PROJEKT

Analiza wybranych czynników wpływających na liczbę rozwodów w Polsce w 2018 roku

Paulina Korzeniowska
Wydział: Zarządzania
Kierunek: Informatyka i Ekonometria
Rok akademicki: 2020/2021
Studia: I stopnia, IV semestr

Spis treści

1. Wstęp

Cel projektu

Celem projektu jest przeprowadzenie analizy ekonometrycznej i zbadanie wpływu 11 czynników na liczbę rozwodów oraz zweryfikowanie powszechnych hipotez na temat przyczyn rozpadu małżeństw.

Źródła danych

Dane zostały pobrane z https://bdl.stat.gov.pl/BDL/dane/podgrup/temat

Swój zestaw dany stworzyłam z różnych zestawów, które znajdowały się w następujących kategoriach

- liczba rozwodów ludność/małżeństwa i rozwody/rozwody
- gęstość zaludnienia LUDNOŚĆ/STAN LUDNOŚCI/Gęstość zaludnienia oraz wskaźniki
- feminizacja ludność/stan ludności/współczynnik feminizacji
- przemoc w rodzinie, alkoholizacja, narkomania OCHRONA ZDROWIA, OPIEKA SPOŁECZNA I ŚWIADCZENIA NA RZECZ RODZINY/ŚWIADCZENIA Z POMOCY SPOŁECZNEJ/Rodziny, którym na podstawie decyzji przyznano pomoc wg przyczyn
- bezrobocie kobiet, bezrobocie mężczyzn RYNEK PRACY/BEZROBOCIE REJESTROWANE/Bezrobotni zarejestrowani wg płci i typu
- przeciętne miesięczne wynagrodzenie brutto WYNAGRODZENIA I ŚWIADCZENIA SPOŁECZNE/WYNAGRODZENIA/Przeciętne miesięczne wynagrodzenia brutto
- przestępczość kryminalna/gospodarcza/przeciwko zdrowiu i życiu - ORGANIZACJA PAŃSTWA I WYMIAR SPRAWIEDLIWOŚCI/ PRZESTĘPSTWA STWIERDZONE PRZEZ POLICJĘ W ZAKOŃCZONYCH POSTĘPOWANIACH PRZYGOTOWAWCZYCH/ Przestępstwa stwierdzone przez Policję w zakończonych postępowaniach przygotowawczych wg powiatów (dane kwartalne)

Wprowadzenie do badań

Badania w projekcie bazują na danych z roku 2018 z podziałem na powiaty. Do wyestymowania modelu zostały użyte dane uczące, które stanowią 90% wszystkich danych, natomiast do prognozy dane testowe czyli pozostałe 10% wszystkich danych.

Analizowane dane są danymi przekrojowymi. Analiza została wykonana z programie GRETL. We wszystkich testach statystycznych poziom istotności został przyjęty w wysokości 5%.

W każdym teście została zbadana normalność rozkładu reszt modelu. W przypadku jeśli test nie wskazywał p-value bliskie zeru powołując się na Centralne Twierdzenie Graniczne, stwierdzam, że w przypadku mojej próby, która jest duża zakładam, że reszty pochodzą z rozkładu normalnego.

Hipotezy

- Przemoc w rodzinie istotnie wpływa na liczbę rozwodów
- Im większe przeciętne wynagrodzenie brutto tym więcej rozwodów
- Im większe bezrobocie wśród kobiet tym więcej rozwodów
- Wskaźnik gęstości zaludnienia na 1km^2 ma istotny wpływ na liczbę rozwodów
- Alkoholizm w rodzinie istotnie wpływa na liczbę rozwodów

2. Teoria

Klasyczna metoda najmniejszych kwadratów

Klasyczną metoda najmniejszych kwadratów (KMNK) jest jedną z najważniejszych i najstarszych metod obliczeniowych w statystyce. Metoda ta ma na celu wyznaczenie linii regresji, linii trendu dla zebranych danych. Jest stosowana ona zarówno do oszacowania zależności liniowej jak również nieliniowej. Ma ona na celu dopasowanie do zebranych danych, pary wyników takiej linii prostej (model liniowy), która jest do nich najlepiej dopasowana (obliczeniowo). Metoda ta wyprowadza taką linię prostą, dla której suma kwadratów tych błędów będzie najniższa, czyli dopasowuje taką linię do zebranych danych, aby ogólny błąd oszacowania (dla wszystkich danych) był jak najmniejszy. Każda inna linia, o innym nachyleniu, wartości początkowej po obliczeniach, dostarczałaby większy błąd oszacowania[1].

Założenia KMNK są następujące:

- Szacowany model ekonometryczny jest liniowy względem parametrów α_j.
- Zmienne objaśniające X_i są wielkościami nielosowymi o ustalonych elementach.
- Rząd macierzy $\boldsymbol{X}$ równy jest liczbie szacowanych parametrów, czyli $r(\boldsymbol{X}) = k + 1$.
- Liczebność próby jest większa niż liczba szacowanych parametrów, tzn. $n \geq k + 1$.
- Nie występuje zjawisko współliniowości pomiędzy zmiennymi objaśniającymi.
- Wartość oczekiwana składnika losowego jest równa zero: $\forall_t \; E(\epsilon_t) = 0$.
- Składnik losowy ma stałą skończoną wariancję $\forall_t \; D^2(\epsilon_t) = \sigma^2$;
- Nie występuje zjawisko autokorelacji składnika losowego, czyli zależności składnika losowego w różnych jednostkach czasu $\forall_{t \neq s} \; cov(\epsilon_t, \epsilon_s) = 0$.
- Składnik losowy ma *n*-wymiarowy rozkład normalny: $\epsilon_t : N(0, \sigma^2)$ dla *t=1,2,…,n.*

Metoda Hellwiga

Metoda wskaźników pojemności informacyjnej (metoda optymalnego wyboru predykat, metoda Hellwiga) jest jedną z metod wyboru zmiennych objaśniających do modelu ekonometrycznego. Jest to jedna z metod wyboru do modelu tych zmiennych, które są silnie skorelowane ze zmienną objaśnianą, a słabo pomiędzy sobą[2] . Wybór ten jest dokonywany się poprzez znalezienie maksimum tak zwanych integralnych wskaźników pojemności informacyjnej, obliczanych dla każdej z kombinacji k potencjalnych zmiennych objaśniających. Indywidualne oraz integralne wskaźniki pojemności informacyjnej są unormowane tzn. przyjmują wartości z przedziału [0,1]. Do modelu

[1] O. Blanchard, Makroekonomia, Oficyna Ekonomiczna Grupa Wolters Kluwer, Warszawa 2011

[2] Hellwig, Z., On the Optimal Choice of Predictors, [w:] Study VI, Toward a System of Quantitative Indicators of Components of Human Resources Development, UNESCO, Paris 1968.

ekonometrycznego wybierana jest kombinacja zmiennych objaśniających, której odpowiada maksymalna wartość integralnego wskaźnika pojemności informacyjnej.[3]

Koincydentność

Zmienna objaśniająca X jest koincydentna w modelu określonym przez regularną parę korelacyjną (R, R0) wtedy i tylko wtedy, gdy

$$r_1 > [R_i^i]^T [R_{ii}]^{-1} R_o^i$$

gdzie R_i^j oznacza i-tą kolumnę macierzy R z pominięciem i-tej współrzędnej, R_0^i powstaje z wektora R0 przez odrzucenie i-tej w spółrzędnej, natomiast R0 oznacza podmacierz otrzymaną z macierzy R przez skreślenie wiersza oraz kolumny o numerze i-tym .

Model jest koincydentny, jeżeli każda zmienna objaśniająca tego modelu ma własność koincydencji. [4]

Współliniowość

Przeprowadzone przez J. Jakubczyca badania potwierdziły przypuszczenia, iż w świetle przyjętych kryteriów „dobroci" miernika współliniowości, miernikami najlepszymi okazały się współczynnik korelacji wielowymiarowej R oraz wskaźnik uwarunkowania macierzy X, który jest równy pierwiastkowi kwadratowemu z ilorazu maksymalnej i minimalnej wartości własnej macierzy XTX. Współliniowość zachodzi wówczas, gdy wartość tego wskaźnika jest odpowiednio wysoka. Niestety, wskaźnik uwarunkowania nie jest miarą unormowaną (wada ta powoduje znaczną uciążliwość w wyciąganiu końcowych wniosków). Jest także zależny od skali (rzędu wartości zmiennych objaśniających). Wady te można w pewnym stopniu wyeliminować. [5]

Efekt Katalizy

Mówimy, że w modelu ekonometrycznym określonym przez regularną parę korelacyjną (R, R0) występuje efekt katalizy, jeżeli istnieje taka para wskaźników (i, j), dla której

$$r_{ij} < 0 \quad \text{lub} \quad r_{ij} > \frac{r_i}{r_j}$$

, gdzie $R = [r_{ij}]_{k \times k}$, $R_o = [r_i]_{k \times 1}$ są odpowiednio macierzą korelacji dla zmiennych objaśniających i wektorem , którego i-tą w spółrzędną jest współczynnik korelacji między i-tą zmienną objaśniającą a zmienną objaśnianą. Ponieważ efekt katalizy może występować w modelu z różnym natężeniem , dlatego został określony miernik $\eta = r^2 - H,$

[3] „The Role of Informatics in Economic and Social Sciences Innovations and Interdisciplinary Implications" - ZBIGNIEW E. ZIELIFSKI

[4] MAKSYMIAK, Elżbieta. "ANN ALES UNI VERSIT ATIS MARIAE CURIE-SKŁODOWSKA LUBLIN—POLONIA."

[5] J. Jakubczyca, Współliniowość statystyczna, Warszawa 1987

gdzie r^2 jest kwadratem współczynnika korelacji wielorakiej, zaś H jest pojemnością integralną informacji służący do mierzenia natężenia efektu katalizy. Na podstawie nierówności $0 \leqslant \eta \leqslant 1$ widzimy, że w modelu nie występuje efekt katalizy gdy n = 0, natomiast natężenie efektu katalizy osiąga wartość największą, gdy n= 1.[6]

Test Ramseya RESET

Test RESET (Regression Specification Error Test) został zaproponowany w 1969 roku przez Ramseya i jest on stosowany jako test diagnostyczny. Układ hipotez w tym teście jest następujący:

H0: zależność liniowa

H1: zależność wielomianowa stopnia k.

Procedura testowania liniowości przebiega w kilku etapach. W etapie pierwszym szacuje się regresję postaci: y = Xβ + ε,

gdzie: X jest macierzą o wymiarach n × d obserwacji na zmiennych objaśniających modelu. Z tego modelu zapisuje się reszty e oraz wartości teoretyczne yˆ . Jeżeli H0 jest prawdziwa, wówczas ε jest procesem o średniej równej 0, w sytuacji gdy nie jest prawdziwa, średnia ta jest niezerowa. W kroku ostatnim testuje się hipotezę o liniowości postaci:

$$H_0: c_2 = c_3 = \ldots = c_k = 0.$$

Statystyka sprawdzająca testu RESET przyjmuje postać:

$$F = \frac{\left(\sum e_t^2 - \sum u_t^2\right)/(k-1)}{\sum u_t^2 /(n-(d+k-1))},$$

która przy założeniu prawdziwości H0 ma asymptotyczny rozkład $F_{k-1,n-(d+k-1)}$.[7]

Uogólniony test Walda

Pozwala zbadać czy uwzględnienie zmiennych w modelu ma podstawy statystyczne.

H0: $a_{k+1}=\ldots=a_{k+m}=0$

H1: co najmniej jeden z parametrów aj (j= k+1,…,k+3) jest różny od zera

[6] Borowiecki R., Kaliszyk J., Kolupa M., Koicydencja i efekt katalizy w liniowych modelach ekonometrycznych, Biblioteka ekonometryczna, PWN Warszawa 1986.

[7] Śliwicki, Dominik. "Jądrowy test liniowości." *Acta Universitatis Nicolai Copernici Ekonomia* 43.2 (2012)

Test liczby serii

Za pomocą testów serii można weryfikować wiele różnorodnych hipotez, na przykład, że:

- obserwacje w próbie są niezależne (testy losowości),

- dwie lub więcej populacji ma ten sam rozkład,

- model regresji jest liniowy. [8]

H0: próba jest losowa

H1: próba nie jest losowa

Test Chowa

Formalna konstrukcja testu jest następująca:

$$F = \frac{e^T \cdot e - e_1^T \cdot e_1 - e_2^T \cdot e_2}{e_1^T \cdot e_1 + e_2^T \cdot e_2} \cdot \frac{n_1 + n_2 - 2 \cdot k}{k} \qquad (3)$$

gdzie:

$e_1^T \cdot e_1$ – suma kwadratów reszt dla modelu (1) z pierwszego podokresu,

$e_2^T \cdot e_2$ – suma kwadratów reszt dla modelu (1) z drugiego podokresu,

n_1 – liczba danych w pierwszym podokresie,

n_2 – liczba danych w drugim podokresie,

k – liczba szacowanych parametrów,

e – wektor reszt modelu, który powstaje przez odjęcie od rzeczywistej wartości ceny akcji ceny obliczonej na podstawie oszacowanego modelu.

Stosując test Chowa, zakłada się, że wariancje w podokresach są równe, a stopy zwrotu charakteryzują się rozkładem normalnym.[9]

H0: parametry modelu są stabilne

H1: parametry modelu nie są stabilne

[8] Domański, Czesław. "Moc testów losowości opartych na liczbie serii wielokrotnych." (2002).
[9] Tarczyński, Waldemar. "O pewnym sposobie wyznaczania współczynnika beta na polskim rynku kapitałowym." *Zeszyty Naukowe Uniwersytetu Szczecińskiego* 561 (2009)

Test White'a

White stworzył test statystyczny, którego technicznie przeprowadza się następująco:

• Szacujemy model (model podstawowy)

• Obliczamy reszty e oraz ich kwadraty . Będą one reprezentować wartości wariancji składnika losowego

• Szacujemy pomocniczy model, w którym zmienną objaśnianą są wartości wariancji (obserwacje reprezentowane są przez kwadraty reszt), a zmiennymi objaśniającymi wszelkie możliwe niepowtarzające się kombinacje iloczynów zmiennych objaśniających modelu podstawowego.

• Obliczamy statystykę White, która ma postać n*R2, gdzie n – liczba obserwacji. Statystyka ta ma rozkład χ^2 z liczbą stopni swobody, równą liczbie zmiennych objaśniających w modelu pomocniczym.[10]

H0: reszty modelu są homoskedastyczne

H1: reszty modelu są hetersoskedastyczne

Test Breuscha – Pagana

W celu porównania modelu z efektami losowymi z modelem klasycznym wykorzystuje się test Breuscha-Pagana. Służy on do weryfikacji założenia o stałości wariancji składnika losowego. W przypadku, gdy wariancja składnika losowego efektów indywidualnych jest różna od zera, właściwszym estymatorem jest estymator z efektami losowymi.[11]

H0: reszty modelu są homoskedastyczne

H1: reszty modelu są hetersoskedastyczne

[10] Czapkiewicz, Anna. "Ekonometria."
[11] Korol, Janusz, and Przemysław Szczuciński. "Ekonometryczne modelowanie zróżnicowania związków w sektorze małych i średnich przedsiębiorstw w przestrzeni regionalnej." *Studia i Prace Wydziału Nauk Ekonomicznych i Zarządzania/Uniwersytetu Szczecińskiego, Szczecin* (2012).

3. Zdefiniowanie analizowanych zmiennych

Opis zmiennych

divorces	zmienna objaśniana - liczba rozwodów
density	gęstość zaludnienia przypadającego na 1km^2 większa niż 100
fem	feminizacja - liczba kobiet przypadająca na 100 mężczyzn
violence	przemoc w rodzinie -Rodziny, którym na podstawie decyzji przyznano pomoc
alcoholism	alkoholizm w rodzinie - Rodziny, którym na podstawie decyzji przyznano pomoc
drugs	narkomania w rodzinie - Rodziny, którym na podstawie decyzji przyznano pomoc
funemployed	bezrobotne zarejestrowane kobiety
munemplyed	bezrobotni zarejestrowani mężczyźni
salary	przeciętne miesięczne wynagrodzenie brutto
c_criminal	Przestępstwa kryminalne stwierdzone przez Policję w zakończonych postępowaniach przygotowawczych
c_economic	Przestępstwa gospodarcze stwierdzone przez Policję w zakończonych postępowaniach przygotowawczych
c_life	Przestępstwa dotyczące życia i zdrowia stwierdzone przez Policję w zakończonych postępowaniach przygotowawczych

Tabela 1

Zmienna objaśniana – divorces

Zmienna ilościowa - liczba rozwodów w Polsce ze względu na powiaty w 2018.

Statystyka	Wartość
Średnia	160,28
Mediana	113
Min	27
Max	3594
Odch. Standardowe	239,09
Skośność	9,8339
Kurtoza	127,36

Tabela 2

Średnia liczba rozwodów w Polsce wynosi 160,28. Wyniki odchylają się od średniej przeciętnie o około 239 rozwodów. Można zaobserwować skośność prawostronną.

Zmienne objaśniające

Każda zmienna określona jest dla powiatów w 2018 roku.

Density

Zmienna binarna - gęstość zaludnienia przypadającego na 1km^2 większa niż 100

1 – większa

0 - mniejsza

Statystyka	Wartość
Średnia	0,4649
Mediana	0
Min	0
Max	1
Odch. Standardowe	0,4995
Skośność	0,14070
Kurtoza	-1,9802

Tabela 3

Możemy zauważyć że przeważa gęstość zaludnienia poniżej 100, ale dane mimo wszystko są prawie równo rozłożone.

Fem

Zmienna ilościowa - feminizacja - liczba kobiet przypadająca na 100 mężczyzn

Statystyka	Wartość
Średnia	104,74
Mediana	104
Min	96
Max	119
Odch. Standardowe	3,8739
Skośność	0,95954
Kurtoza	0,66356

Tabela 4

Możemy stwierdzić, że średnio na 100 mężczyzn przypada 104 kobiet, więc więcej jest kobiet. Wyniki odchylają się od średniej przeciętnie o około 3,8739.

Violence

Zmienna ilościowa - przemoc w rodzinie -Rodziny, którym na podstawie decyzji przyznano pomoc

Statystyka	Wartość
Średnia	38,708
Mediana	23
Min	1
Max	610
Odch. Standardowe	53,117
Skośność	5,3463
Kurtoza	43,579

Tabela 5

Średnia liczba przypadków przemocy w rodzinie wynosi 38,708 . Wyniki odchylają się od średniej przeciętnie o około 53,117. Można zaobserwować skośność prawostronną.

Alcoholism

Zmienna ilościowa - alkoholizm w rodzinie - Rodziny, którym na podstawie decyzji przyznano pomoc

Statystyka	Wartość
Średnia	183,99
Mediana	153,50
Min	9
Max	1795
Odch. Standardowe	153,98
Skośność	4,6031
Kurtoza	37,643

Tabela 6

Średnia liczba przypadków alkoholizmu w rodzinie wynosi 183,99. Wyniki odchylają się od średniej przeciętnie o około 153,98. Można zaobserwować skośność prawostronną.

Drugs

Zmienna ilościowa - narkomania w rodzinie - Rodziny, którym na podstawie decyzji przyznano pomoc

Statystyka	Wartość
Średnia	11,345
Mediana	6
Min	0
Max	410
Odch. Standardowe	25,310
Skośność	11,921
Kurtoza	179,48

Tabela 7

Średnia liczba przypadków narkomanii w rodzinie wynosi 11,345. Wyniki odchylają się od średniej przeciętnie o około 25,31. Można zaobserwować skośność prawostronną.

Funemployed

Zmienna ilościowa - bezrobotne zarejestrowane kobiety

Statystyka	Wartość
Średnia	1160,2
Mediana	907,00
Min	165,00
Max	9685
Odch. Standardowe	1013,6
Skośność	4,5148
Kurtoza	30,331

Tabela 8

Średnia liczba bezrobotnych kobiet wynosi 1160,2. Wyniki odchylają się od średniej przeciętnie o około 1013,6. Można zaobserwować skośność prawostronną.

Munemplyed

Zmienna ilościowa - bezrobotni zarejestrowani mężczyźni

Statystyka	Wartość
Średnia	1458,1
Mediana	1204,5
Min	181
Max	9707,0
Odch. Standardowe	1059,6
Skośność	3,6149
Kurtoza	21,269

Tabela 9

Średnia liczba bezrobotnych mężczyzn wynosi 1458,1. Wyniki odchylają się od średniej przeciętnie o około 1059,6. Można zaobserwować skośność prawostronną.

Salary

Zmienna ilościowa - przeciętne miesięczne wynagrodzenie brutto

Statystyka	Wartość
Średnia	4144,5
Mediana	4015,4
Min	3183,3
Max	8121,1
Odch. Standardowe	577,22
Skośność	2,7284
Kurtoza	12,366

Tabela 10

Średnie miesięczne wynagrodzenie brutto wynosi 4144,5. Wyniki odchylają się od średniej przeciętnie o około 577,22. Można zaobserwować skośność prawostronną.

c_criminal

Zmienna ilościowa - Przestępstwa kryminalne stwierdzone przez Policję w zakończonych postępowaniach przygotowawczych

Statystyka	Wartość
Średnia	1284,0
Mediana	802
Min	166
Max	35618
Odch. Standardowe	2400,7
Skośność	9,9978
Kurtoza	127,34

Tabela 11

Średnia liczba przestępstw kryminalnych wynosi 1284,0. Wyniki odchylają się od średniej przeciętnie o około 2400,7. Można zaobserwować skośność prawostronną.

c_economic

Zmienna ilościowa - Przestępstwa gospodarcze stwierdzone przez Policję w zakończonych postępowaniach przygotowawczych

Statystyka	Wartość
Średnia	480,87
Mediana	208,5
Min	27
Max	8531
Odch. Standardowe	860,89
Skośność	5,6520
Kurtoza	42,191

Tabela 12

Średnia liczba przestępstw gospodarczych wynosi 480,87. Wyniki odchylają się od średniej przeciętnie o około 860,89. Można zaobserwować skośność prawostronną.

c_life

Zmienna ilościowa - Przestępstwa dotyczące życia i zdrowia stwierdzone przez Policję w zakończonych postępowaniach przygotowawczych

Statystyka	Wartość
Średnia	46,202
Mediana	32
Min	3
Max	839
Odch. Standardowe	62,799
Skośność	7,4600
Kurtoza	78,949

Tabela 13

Średnia liczba przestępstw dotyczących zdrowia i życia wynosi 46,202. Wyniki odchylają się od średniej przeciętnie o około 62,799. Można zaobserwować skośność prawostronną.

Współczynnik zmienności

Współczynnik zmienności, podobnie jak odchylenie standardowe, jest miarą rozproszenia i informuje o stopniu zróżnicowania wartości zmiennej.

Zmienna	Współczynnik zmienności
divorces	1,4917
density	1,0744
fem	0,036985
violence	1,3722
alcoholism	0,83688
drugs	2,2309
funemployed	0,87369
munemplyed	0,72669
salary	0,13927
c_criminal	1,8697
c_economic	1,7903
c_life	1,3592

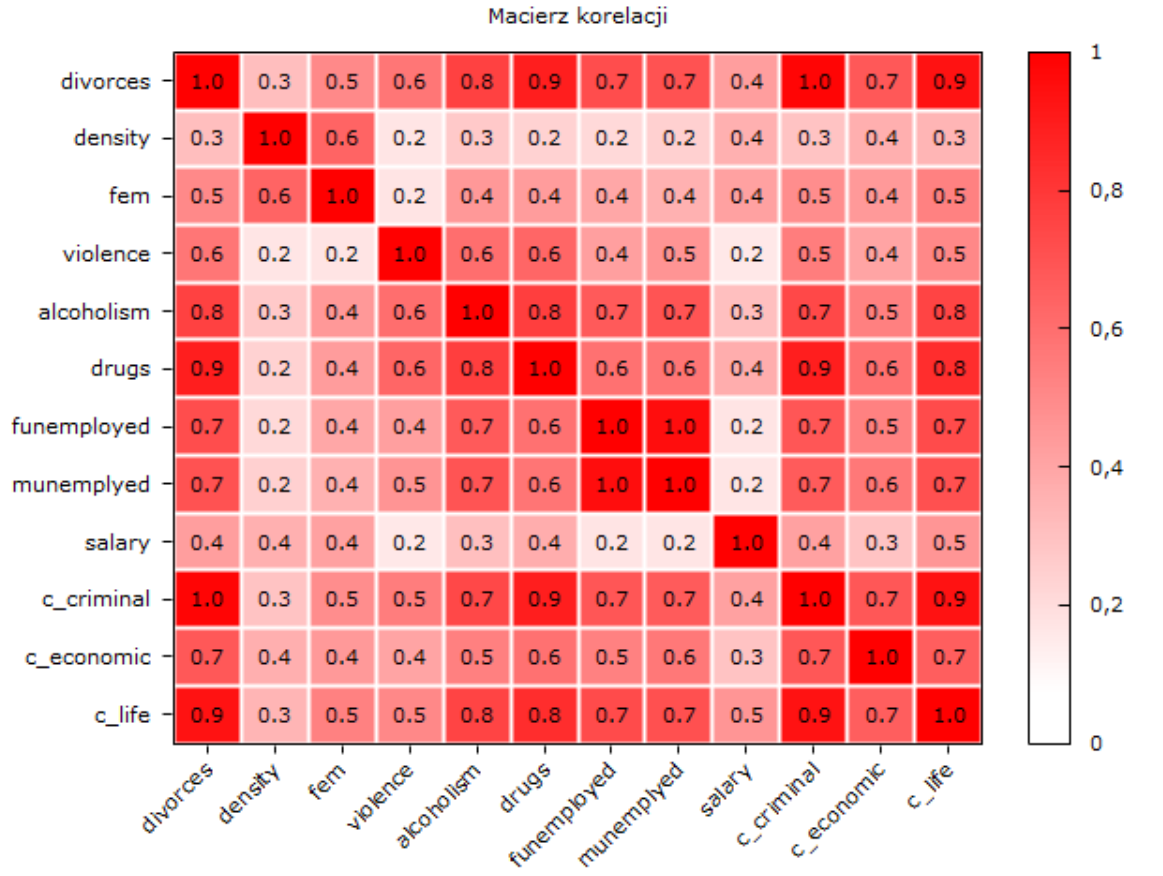
Współczynnik zmienności

Żeby uwzględnić zmienną w modelu, wartość współczynnika zmienności nie powinna być mniejsza niż 10%, w przeciwnym wypadku świadczy to o jednorodności badanej cechy to znaczy, że zmienna

jest quasi-stała. Wśród naszych zmiennych największą zmiennością wykazuje zmienna drugs, ponieważ aż 223%, a najmniejszą zmienna fem, której wartość wynosi zaledwie 3,7% więc nie będziemy jej brać pod uwagę przy tworzeniu modelu. Reszta zmiennych ma wartość współczynnika większy niż 10% więc na ten moment żadnej innej zmiennej nie odrzucamy z modelu.

Macierz korelacji

Zmienne w modelu ekonometrycznym powinny charakteryzować się słabą korelacją między sobą oraz silnym skorelowaniem między zmienną objaśnianą. Na podstawie mapy ciepła korelacji między danymi możemy wstępnie zobaczyć, które dane będą brane pod uwagę w modelu. Im bardziej czerwony kolor tym bardziej dana zmienna jest skorelowana z drugą.



Macierz korelacji

Zmienne najsilniej skorelowane ze zmienną objaśnianą to: c_criminal, c_life, drugs, ale są również mocno skorelowane między sobą dlatego jest małe prawdopodobieństwo, ze w ostateczności znajdą się wszystkie w modelu.

Najmniej skorelowane ze zmienną objaśnianą to: density co może wskazywać na to, że ta zmienna nie pojawi się w modelu.

4. Budowanie modelu

Pierwsza estymacja

Do wykonania pierwszej estymacji biorę pod uwagę wszystkie zmienne.

```
Model 1: Estymacja KMNK, wykorzystane obserwacje 1-342
Zmienna zależna (Y): divorces

              współczynnik   błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------
  const        -94,1646         90,7675          -1,037     0,3003
  density        4,84111         6,19491          0,7815    0,4351
  fem            0,516226        0,888437         0,5810    0,5616
  violence       0,114492        0,0586775        1,951     0,0519    *
  alcoholism     0,0425231       0,0277483        1,532     0,1264
  drugs          0,598469        0,231480         2,585     0,0102    **
  funemployed    0,0141482       0,00795441       1,779     0,0762    *
  munemplyed     0,00432432      0,00770710       0,5611    0,5751
  salary         0,0121233       0,00475395       2,550     0,0112    **
  c_criminal     0,0788424       0,00335434      23,50      1,81e-072 ***
  c_economic     0,000112380     0,00379746       0,02959   0,9764
  c_life         0,104502        0,116314         0,8985    0,3696

Średn.aryt.zm.zależnej  160,2836   Odch.stand.zm.zależnej  239,0940
Suma kwadratów reszt    573197,4   Błąd standardowy reszt  41,67688
Wsp. determ. R-kwadrat  0,970596   Skorygowany R-kwadrat   0,969615
F(11, 330)              990,2552   Wartość p dla testu F   3,3e-245
Logarytm wiarygodności -1754,811   Kryt. inform. Akaike'a  3533,622
Kryt. bayes. Schwarza   3579,639   Kryt. Hannana-Quinna    3551,954

Wyłączając stałą, największa wartość p jest dla zmiennej 11 (c_economic)
```

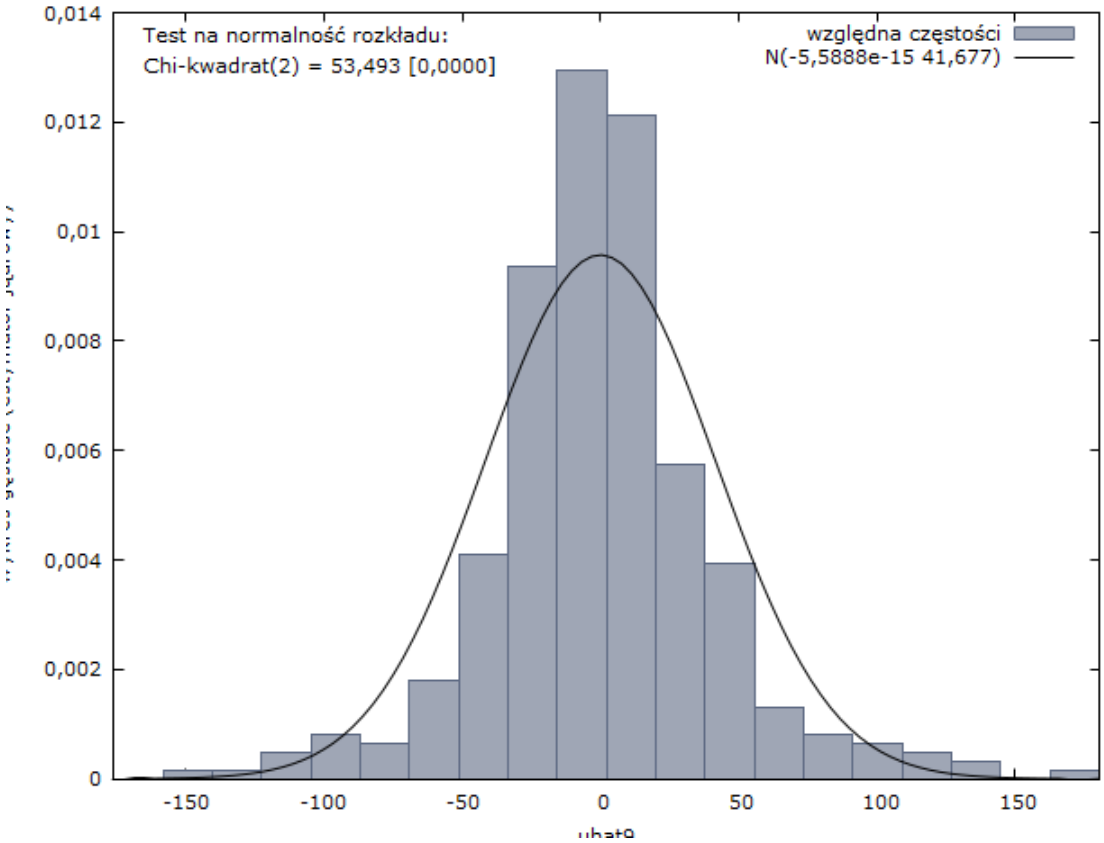
Pierwszy model MNK

Aby upewnić się, że interpretacja naszego modelu będzie poprawna sprawdzamy normalność rozkładu reszt.

Test Doornika-Hansena (1994)

H0: Reszty mają rozkład normalny

H1: Reszty nie maja rozkładu normalnego



Test rozkładu reszt pierwszego modelu MNK

Powołując się na Centralne Twierdzenie Graniczne mogę stwierdzić, ze przy tak licznej próbie reszty mają rozkład normalny.

Z wcześniejszych spostrzeżeń na temat współczynnika zmienności zmiennej fem należy odrzucić tą zmienną z modelu, ponieważ p-value w modelu MNK wynosi prawie 1.

Metody doboru zmiennych

Model ekonometryczny będę budować na podstawie dwóch metod: metody krokowej – wstecz oraz metody Hellwiga

Metoda krokowa – wstecz

Metoda krokowa – wstecz polega na kolejnym usuwaniu zmiennych z modelu, które są nieistotne. Zaczynam od modelu, w którym nie występuje już zmienna fem. Następnie po każdorazowym sprawdzeniu normalności rozkładu reszt (również mogę założyć, że dla tak dużej próbki reszty pochodzą z rozkładu normlanego) usuwam z modelu: c_economic, munemployed, c_life, density,. Ostateczny wygenerowano następujący model

```
Model 6: Estymacja KMNK, wykorzystane obserwacje 1-342
Zmienna zależna (Y): divorces

              współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------
  const        -51,4488        18,4153          -2,794    0,0055    ***
  violence       0,111517       0,0563640        1,979    0,0487    **
  alcoholism     0,0556404      0,0262489        2,120    0,0348    **
  drugs          0,549762       0,226695         2,425    0,0158    **
  funemployed    0,0194210      0,00339778       5,716    2,42e-08  ***
  salary         0,0154099      0,00438362       3,515    0,0005    ***
  c_criminal     0,0814192      0,00237897      34,22     2,17e-111 ***

Średn.aryt.zm.zależnej  160,2836   Odch.stand.zm.zależnej  239,0940
Suma kwadratów reszt    580028,1   Błąd standardowy reszt  41,61042
Wsp. determ. R-kwadrat  0,970245   Skorygowany R-kwadrat   0,969712
F(6, 335)               1820,614   Wartość p dla testu F   2,8e-252
Logarytm wiarygodności -1756,837   Kryt. inform. Akaike'a  3527,673
Kryt. bayes. Schwarza   3554,517   Kryt. Hannana-Quinna    3538,367
```

Wyestymowany model MNK

Ostateczny model:

$$Divorces = 0{,}111517 * violence + 0{,}0556404 * alcoholism + 0{,}549762 * drugs + 0{,}0194210 * funemployed + 0{,}0154099 * salary + 0{,}0814192 * c\_criminal - 51{,}4488 + \varepsilon$$

Metoda Hellwiga

Wykorzystuję do tej metody skrypt Gretla:

```
? H_max
0,96237225
? najlepszalista
c_criminal
```

Metoda Hellwiga

Metoda Hellwiga wskazuje nam, że najlepsze wyniki estymacji zostaną osiągnięte po uwzględnieniu tylko zmiennej c_criminal. Reszty modelu pochodzą z rozkładu normalnego:

```
Model 7: Estymacja KMNK, wykorzystane obserwacje 1-342
Zmienna zależna (Y): divorces

              współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------
  const         34,8352          2,84917          12,23     9,69e-029 ***
  c_criminal     0,0977006       0,00104771       93,25     3,02e-244 ***

Średn.aryt.zm.zależnej  160,2836   Odch.stand.zm.zależnej  239,0940
Suma kwadratów reszt    733499,9   Błąd standardowy reszt  46,44731
Wsp. determ. R-kwadrat  0,962372   Skorygowany R-kwadrat   0,962262
F(1, 340)               8695,884   Wartość p dla testu F   3,0e-244
Logarytm wiarygodności -1796,979   Kryt. inform. Akaike'a  3597,958
Kryt. bayes. Schwarza   3605,628   Kryt. Hannana-Quinna    3601,013
```

Model – Metoda Hellwiga

Postać modelu:

$$Divorces = 0{,}0977006 * c\_criminal + 34{,}8352 + \varepsilon$$

Porównywanie modeli

Przy wyborze najbardziej optymalnego modelu porównam kryteria informacyjne oraz skorygowaną wartość współczynnika R^2 w modelach wyestymowanych wcześniej.

Statystyka	Metoda Krokowa	Metoda Hellwiga
R^2	0,970245	0,962372
Skorygowane R^2	0,969712	0,962262
Kryt. inform. Akaike'a	3527,673	3597,958
Kryt. Hannana-Quinna	3538,367	3601,013
Kryt. bayes. Schwarza	3554,517	3605,628

Porównanie modelu

W przypadku modelu wyestymowanego za pomocą metody krokowej współczynnik determinacji R^2 i skorygowany R^2 jest wyższy niż w modelu stworzonego za pomocą metody Hellwiga. Również kryteria informacyjne w każdym przypadku są niższe w metodzie krokowej niż w metodzie Hellwiga co przeważa na korzyść modelu z metody krokowej tak jak wyższy współczynnik R^2 dla metody krokowej. Na tej podstawie możemy wybrać model wyestymowany przy pomocy metody krokowej również ze względu na większą ilość zmiennych.

Koincydencja i współliniowość

Koincydentność

Weryfikacja koincydentności w modelu

Zmienna	Znak współczynnika	Znak korelacji
violence	+	+
alcoholism	+	+
drugs	+	+
funemplyed	+	+
salary	+	+
c_criminal	+	+

Koincydentność

Nasz model ma takie same znaki dla zmiennych więc jest koincydentny.

Współliniowość

Weryfikacja współliniowości za pomocą wskaźnika VIF. Wartości > 10.0 mogą wskazywać na problem współliniowości.

```
    violence    1,765
  alcoholism    3,217
       drugs    6,484
 funemployed    2,336
      salary    1,261
  c_criminal    6,424

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

Współliniowość

Dla żadnej zmiennej wskaźnik vif jest większy niż 10 dlatego możemy stwierdzić, że nie zachodzi w współliniowość. Model pozostaje bez zmian.

Efekt katalizy

W celu sprawdzenia efektu katalizy korzystam ze skryptu Gretla. Ostateczny wynik skryptu wskazuje na brak katalizatorów.

```
> print "KATALIZATOR:"
> katalizator
> print "W PARZE:"
> w_parze
> endif
> endif
> endif
> endloop
> endloop
```

Efekt katalizy

Ostateczna transformacja modelu

Mimo, że w modelu pozostała spora ilość zmiennych egzogenicznych próbuję zweryfikować czy mogę dołączyć do modelu statystycznie istotną odrzuconą wcześniej zmienną.

Zmienną c_economic, c_life, munemployed podniosłam do kwadratu. Przy próbie dołączenia ich do modelu zmienne c_economic, c_life, munemployed w dalszym ciągu okazały się nieistotne natomiast zmienna c_life mogłaby zostać użyta w modelu, ale ze względu na wystarczającą ilość zmiennych i przebudowanie całkowite modelu, które ta zmienna powoduje (gdzie współcznnik R^2 nie zmienia się praktycznie) nie włączam ten zmiennej do modelu.

Pomijam także odrzuconą wcześniej zmienną binarną density, ponieważ przekształcenie tej zmiennej niewiele zmieni w modelu.

Ostatecznie model zostaje niezmienny.

5. Weryfikacja statystyczna modelu

Test Ramseya – RESET

```
Test RESET na specyfikację (kwadrat i sześcian zmiennej)
Statystyka testu: F = 5,942308,
z wartością p = P(F(2,333) > 5,94231) = 0,00291

Test RESET na specyfikację (tylko kwadrat zmiennej)
Statystyka testu: F = 9,717377,
z wartością p = P(F(1,334) > 9,71738) = 0,00198

Test RESET na specyfikację (tylko sześcian zmiennej)
Statystyka testu: F = 11,109272,
z wartością p = P(F(1,334) > 11,1093) = 0,000955
```

Test Ramseya

W każdym przypadku p-value odrzucamy hipotezę zerową, więc musimy dokonać transformacji naszego modelu.

Do modelu dodaję kwadraty zmiennych objaśniających i po ostatecznym wyborze istotnych zmiennych przy pomocy MNK i po sprawdzeniu koincydentności(nie włączam do modelu sq-funemployed), współliniowości i efektu katalizy model prezentuje się następująco:

```
Model 15: Estymacja KMNK, wykorzystane obserwacje 1-342
Zmienna zależna (Y): divorces

               współczynnik   błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------
  const         -50,2384         18,1691           -2,765    0,0060    ***
  funemployed     0,0189253       0,00312921        6,048    3,90e-09  ***
  c_criminal      0,0807701       0,00208018       38,83     1,91e-126 ***
  salary          0,0185628       0,00436256        4,255    2,71e-05  ***
  alcoholismsq    0,000137792     2,33320e-05       5,906    8,58e-09  ***

Średn.aryt.zm.zależnej  160,2836   Odch.stand.zm.zależnej  239,0940
Suma kwadratów reszt    577069,4   Błąd standardowy reszt  41,38082
Wsp. determ. R-kwadrat  0,970397   Skorygowany R-kwadrat   0,970046
F(4, 337)               2761,742   Wartość p dla testu F   4,3e-256
Logarytm wiarygodności -1755,962   Kryt. inform. Akaike'a  3521,924
Kryt. bayes. Schwarza   3541,098   Kryt. Hannana-Quinna    3529,563
```

Model MNK po transformacji po teście RESET

Wykonuję ponowny test Ramseya – RESET

```
Test RESET na specyfikację (kwadrat i sześcian zmiennej)
Statystyka testu: F = 2,477412,
z wartością p = P(F(2,335) > 2,47741) = 0,0855

Test RESET na specyfikację (tylko kwadrat zmiennej)
Statystyka testu: F = 0,371673,
z wartością p = P(F(1,336) > 0,371673) = 0,543

Test RESET na specyfikację (tylko sześcian zmiennej)
Statystyka testu: F = 1,036273,
z wartością p = P(F(1,336) > 1,03627) = 0,309
```

Ponowny test Ramseya – RESET

Jak widać po transformacji modelu wartość p-value dla tego testu nie daje powodu do odrzucenia hipotezy zerowej, więc postać naszego modelu jest dobrze dobrana.

Model prezentuje się następująco:

Badanie stabilności parametrów

Test został wykonany dwukrotnie. W podziale względem zmiennej density oraz przy podzieleniu próbki na pół

```
Test Chowa na zmiany strukturalne przy podziale próby w obserwacji 171
  F(5, 332) = 1,5994 z wartością p 0,1597
```

```
Test Chowa na strukturalne różnice poziomów ze względu na zmienną: density
  F(5, 332) = 1,92996 z wartością p 0,0889
```

Test Chowa

W obu przypadkach nie odrzucamy hipotezy zerowej co mówi nam, że parametry modelu są stabilne.

Uogólniony test Walda

Pozwala zweryfikować 2 rzeczy- pierwszą z nich jest istotność podzbioru zmiennych objaśniających, drugą natomiast istotność współczynnika determinacji.

```
Hipoteza zerowa: parametry regresji dla wskazanych zmiennych są równe zero
  const, funemployed, c_criminal, alcoholismsq
Statystyka testu: F(4, 337) = 2555,31, wartość p 1,3983le-250
```

Test Walda

Odrzucamy hipotezę zerową, więc wykorzystanie tych danych w modelu ma podstawy statystyczne.

Badanie heteroskedatyczności

Test Breusha-Pagana

```
Test Breuscha-Pagana na heteroskedastyczność
Estymacja KMNK, wykorzystane obserwacje 1-342
Zmienna zależna (Y): standaryzowane uhat^2

                współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------
  const          -2,55031        0,923157        -2,763     0,0060    ***
  funemployed     0,000703733    0,000158992      4,426     1,30e-05  ***
  c_criminal      0,000204079    0,000105692      1,931     0,0543    *
  salary          0,000653777    0,000221657      2,949     0,0034    ***
  alcoholismsq   -4,13563e-06    1,18548e-06     -3,489     0,0006    ***

  Wyjaśniona suma kwadr. = 222,217

Statystyka testu: LM = 111,108294,
z wartością p = P(Chi-kwadrat(4) > 111,108294) = 0,000000
```

Test Breusha-Pagana

Test White'a

```
Test White'a na heteroskedastyczność reszt (zmienność wariancji resztowej)
Estymacja KMNK, wykorzystane obserwacje 1-342
Zmienna zależna (Y): uhat^2

                  współczynnik   błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------------
  const            620,234         7393,94           0,08388    0,9332
  funemployed       -1,08022          3,84467        -0,2810     0,7789
  c_criminal         4,63412          1,90320         2,435      0,0154   **
  salary            -0,880285         2,86927        -0,3068     0,7592
  alcoholismsq      -0,0278398        0,0279664      -0,9955     0,3202
  sq_funemployed     0,000143445      0,000226472     0,6334     0,5269
  X2_X3              0,000130120      0,000383268     0,3395     0,7345
  X2_X4              0,000268290      0,000931623     0,2880     0,7735
  X2_X5             -2,98749e-06      2,57905e-06    -1,158      0,2476
  sq_c_criminal     -0,000160322      7,05434e-05    -2,273      0,0237   **
  X3_X4             -0,000491603      0,000415503    -1,183      0,2376
  X3_X5             -6,53939e-07      1,32972e-06    -0,4918     0,6232
  sq_salary          0,000136963      0,000266789     0,5134     0,6080
  X4_X5              4,12952e-06      6,54423e-06     0,6310     0,5285
  sq_alcoholismsq    2,48875e-08      1,40106e-08     1,776      0,0766   *

  Wsp. determ. R-kwadrat = 0,264829

Statystyka testu: TR^2 = 90,571385,
z wartością p = P(Chi-kwadrat(14) > 90,571385) = 0,000000
```

Test White'a

W obu przypadkach widzimy, że występuje problem heteroskedastyczności. Powoduje ją nasza zmienna c_criminal.

Test liczby serii

```
Test serii

Liczba serii (R) dla zmiennej 'uhat19_aaa' = 138
Test niezależności oparty na liczbie dodatnich i ujemnych serii.
Hipoteza zerowa: próba jest losowa, dla R odpowiednio N(172, 9,23309),
test z-score = -3,68241, przy dwustronym obszarze krytycznym p = 0,000231043
```

Test serii

W naszym przypadku odrzucamy hipotezę zerową co mówi nam, że próbka nie została dobrana losowo.

Pierwsza transformcja modelu

Z powodu wykazania nielosowości próby wymagana jest transformacja modelu. Za punkt początkowy biorę model wyestymowany za pomocą Hellwiga i powoli dodaję do niego zmienne. Model, który posiada zmienne takie jak c_criminal i drugs jest koincydentny, nie występują katalizatory, nie zachodzi współliniowość,

```
  c_criminal     4,832
       drugs     4,832

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

Współliniowość

test liczby serii wskazuje na losowość próby,

```
Test serii

Liczba serii (R) dla zmiennej 'uhat9_aaa' = 161
Test niezależności oparty na liczbie dodatnich i ujemnych serii.
Hipoteza zerowa: próba jest losowa, dla R odpowiednio N(172, 9,23309),
test z-score = -1,19137, przy dwustronym obszarze krytycznym p = 0,23351
```

Test serii

Test Ramseya RESET wskazuje na poprawność dobranego modelu

```
Test RESET na specyfikację (kwadrat i sześcian zmiennej)
Statystyka testu: F = 2,957357,
z wartością p = P(F(2,337) > 2,95736) = 0,0533

Test RESET na specyfikację (tylko kwadrat zmiennej)
Statystyka testu: F = 0,001283,
z wartością p = P(F(1,338) > 0,0012828) = 0,971

Test RESET na specyfikację (tylko sześcian zmiennej)
Statystyka testu: F = 0,170255,
z wartością p = P(F(1,338) > 0,170255) = 0,68
```

Test Ramseya RESET

Natomiast test Chowa wykazuje niestabilność parametrów przy podziale próbki względem zmiennej binarnej

```
Test Chowa na zmiany strukturalne przy podziale próby w obserwacji 171 -
  Hipoteza zerowa: brak zmian strukturalnych
  Statystyka testu: F(4, 334) = 1,22293
  z wartością p = P(F(4, 334) > 1,22293) = 0,30082

Test Chowa na strukturalne różnice poziomów ze względu na zmienną: density -
  Hipoteza zerowa: brak strukturalnych różnic
  Statystyka testu: F(4, 334) = 3,29246
  z wartością p = P(F(4, 334) > 3,29246) = 0,0114962
```

Test Chowa

Próba doprowadzenia modelu do właściwej postaci

Ostatecznie udało się dobrać zmienne, które spełniają wszystkie założenia natomiast problemem jest występowanie heteroskedastyczności

```
Model 16: Estymacja KMNK, wykorzystane obserwacje 1-342
Zmienna zależna (Y): divorces

                współczynnik   błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------
  const          -23,9406         18,2490          -1,312    0,1905
  c_criminal       0,0851943       0,00204507      41,66     5,69e-135 ***
  alcoholism       0,0773643       0,0292388        2,646    0,0085    ***
  sq_alcoholism    0,000108348     3,13023e-05      3,461    0,0006    ***
  salary           0,0131187       0,00453658       2,892    0,0041    ***

Średn.aryt.zm.zależnej  160,2836   Odch.stand.zm.zależnej  239,0940
Suma kwadratów reszt    626685,0   Błąd standardowy reszt  43,12307
Wsp. determ. R-kwadrat  0,967852   Skorygowany R-kwadrat   0,967470
F(4, 337)               2536,420   Wartość p dla testu F   4,7e-250
Logarytm wiarygodności -1770,066   Kryt. inform. Akaike'a  3550,133
Kryt. bayes. Schwarza   3569,307   Kryt. Hannana-Quinna    3557,771
```

Model MNK

W celu jej zlikwidowania próbuję przekształcać moje zmienne:

```
Model 3: Estymacja KMNK, wykorzystane obserwacje 1-342
Zmienna zależna (Y): divorcesh

              współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------
  const         0,423414        0,461364         0,9177    0,3594
  alcoholismh   0,0953147       0,0225969        4,218     3,17e-05  ***
  c_criminalh   0,0887543       0,00560285      15,84      1,17e-042 ***
  salaryh       0,00370421      0,00184631       2,006     0,0456    **

Średn.aryt.zm.zależnej  4,338813   Odch.stand.zm.zależnej  1,813220
Suma kwadratów reszt    462,5278   Błąd standardowy reszt  1,169797
Wsp. determ. R-kwadrat  0,587444   Skorygowany R-kwadrat   0,583783
F(3, 338)               160,4279   Wartość p dla testu F   1,17e-64
Logarytm wiarygodności -536,9012   Kryt. inform. Akaike'a  1081,802
Kryt. bayes. Schwarza   1097,142   Kryt. Hannana-Quinna    1087,913

Test White'a na heteroskedastyczność reszt (zmienność wariancji resztowej) -
  Hipoteza zerowa: heteroskedastyczność reszt nie występuje
  Statystyka testu: LM = 24,3284
  z wartością p = P(Chi-kwadrat(9) > 24,3284) = 0,00381129

Test Chowa na strukturalne różnice poziomów ze względu na zmienną: density -
  Hipoteza zerowa: brak strukturalnych różnic
  Statystyka testu: F(4, 334) = 2,54173
  z wartością p = P(F(4, 334) > 2,54173) = 0,0396747
```

Model MNK z wagami

Po przeprowadzeniu transformacji danych przez wagę zmienna/sqrt(c_criminal) model wygląda następująco. Okazuje się, że parametry modelu są niestabilne oraz dalej zachodzi heteroskedastyczność.

Przy transformacji zmiennych objaśniających na kwadraty okazuje się że test Ramseya wskazuje na niepoprawnie dobrany model oraz dalej zachodzi heteroskedastyczność.

Przy zlogarytmowaniu zmiennych objaśnianej i objaśniających test Ramseya wszystkie założenia oprócz testu Ramseya i Chowa zostają spełnione.

```
Test RESET na specyfikację (kwadrat i sześcian zmiennej)
Statystyka testu: F = 2,229259,
z wartością p = P(F(2,336) > 2,22926) = 0,109

Test RESET na specyfikację (tylko kwadrat zmiennej)
Statystyka testu: F = 4,432641,
z wartością p = P(F(1,337) > 4,43264) = 0,036

Test RESET na specyfikację (tylko sześcian zmiennej)
Statystyka testu: F = 4,471125,
z wartością p = P(F(1,337) > 4,47112) = 0,0352
```

Test Ramseya RESET

Test Chowa

```
Test Chowa na zmiany strukturalne przy podziale próby w obserwacji 171
  F(4, 334) = 3,20989 z wartością p 0,0132
```

Test Chowa

W dalszym ciągu nie uzyskałam satysfakcjonującego mnie modelu więc kontynuuję transformację modelu.

Ostateczny model

Do modelu dodaję kwadrat zlogarytmowanej zmiennej c_crminal.

```
Model 18: Estymacja KMNK, wykorzystane obserwacje 1-342
Zmienna zależna (Y): l_divorces

                  współczynnik  błąd standardowy  t-Studenta  wartość p
  ---------------------------------------------------------------------
  const             -1,44934        1,19525          -1,213    0,2261
  l_alcoholism       0,149024       0,0289762         5,143    4,59e-07  ***
  l_salary           0,390192       0,146161          2,670    0,0080    ***
  sq_l_c_criminal    0,0483243      0,00200938       24,05     3,41e-075 ***

Średn.aryt.zm.zależnej  4,762955   Odch.stand.zm.zależnej  0,699266
Suma kwadratów reszt    27,00421   Błąd standardowy reszt  0,282655
Wsp. determ. R-kwadrat  0,838046   Skorygowany R-kwadrat   0,836608
F(3, 338)               583,0037   Wartość p dla testu F   3,3e-133
Logarytm wiarygodności -51,13910   Kryt. inform. Akaike'a  110,2782
Kryt. bayes. Schwarza   125,6174    Kryt. Hannana-Quinna    116,3889
```

Model MNK – ostateczny model

- Model jest koincydenty
- Badanie współliniowości

```
   l_alcoholism    1,717
      l_salary     1,374
sq_l_c_criminal    2,184

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

Współliniowość

Nie zachodzi współliniowość

- Efekty katalizy – brak katalizatorów

- Test serii

```
Test serii

Liczba serii (R) dla zmiennej 'uhat18_aab' = 163
Test niezależności oparty na liczbie dodatnich i ujemnych serii.
Hipoteza zerowa: próba jest losowa, dla R odpowiednio N(172, 9,23309),
test z-score = -0,974755, przy dwustronym obszarze krytycznym p = 0,329682
```

Test serii

Brak podstaw do odrzucenia hipotezy zerowej – próba jest losowa

- Test Chowa

```
Test Chowa na zmiany strukturalne przy podziale próby w obserwacji 171
  F(4, 334) = 2,87579 z wartością p 0,0230


Test Chowa na strukturalne różnice poziomów ze względu na zmienną: density
  F(4, 334) = 2,33138 z wartością p 0,0557
```

Test Chowa

Przy podziale względem zmiennej binarnej denisity paramtery modelu są stabline natomiast przy podziale próby na pół odrzucamy hipotezę zerową. Przy modelu z danymi przekrojowymi ważniejsza i bardziej sensowna jest statystyka podzialu względem zmiennej binarnej, której wartość w tym drugim przypadku również jest na pograniczu. Mimo wszystko przyjmuję, że parametry modelu są stabline. Niskie wartości statystyk mogą wpłynąc na późnijesze niedoszacowania lub przesacowania.

- Test Ramseya - RESET

```
Test RESET na specyfikację (kwadrat i sześcian zmiennej)
Statystyka testu: F = 0,920653,
z wartością p = P(F(2,336) > 0,920653) = 0,399

Test RESET na specyfikację (tylko kwadrat zmiennej)
Statystyka testu: F = 1,660703,
z wartością p = P(F(1,337) > 1,6607) = 0,198

Test RESET na specyfikację (tylko sześcian zmiennej)
Statystyka testu: F = 1,559383,
z wartością p = P(F(1,337) > 1,55938) = 0,213
```

Test Ramseya RESET

Brak podstaw do odrzucenia hipotezy zerowej – model został dobrze dobrany

- Test normalności reszt

Możemy założyć z centralnego twierdzenia granicznego przy tak dużej próbie, że reszty pochodzą z rozkładu normalnego.

- Badanie heteroskedastyczności

    - Test White'a

```
Test White'a na heteroskedastyczność reszt (zmienność wariancji resztowej)
Estymacja KMNK, wykorzystane obserwacje 1-342
Zmienna zależna (Y): uhat^2

                      współczynnik  błąd standardowy  t-Studenta  wartość p
  ---------------------------------------------------------------------------
  const                17,5119         17,9903           0,9734     0,3311
  l_alcoholism          1,57561         1,06341          1,482      0,1394
  l_salary             -5,11099         4,39253         -1,164      0,2454
  sq_l_c_criminal       0,00481599      0,0770028        0,06254    0,9502
  sq_l_alcoholism       0,000769019     0,0156331        0,04919    0,9608
  X2_X3                -0,195176        0,130723        -1,493      0,1364
  X2_X4                 0,000215833     0,00201810       0,1069     0,9149
  sq_l_salary           0,364507        0,272623         1,337      0,1821
  X3_X4                -0,000461621     0,00952845      -0,04845    0,9614
  sq_sq_l_c_crimin~    -7,59579e-06     8,57809e-05     -0,08855    0,9295

  Wsp. determ. R-kwadrat = 0,035776

Statystyka testu: TR^2 = 12,235350,
z wartością p = P(Chi-kwadrat(9) > 12,235350) = 0,200364
```

Test White'a

Brak podstaw do odrzucenia hipotezy zerowej – model jest homoskedastyczny

    - Test Breusha-Pagana

```
Test Breuscha-Pagana na heteroskedastyczność
Estymacja KMNK, wykorzystane obserwacje 1-342
Zmienna zależna (Y): standaryzowane uhat^2

                  współczynnik  błąd standardowy  t-Studenta  wartość p
  -------------------------------------------------------------------------
  const           -0,137005        6,65655         -0,02058    0,9836
  l_alcoholism    -0,354644        0,161374        -2,198      0,0287   **
  l_salary         0,310410        0,813998         0,3813     0,7032
  sq_l_c_criminal  0,00695851      0,0111906        0,6218     0,5345

  Wyjaśniona suma kwadr. = 14,2247

Statystyka testu: LM = 7,112365,
z wartością p = P(Chi-kwadrat(3) > 7,112365) = 0,068401
```

Test Breausha - Pagana

Brak podstaw do odrzucenia hipotezy zerowej – model jest homoskedastyczny

Założenie MNK o stałości wariancji składnika losowego zostało spełnione.

Wnioski: Ostateczna postać modelu spełnia wszystkie założenia MNK i wszystkie istotne testy statystyczne. Na tej podstawie będzie można dokonać prognoz.

Ostateczna postać modelu

l_divorces = 0,149024*_l_alcoholism_ +0,390192* _l_salary_ + 0,0483243* _sq_l_c_criminal_ −1,44934 + ε

6. Prognoza

Prognoza ex ante

Na podstawie wartości średnich l_alcoholism, l_salary oraz sq_l_c_criminal dokonuję prognozę:

Dane wejściowe

- me_l_alcoholism = 4,98014
- me_l_salary = 8,32142
- me_sq_l_c_criminal = 46,0055

Prognoza

- Prognoza punktowa: 4,76295
- Błąd prognozy: 0,297781
- Przedział ufności: (4,17722; 5,34869)

Na podstawie prognozy możemy zauważyć, że kiedy liczba rodzin o stwierdzonym alkoholizmie w rodzinie wynoszącym 145 oraz przeciętne miesięczne wynagrodzenie brutto wynosi 4111 i liczba przestępstw wynosi 882 to liczba rozwodów w powiecie powinna wynosić około 117 przy błędzie prognozy równym 0,3.

Prognoza ex post

	l_divorces
343	5,723585
344	4,343805
345	6,598509
346	4,644391
347	4,574711
348	5,036953
349	4,691348
350	4,343805
351	4,828314
352	4,882802
353	4,189655
354	4,912655
355	4,744932
356	5,204007
357	4,912655
358	5,105945
359	6,914731
360	4,394449
361	4,465908
362	4,532599
363	5,176150
364	4,605170
365	5,129899
366	4,624973
367	4,948760
368	4,624973
369	4,875197
370	5,117994
371	4,406719
372	4,343805
373	5,375278
374	4,770685
375	4,276666
376	4,499810
377	4,248495
378	5,459586
379	6,771936
380	4,442651

Wartości rzeczywiste

	Y_prog
343	5,342623
344	4,258277
345	5,905361
346	4,340517
347	3,883821
348	4,797812
349	4,548680
350	4,590846
351	4,549550
352	4,714509
353	4,089047
354	4,443925
355	4,558535
356	5,095100
357	5,030759
358	4,752009
359	6,973516
360	4,557914
361	4,679822
362	4,681062
363	4,959289
364	5,063265
365	4,891312
366	4,484301
367	4,912016
368	4,679763
369	4,886248
370	4,911638
371	4,236051
372	4,373254
373	5,085546
374	4,657410
375	4,344472
376	4,595770
377	4,412135
378	5,270490
379	6,841445
380	4,775426

Prognoza

Błędy prognozy:

- ME = 0,0939734
- MAE = 0,211492
- RMSE = 0,264582
- MAPE = 0,0429702

Prognoza różni się od wartości rzeczywistej średnio o 4%.

Model ma tendencję do zaniżania ilości rozwodów, ponieważ średni błąd prognozy ex post ME jest dodatni. Świadczy to o niedoszacowaniu wartości prognozowanych, czyli są one przeciętnie niższe niż realna wartość zmiennej objaśnianej.

7. Weryfikacja hipotez

- Przemoc w rodzinie istotnie wpływa na liczbę rozwodów
- Im większe przeciętne wynagrodzenie brutto tym więcej rozwodów
- Im większe bezrobocie wśród kobiet tym więcej rozwodów
- Wskaźnik gęstości zaludnienia na 1km^2 ma istotny wpływ na liczbę rozwodów
- Alkoholizm w rodzinie istotnie wpływa na liczbę rozwodów

8. Interpretacja parametrów

Ostatecznie model przyjmuje postać

l_divorces = 0,149024*_l_alcoholism_ +0,390192* _l_salary_ + 0,0483243* _sq_l_c_criminal_ −1,44934 + ε

co oznacza, interpretacja niejednoznaczna spowodowana logarytmowaniem oraz potęgowaniem zmiennych, ale wiemy, że

- Wraz ze wzrostem alkoholizmu wśród rodzin liczba rozwodów rośnie o około 0,15 jednostki (ceteris paribus)
- Wraz ze wzrostem przeciętnego miesięcznego wynagrodzenia brutto liczba rozwodów rośnie o około 0,39 jednostki (ceteris paribus)
- Wraz ze wzrostem liczby przestępstw kryminalnych liczba rozwodów rośnie o około 0,05 jednostki (ceteris paribus)

9. Wnioski i podsumowanie

Przeprowadzone przeze mnie badania wykazały, że nie wszystkie, początkowo przyjęte zmienne do modelu są istotne. Ostatecznie w modelu pozostały zmienne takie jak c_criminal, alcoholizm oraz salary. Możemy więc wnioskować, że liczba rozwodów w Polsce w 2018 roku w zależności od powiatów zależy od przeciętnego miesięcznego wynagrodzenia brutto, liczby przestępstw kryminalnych stwierdzonych przez Policję w zakończonych postępowaniach przygotowawczych oraz od liczby rodzin, którym udzielono pomocy z zakresie alkoholizmu w rodzinie. Sam model jest modelem, który w ok. 85% opisuje rzeczywistość, w 85% objaśnia danych problem, co jest bardzo dobrym wynikiem.

Bibliografia

- O. Blanchard, Makroekonomia, Oficyna Ekonomiczna Grupa Wolters Kluwer, Warszawa 2011
- Hellwig, Z., On the Optimal Choice of Predictors, [w:] Study VI, Toward a System of Quantitative Indicators of Components of Human Resources Development, UNESCO, Paris 1968.
- „The Role of Informatics in Economic and Social Sciences Innovations and Interdisciplinary Implications" - ZBIGNIEW E. ZIELIFSKI
- MAKSYMIAK, Elżbieta. "ANN ALES UNI VERSIT ATIS MARIAE CURIE-SKŁODOWSKA LUBLIN—POLONIA."
- J. Jakubczyca, Współliniowość statystyczna, Warszawa 1987
- Borowiecki R., Kaliszyk J., Kolupa M., Koicydencja i efekt katalizy w liniowych modelach ekonometrycznych, Biblioteka ekonometryczna, PWN Warszawa 1986.
- Śliwicki, Dominik. "Jądrowy test liniowości." *Acta Universitatis Nicolai Copernici Ekonomia* 43.2 (2012)
- Domański, Czesław. "Moc testów losowości opartych na liczbie serii wielokrotnych." (2002).
- Tarczyński, Waldemar. "O pewnym sposobie wyznaczania współczynnika beta na polskim rynku kapitałowym." *Zeszyty Naukowe Uniwersytetu Szczecińskiego* 561 (2009)
- Czapkiewicz, Anna. "Ekonometria."
- Korol, Janusz, and Przemysław Szczuciński. "Ekonometryczne modelowanie zróżnicowania związków w sektorze małych i średnich przedsiębiorstw w przestrzeni regionalnej." *Studia i Prace Wydziału Nauk Ekonomicznych i Zarządzania/Uniwersytetu Szczecińskiego, Szczecin* (2012).

Spis tabel i rysunków

- Test Chowa
- Test Ramseya RESET
- Test White'a
- Test Breausha – Pagana
- Wartości rzeczywiste
- Prognoza