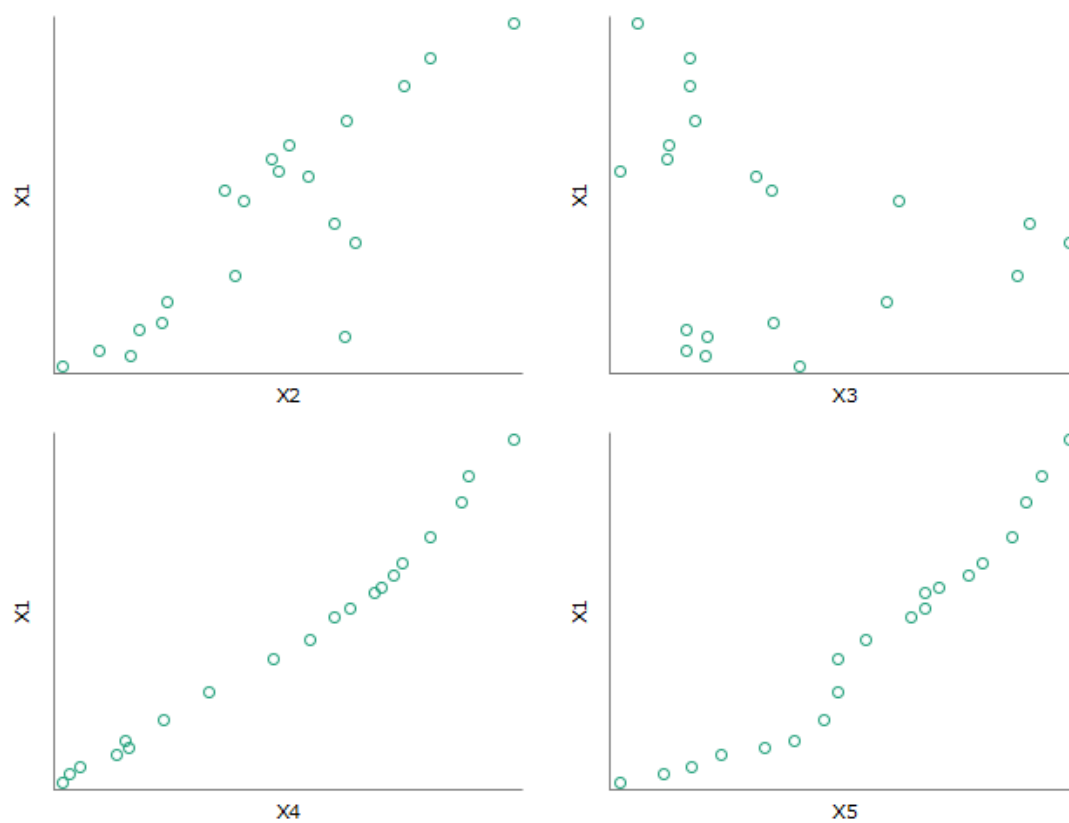
Wśród zmiennych objaśniających średnie wartości w latach 2000 – 2019 prezentują się następująco: miesięczny dochód rozporządzalny na 1 osobę (X1) - 1 118 zł, liczba mieszkań oddanych do użytku (X2) – 142 016, liczba zawartych małżeństw (X3) – 206 701, miesięczne wydatki na 1 osobę (X4) - 917 zł oraz wiek kobiet podczas porodu pierwszego dziecka (X5) – 26 lat. Dla wszystkich zmiennych oprócz X2 średnia jest większa od mediany. Podobnie jak dla zmiennej Y wielkości zmiennych objaśniających są daleko rozrzucone wokół średniej, jednym wyjątkiem jest zmienna X5 dla której wartość odchylenia standardowego wynosi 0,91. Wszystkie zmienne objaśniające posiadają spłaszczony rozkład, gdzie zmienne X4 oraz X5 posiadają rozkład lewoskośny, natomiast pozostałe prawoskośny. Największy współczynnik zmienności posiada zmienna X1 – 33%, a najmniejszy zmienna X5 – 3%. Reszta zmiennych posiada współczynnik zmienności większy od 10%. Z powodu niskiej wartości współczynnika zmienności zmienna X5 powinna być usunięta z modelu.

4. Wykresy zależności zmiennych

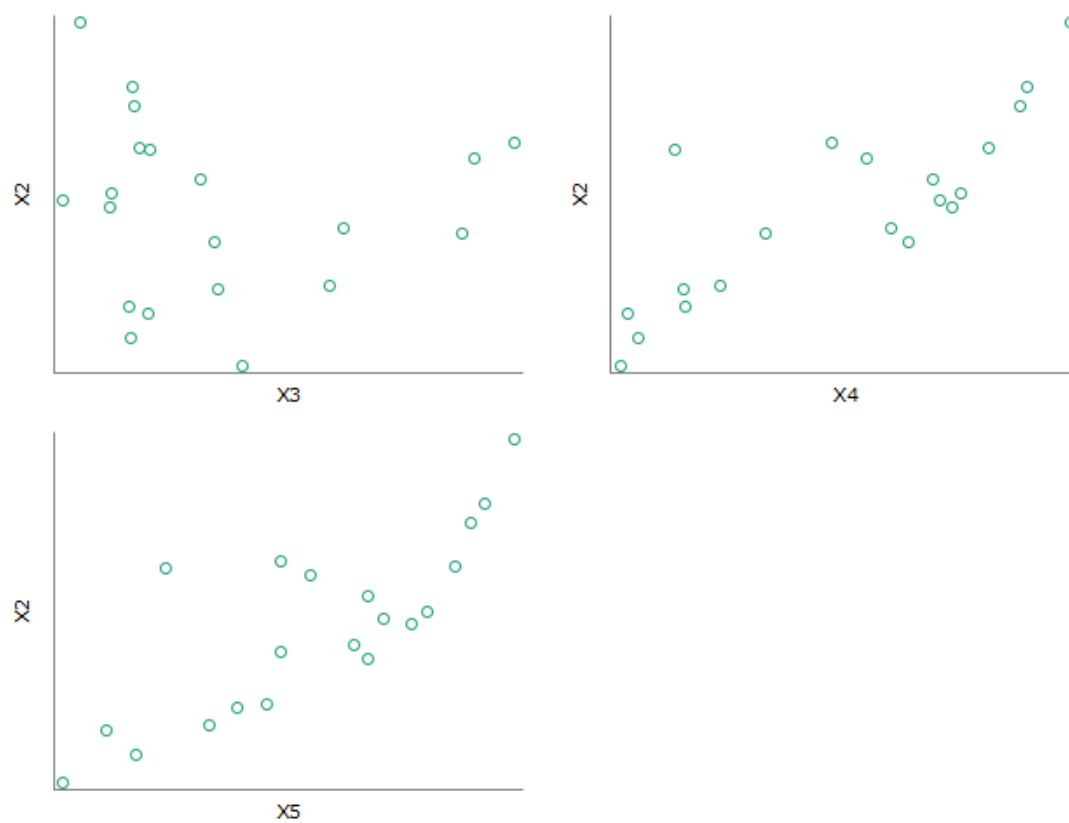
a) Wykres zależności zmiennej Y od zmiennych X1 – X5



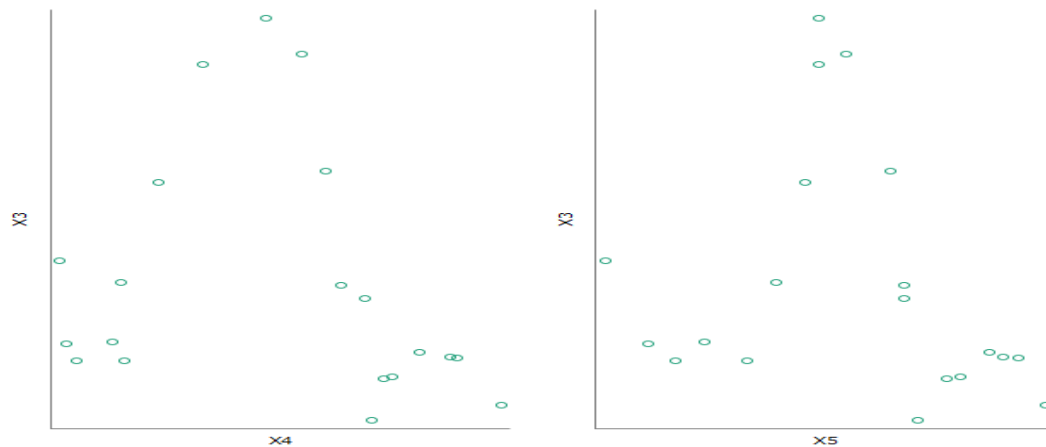
b) Wykres zależności zmiennej X1 od zmiennych X2 – X5



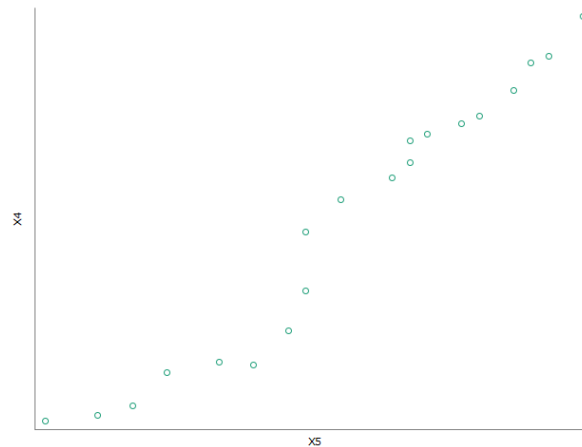
c) Wykres zależności zmiennej X2 od zmiennych X3 – X5



d) Wykres zależności zmiennej X3 od zmiennych X4 – X5

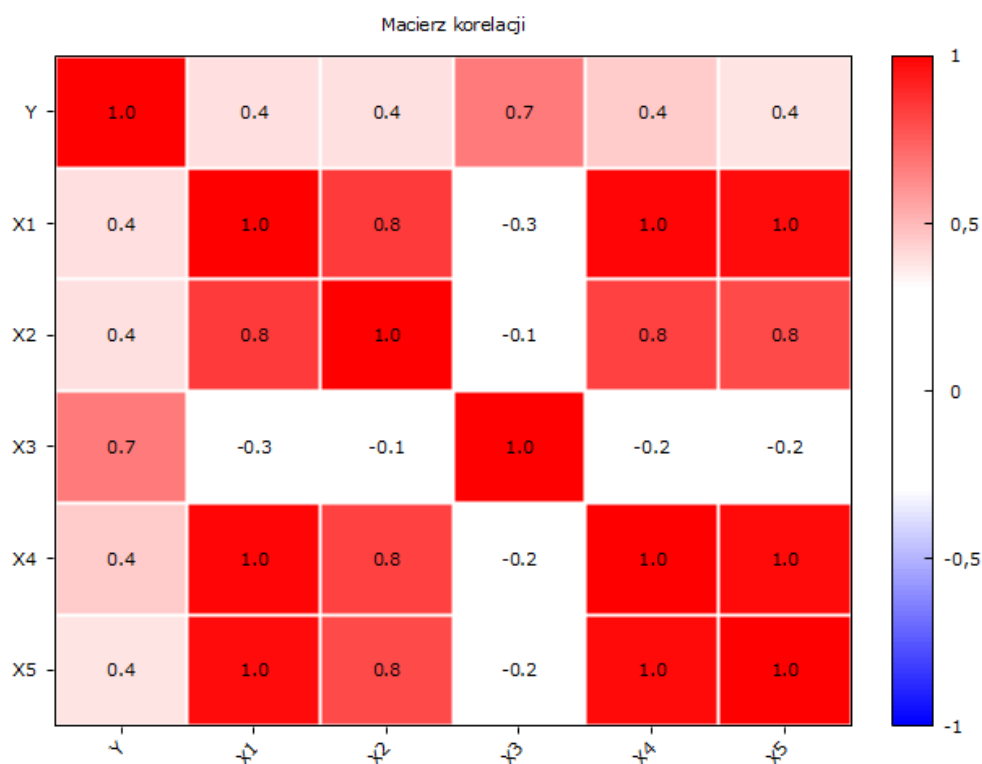


e) Wykres zależności zmiennej X4 od zmiennej X5



Z wykresów wynika, że zmienne objaśniające są skorelowane ze zmienną objaśnianą co wskazuje na poprawny dobór zmiennych objaśniających. Problem pojawia się przy korelacji pomiędzy zmiennymi objaśniającymi, gdzie można zauważyć silną zależność pomiędzy X1, X2, X4 oraz X5. Zależność liniowa między zmiennymi objaśniającymi jest niepożądanym zjawiskiem w modelu.

5. Macierz korelacji



Rysunek 1- mapa ciepła korelacji zmiennych

Współczynniki korelacji, wykorzystane obserwacje 1 - 20

Wartość krytyczna (przy dwustronnym 5% obszarze krytycznym) = 0,4438 dla n = 20

Y	X1	X2	X3	X4	X5	
1,0000	0,3880	0,3842	0,6648	0,4426	0,3726	Y
	1,0000	0,8398	-0,2774	0,9876	0,9706	X1
		1,0000	-0,0794	0,8207	0,7976	X2
			1,0000	-0,2205	-0,2241	X3
				1,0000	0,9764	X4
					1,0000	X5

Rysunek 2- macierz korelacji zmiennych

Jak widać z macierzy korelacji, zmienne objaśniające są skorelowane ze zmienną objaśnianą, wartość współczynników korelacji między zmiennymi X_n a Y wynosi co najmniej 0,4. Najsilniejsza korelacja występuje między zmienną X_3 a Y . W modelu ekonometrycznym nie powinna występować korelacja między zmiennymi objaśniającymi, niestety jak widać silna zależność występuje między zmiennymi X_1 , X_2 , X_4 oraz X_5 .

Biorąc pod uwagę wartości współczynnika zmienności oraz współczynnika korelacji zmienna X_5 zostanie usunięta z modelu. Natomiast pozostałe zmienne czyli X_1 , X_2 oraz X_4 pozostaną w modelu.

II. Część druga – model ściśle liniowy

1. Szacowanie parametrów regresji

Do oszacowania parametrów regresji wykorzystana została Klasyczna Metoda Najmniejszych Kwadratów, gdzie wcześniej z modelu została usunięta zmienna X5.

Model 8: Estymacja KMNK, wykorzystane obserwacje 1-20
Zmienna zależna (Y): Y

	współczynnik	błąd standardowy	t-Studenta	wartość p	
const	195193	25116,6	7,771	1,23e-06	***
X1	17,3997	44,1448	0,3942	0,6990	
X2	-0,143261	0,135625	-1,056	0,3075	
X3	0,706606	0,108806	6,494	1,01e-05	***
X4	44,0550	68,9559	0,6389	0,5325	
Średn. aryt. zm. zależnej	380774,8	Odch. stand. zm. zależnej	19483,75		
Suma kwadratów reszt	1,30e+09	Błąd standardowy reszt	9295,638		
Wsp. determ. R-kwadrat	0,820299	Skorygowany R-kwadrat	0,772379		
F(4, 15)	17,11798	Wartość p dla testu F	0,000018		
Logarytm wiarygodności	-208,2480	Kryt. inform. Akaike'a	426,4959		
Kryt. bayes. Schwarza	431,4746	Kryt. Hannana-Quinna	427,4678		

Wyłączając stałą, największa wartość p jest dla zmiennej 3 (X1)

Rysunek 3- wynik estymacji parametrów regresji za pomocą KMNK

2. Wstępna analiza modelu

Analizę modelu rozpocznę od współczynnika determinacji $R^2 = 0,8203$ oraz skorygowanego $R^2 = 0,7724$, liczba szacowanych zmiennych w modelu jest mniejsza niż liczba obserwacji, dlatego pod uwagę został wzięty współczynnik determinacji, który informuje, że zmienna Y zostaje wyjaśniana przez model w 82%.

Istotność zmiennych objaśnianych została sprawdzona testem t-Studenta¹, gdzie za najbardziej istotne (na poziomie istotności poniżej 0,01) zostały wskazane zmienne const oraz X3. Z kolei największą wartość p posiada zmienna X1, co sugeruje, że zmienna ta powinna być wykluczona z modelu. Porównując wartości p zmiennych X2 oraz X4 okazuje się, wynika, że zmienna X4 powinna być wykluczona jako kolejna z modelu.

Do badania łącznej istotności wszystkich parametrów posłużę się wynikami testu F. Oto hipotezy wykorzystane do testu F: H_0 : wszystkie zmienne są nieistotne oraz H_1 : przynajmniej jedna zmienna jest istotna. Wartość p dla testu F jest mniejsza od poziomu istotności $\alpha = 0,01$, co pozwala na odrzucenie hipotezy zerowej i przyjęciu hipotezy alternatywnej, że co najmniej jedna zmienna w modelu jest istotna.

¹ G.S. Maddala, Introduction to Econometrics, wyd.2, Nowy York 1992

3. Redukcja ilości zmiennych

W celu doboru zmiennych, które będą najlepiej pasować do budowanego modelu, wykorzystane zostaną metody Hellwiga oraz krokowa – wstecz.

a) Metoda Hellwiga

Metoda pojemności integralnej Hellwiga została wykonana z wykorzystaniem skryptu udostępnionego w kursie. Z spośród czterech zmiennych objaśniających metoda wskazała najbardziej istotne zmienne X2 oraz X3 przy pojemności integralnej na poziomie 0,5462. Należy też zwrócić uwagę na kombinacje z niewiele mniejszą wartością pojemności integralnej. Takimi kombinacjami są {X3, X4} oraz {X2, X3, X4} gdzie ich wartość pojemności integralnej wynosi odpowiednio 0,52 oraz 0,51.

b) Metoda krokowa – wstecz

Metodę krokową wstecz wykonałem z wykorzystaniem polecenia *omit* z opcją *--auto*. Metoda wyeliminowała zmienne X1 oraz X2 z modelu, a następnie dokonała estymacji parametrów modelu ze zmiennymi objaśniającymi X3 oraz X4.

```
Sekwencyjna eliminacja nieistotnych zmiennych przy dwustronnym obszarze krytycznym, alfa = 0,10

Wyeliminowano nieistotną zmienną: X1                (wartość p = 0,699)
Wyeliminowano nieistotną zmienną: X2                (wartość p = 0,327)

Hipoteza zerowa: parametry regresji dla wskazanych zmiennych są równe zero
X1, X2
Statystyka testu: F(2, 15) = 0,560891, wartość p 0,582219
Pominięcie zmiennych poprawiło 3 z 3 kryteriów informacyjnych (AIC, BIC, HQC).

Model 7: Estymacja KMNK, wykorzystane obserwacje 1-20
Zmienna zależna (Y): Y
```

	współczynnik	błąd standardowy	t-Studenta	wartość p	
const	191389	22714,3	8,426	1,79e-07	***
X3	0,668816	0,0912027	7,333	1,17e-06	***
X4	55,7441	9,83684	5,667	2,78e-05	***

Średn. aryt. zm. zależnej	380774,8	Odch. stand. zm. zależnej	19483,75
Suma kwadratów reszt	1,39e+09	Błąd standardowy reszt	9052,349
Wsp. determ. R-kwadrat	0,806860	Skorygowany R-kwadrat	0,784137
F(2, 17)	35,50948	Wartość p dla testu F	8,51e-07
Logarytm wiarygodności	-208,9692	Kryt. inform. Akaike'a	423,9383
Kryt. bayes. Schwarza	426,9255	Kryt. Hannana-Quinna	424,5215

Rysunek 4- wydruk polecenia *omit --auto* dla modelu ze zmiennymi objaśniającymi X3 oraz X4

Model ze zmiennymi objaśniającymi X3 oraz X4 posiada wysoki współczynnik determinacji ($R^2=0,81$), co pokazuje, że zmienna Y jest wyjaśniana przez model w 81%. Dodatkowo wszystkie zmienne objaśniające są bardzo istotne, a ich p-value jest mniejsze od 0,01.

III. Część trzecia – wybór ostatecznej postaci modelu

Na początku stworzony model ekonometryczny składał się z 5 zmiennych objaśniających. Z powodu silnej korelacji ze zmiennymi X1 i X4 oraz bardzo niskiej wartości współczynnika zmienności zmienna X5 została usunięta z modelu. Wykonując redukcję zmiennych metodą Hellwiga oraz krokową – wstecz, każda z tych metod wykluczała zmienną X1 z modelu. Również podczas estymacji wstępnego modelu składającego się z zmiennych objaśniających X1, X2, X3 oraz X4 – to właśnie zmienna X1 miała największą wartość p (czyli była najmniej istotną zmienną w modelu). Najbardziej istotną zmienną w modelu jest X3, która została wskazana przez obydwie wykonane metody. Silna zależność X3 ze zmienną objaśnianą oraz słabe korelacje ze zmiennymi objaśniającymi potwierdzają, że zmienna X3 jest najbardziej pożądaną zmienną w modelu. Ze względu na silną korelację pomiędzy X2 oraz X4 jedna z tych zmiennych powinna być usunięta z modelu. Metoda krokowa – wstecz wskazała na model ze zmiennymi X3 oraz X4, natomiast metoda Hellwiga wskazała na model ze zmiennymi X2 oraz X3. Tak jak wspominałem wcześniej wartość pojemności integralnej dla kombinacji {X3, X4} jest niewiele niższa niż wartość pojemności integralnej w kombinacji wskazanej przez metodę Hellwiga. Dodatkowo wartość współczynnika determinacji dla modelu ze zmiennymi X2 oraz X3 wynosi 0,63, gdzie w modelu ze zmiennymi X3 oraz X4 wartość R^2 wynosi 0,81 – dlatego zmienność Y będzie lepiej wyznaczana przez model ze zmiennymi X3 oraz X4.

Ostateczna wersja modelu:

$$Y = \alpha_0 + \alpha_1 X_3 + \alpha_2 X_4$$

Po uwzględnieniu współczynników:

$$Y = 191389 + 0,67 \cdot X_3 + 55,74 \cdot X_4$$

gdzie:

Y - liczba urodzonych dzieci

X3 - liczba zawartych małżeństw

X4 - przeciętne miesięczne wydatki na 1 osobę [zł]

IV. Część czwarta – opis oraz testowanie własności modelu

1. Współczynnik determinacji

W ostatecznej wersji modelu współczynnik determinacji R^2 wynosi 0,81 oraz skorygowany R^2 równa się 0,78. Różnica między R^2 a R^2 skorygowanym wynosi około 0,03 – model nie jest przeparametryzowany.

2. Efekt katalizy

Do sprawdzenia występowania efektu katalizy wykorzystam skrypt udostępniony w kursie. W badanym modelu występuje efekt katalizy, zmienną katalizatorem jest X4, a natężenie efektu katalizy wynosi 0,28. Występowanie efektu katalizy jest złym zjawiskiem, ponieważ zmienne zawyżają wartość współczynnika R^2 .

3. Normalność rozkładu składnika losowego

Do zbadania normalności rozkładu składnika losowego wykorzystałem polecenie *modtest --normality*. Wartość p dla testu wynosi 0,7 czyli jest znacznie większa od 0,05 co świadczy o tym, że reszty mają rozkład normalny.

Rozkład częstości dla uhat0, obserwacje 1-20
liczba przedziałów = 7, średnia = 2,61934e-011, odch.std. = 9052,35

Przedziały	średnia	liczba	częstość	skumulowana
< -12400,	-15002,	1	5,00%	5,00% *
-12400, - -7197,3	-9798,8	4	20,00%	25,00% *****
-7197,3 - -1994,4	-4595,9	3	15,00%	40,00% *****
-1994,4 - 3208,4	607,01	7	35,00%	75,00% *****
3208,4 - 8411,3	5809,9	1	5,00%	80,00% *
8411,3 - 13614,	11013,	2	10,00%	90,00% ***
>= 13614,	16216,	2	10,00%	100,00% ***

Hipoteza zerowa: dystrybuanta empiryczna posiada rozkład normalny. Test Doornika-Hansena (1994) - transformowana skośność i kurtoza.:
Chi-kwadrat(2) = 0,706 z wartością p 0,70274

Rysunek 5- wydruk polecenia *modtest --normality*

4. Istotność zmiennych

	współczynnik	błąd standardowy	t-Studenta	wartość p	
const	191389	22714,3	8,426	1,79e-07	***
X3	0,668816	0,0912027	7,333	1,17e-06	***
X4	55,7441	9,83684	5,667	2,78e-05	***

Rysunek 6- wydruk estymacji za pomocą KMNK

Badanie istotności zmiennych w modelu wykonuję się z pomocą testu t-Studenta. Wartości testu, p-value oraz oznaczenia istotności zmiennych (*) można odczytać w ostatnich trzech kolumnach w *Rysunek 6*. W modelu występują zmienne istotne na poziomie ufności 99%. Występowanie takiego zjawiska jest bardzo pożądane w modelu.

5. Testy dodanych zmiennych

Z wykorzystaniem polecenia *add* wykonany został test dodanych zmiennych. Do modelu została dodana zmienna X2.

Hipoteza zerowa: parametr regresji jest równy zero dla X2
Statystyka testu: F(1, 16) = 1,02029, wartość p 0,327483
Dodanie zmiennych poprawiło 0 z 3 kryteriów informacyjnych (AIC, BIC, HQC).

Model 31: Estymacja KMNK, wykorzystane obserwacje 1-20
Zmienna zależna (Y): Y

	współczynnik	błąd standardowy	t-Studenta	wartość p	
const	191521	22701,2	8,437	2,76e-07	***
X3	0,685874	0,0926996	7,399	1,50e-06	***
X4	70,2998	17,4443	4,030	0,0010	***
X2	-0,119791	0,118594	-1,010	0,3275	

Średn.aryt.zm.zależnej	380774,8	Odch.stand.zm.zależnej	19483,75
Suma kwadratów reszt	1,31e+09	Błąd standardowy reszt	9046,952
Wsp. determ. R-kwadrat	0,818438	Skorygowany R-kwadrat	0,784395
F(3, 16)	24,04133	Wartość p dla testu F	3,61e-06
Logarytm wiarygodności	-208,3510	Kryt. inform. Akaike'a	424,7020
Kryt. bayes. Schwarz	428,6849	Kryt. Hannana-Quinna	425,4795

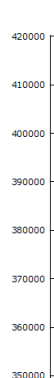
Wyluczając stałą, największa wartość p jest dla zmiennej 4 (X2)

Rysunek 7- wydruk polecenia *add*

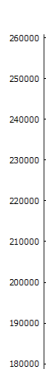
Jak widać na *Rysunek 7* dodanie zmiennej X2 nie poprawiło żadnego z 3 kryteriów informacyjnych (AIC, BIC oraz HQC). Podczas estymacji KNMK zmienna X2 nie znacznie poprawiła wartość współczynnika determinacji oraz zmienna ta okazała się najmniej istotną zmienną w modelu.

6. Obserwacje odstające

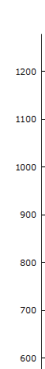
W celu znalezienia obserwacji odstających wygenerowane zostały wykresy pudełkowe dla zmiennych Y, X3 oraz X4.



Rysunek 8- wykres pudełkowy zmiennej Y



Rysunek 9 - wykres pudełkowy zmiennej X3



Rysunek 10- wykres pudełkowy zmiennej X4

Jak widać na załączonych wykresach, zmienne nie posiadają wartości odstających.

7. Test RESET

Z wykorzystaniem polecenia *reset* wykonany zostanie test, który sprawdzi poprawność formy funkcyjnej modelu, gdzie hipotezą zerową tego testu jest założenie, że postać liniowa modelu jest poprawna.

```
Pomocnicze równanie regresji dla testu specyfikacji RESET
Estymacja KMNK, wykorzystane obserwacje 1-20
Zmienna zależna (Y): Y
```

	współczynnik	błąd standardowy	t-Studenta	wartość p
const	-5,56604e+06	1,42136e+07	-0,3916	0,7009
X3	-63,1536	149,350	-0,4229	0,6784
X4	-5255,41	12447,0	-0,4222	0,6789
yhat^2	0,000242123	0,000581761	0,4162	0,6832
yhat^3	-2,04413e-010	5,04570e-010	-0,4051	0,6911

```
Statystyka testu: F = 0,346046,
z wartością p = P(F(2,15) > 0,346046) = 0,713
```

Rysunek 11- wydruk polecenia reset

Wartość p wynosi 0,713 jest zatem większa od 0,05 co oznacza, że model jest liniowy.

8. Testowanie heteroskeastyczności

Do testowania heteroskeastyczności wykorzystałem test White'a oraz test Breuscha-Pagana. Hipotezą zerową w obydwu testach jest brak heteroskedastyczności².

```
Test White'a na heteroskedastyczność reszt (zmiennosc wariacji resztowej)
Estymacja KMNK, wykorzystane obserwacje 1-20
Zmienna zależna (Y): uhat^2
```

	współczynnik	błąd standardowy	t-Studenta	wartość p
const	-3,31282e+09	2,96663e+09	-1,117	0,2829
X3	30489,4	22513,0	1,354	0,1971
X4	-185426	2,60148e+06	-0,07128	0,9442
sq_X3	-0,0504006	0,0468192	-1,076	0,2999
X2_X3	-7,42266	8,76082	-0,8473	0,4111
sq_X4	936,148	799,522	1,171	0,2612

Wsp. determ. R-kwadrat = 0,258852

```
Statystyka testu: TR^2 = 5,177044,
z wartością p = P(Chi-kwadrat(5) > 5,177044) = 0,394658
```

Rysunek 12- wydruk testu White'a

```
Test Breuscha-Pagana na heteroskedastyczność
Estymacja KMNK, wykorzystane obserwacje 1-20
Zmienna zależna (Y): standaryzowane uhat^2
```

	współczynnik	błąd standardowy	t-Studenta	wartość p
const	-0,733569	3,16354	-0,2319	0,8194
X3	9,64571e-06	1,27023e-05	0,7594	0,4580
X4	-0,000283639	0,00137003	-0,2070	0,8384

Wyjaśniona suma kwadr. = 1,15084

```
Statystyka testu: LM = 0,575421,
z wartością p = P(Chi-kwadrat(2) > 0,575421) = 0,749979
```

Rysunek 13- wydruk testu Breuscha-Pagana

W obu testach wartość jest większa od poziomu istotności 0,05 co wskazuje na brak podstaw do odrzucenia hipotezy zerowej. W modelu nie występuje heteroskedastyczność.

9. Test Chowa

Do ocen stabilności parametrów strukturalnych przy znanym punkcie załamania zostanie wykonany test Chowa, gdzie hipotezą zerową testu jest brak zmian strukturalnych. W teście zostaną sprawdzona stabilność parametrów strukturalnych pomiędzy dwoma przedziałami czasowymi 2000 – 2009 i 2010 – 2019.

² G.S. Maddala, Introduction to Econometrics, wyd.2, Nowy York 1992

Pomocnicze równanie regresji dla testu Chowa
 Estymacja KMNK, wykorzystane obserwacje 1-20
 Zmienna zależna (Y): Y

	współczynnik	błąd standardowy	t-Studenta	wartość p	
const	212738	24966,9	8,521	6,52e-07	***
X3	0,872460	0,240304	3,631	0,0027	***
X4	-38,6443	60,1096	-0,6429	0,5307	
splitdum	-80810,8	79944,9	-1,011	0,3292	
sd_X3	0,0255458	0,298825	0,08549	0,9331	
sd_X4	108,028	73,3708	1,472	0,1630	
Średn.aryt.zm.zależnej	380774,8	Odch.stand.zm.zależnej	19483,75		
Suma kwadratów reszt	9,32e+08	Błąd standardowy reszt	8161,260		
Wsp. determ. R-kwadrat	0,870716	Skorygowany R-kwadrat	0,824544		
F(5, 14)	18,85780	Wartość p dla testu F	8,98e-06		
Logarytm wiarygodności	-204,9551	Kryt. inform. Akaike'a	421,9102		
Kryt. bayes. Schwarza	427,8846	Kryt. Hannana-Quinna	423,0765		
Test Chowa na zmiany strukturalne przy podziale próby w obserwacji 10					
F(3, 14) = 2,30499 z wartością p 0,1213					

Rysunek 14 - wydruk testu Chowa

Wartość $p = 0,12$ jest większa od 0,05 co oznacza brak podstaw do odrzucenia hipotezy zerowej co pokazuje, że parametry są takie same w podpróbkach.

10. Współliniowość

Do sprawdzenia współliniowości wykorzystam polecenie *vif*.

Ocena współliniowości VIF(j) - czynnik rozdęcia wariancji

VIF (Variance Inflation Factors) - minimalna możliwa wartość = 1.0

Wartości > 10.0 mogą wskazywać na problem współliniowości - rozdęcia wariancji

```
X3    1,051
X4    1,051
```

$VIF(j) = 1/(1 - R(j)^2)$, gdzie $R(j)$ jest współczynnikiem korelacji wielorakiej pomiędzy zmienną 'j' a pozostałymi zmiennymi niezależnymi modelu.

Rysunek 15- wydruk polecenia *vif*

Żadna ze zmiennych nie posiada wartości *vif* większej od 10, wręcz są bardzo blisko wartości minimalnej 1. Można stwierdzić, że w modelu nie występuje współliniowość zmiennych objaśniających.

11. Koincydencja

Model ekonometryczny posiada własność koincydencji, jeśli dla każdej zmiennej objaśniającej znak współczynnika stojącego przy zmiennej w modelu jest równy znakowi współczynnika korelacji ze zmienną objaśnianą.

	współczynnik	korelacja
X3	0,67	0,66
X4	55,74	0,44

Tabela 2- wartości współczynników oraz korelacji zmiennych objaśniających ze zmienną objaśnianą

Jak widać w *Tabela 2* znak współczynnika stojącego przy zmiennej jest równy znakowi współczynnika korelacji, więc można stwierdzić, że w modelu występuje zjawisko koincydencji.

12. Interpretacja parametrów modelu

Wraz ze wzrostem liczby małżeństw o 1, liczba urodzeń zwiększa się o 0,67 natomiast, gdy przeciętne miesięczne wydatki na jedną osobę wzrosną o 1 złoty to liczba urodzeń wzrasta o 55,74.

13. Predykcja

Do wykonania predykcji wraz z 95% przedziałem ufności wykorzystałem skrypt udostępniony w kursie. Po wykonaniu skryptu otrzymałem następujące wartości:

- wartość prognozy: 380 775
- błąd prognozy: 9275,9
- skalar dolny: 361 204
- skalar górny: 400 345

Dla otrzymanych wyników można stwierdzić, że na 95% w przyszłym roku urodzi się 380 775 dzieci.

V. Część piąta – podsumowanie

Model miał na celu wyłonienie czynników, które mają wpływ na liczbę urodzonych dzieci w Polsce. Ostatecznie okazało się, że największy wpływ mają liczba zawieranych małżeństw oraz przeciętne miesięczne wydatki w przeliczeniu na jedną osobę. Na starcie projektu zakładałem, że przeciętny miesięczny dochód w przeliczeniu na jedną osobę oraz średni wiek matki podczas narodzin pierwszego dziecka będzie również ważny. Choć liczba mieszkań oddanych do użytku nie została dodana do modelu to i tak wbrew moim przewidywaniom okazała się bardziej istotna niż dane o przychodach i wieku matki.

VI. Bibliografia

- G.S. Maddala, *Introduction to Econometrics*, wyd.2, Nowy York 1992
- K. Jankowska-Kowalska, *Wybór testów do weryfikacji liniowych modeli ekonometrycznych*, Wydawnictwo Politechniki Śląskiej, Gliwice 2012
- T. Kufel, *Ekonometria. Rozwiązywanie problemów z wykorzystaniem programu GRETL*, PWN, 2011
- E. W. Nowakowski, *Podstawy ekonometrii z elementami algebry liniowej*, Wszechnica Polska Szkoła Wyższa TWP w Warszawie, 2011