

Projekt zaliczeniowy Ekonometria

Analiza czyników wpływających na przestępczość w 50 stanach Usa

Maciej Śmiałowski

2019

Spis treści

[Wstęp 2](#_Toc11081079)

[Opis danych 3](#_Toc11081080)

[Podstawowe statystyki opisowe 4](#_Toc11081081)

[Model podstawowy 5](#_Toc11081082)

[Korelacja 5](#_Toc11081083)

[Model KMNK 6](#_Toc11081084)

[Metoda Hellwiga 6](#_Toc11081085)

[Metoda krokowo wsteczna 7](#_Toc11081086)

[Weryfikacja statystyczna modelu 7](#_Toc11081087)

[Test istotności zmiennych jako całości(Współczynnik R2) 7](#_Toc11081088)

[Efekt Katalizy 8](#_Toc11081089)

[Model po usunięciu katlizatora 8](#_Toc11081090)

[Koincydencja 9](#_Toc11081091)

[Normalność rozkładu reszt 9](#_Toc11081092)

[Stabilność parametrów 10](#_Toc11081093)

[Heteroskedastyczność składnika losowego 10](#_Toc11081094)

[Losowość próby 11](#_Toc11081095)

[Liniowość modelu 12](#_Toc11081096)

[Test RESET-Ramseya 12](#_Toc11081097)

[Współliniowość 12](#_Toc11081098)

[Obserwacja odstająca 13](#_Toc11081099)

[Wykresy pudełkowe 14](#_Toc11081100)

[Prognoza 15](#_Toc11081101)

[Podsumowanie 15](#_Toc11081102)

[Postać liniowa modelu 15](#_Toc11081103)

[Wnioski 16](#_Toc11081104)

[Bibliografia 16](#_Toc11081105)

# Wstęp

Stany zjednoczone to kraj o największym PKB na świecie świadczy to o dużej liczbie działalności, jest to jeden z powodów dla których państwo to zawsze charakteryzowało się wysoką przestępczością. Jednak tendencja w ciągu ostatniej dekady poprawiała się. Warto również wspomnieć ,iż w stanach zjednoczonych aż 2,3 mln mieszkańców znajduje się w więzieniach stanowi to około 2% całej populacji tego kraju, można zatem wnioskować ,że rząd dobrze radzi sobie ze zwalczanie przestępczości jednak taka liczba osadzonych wiąże się z olbrzymimi wydatkami na utrzymanie zakładów karnych. W niektórych jednak miastach mimo takiej liczby osadzonych wciąż odnotowywana jest wysoka ilość przestępstw.

|  |  |  |  |
| --- | --- | --- | --- |
| **State** | **City** | **Population** |  |
| **Total** |
|
| Missouri | St. Louis | 310284 | 2082.29 |
| Michigan | Detroit | 670792 | 2056.67 |
| Maryland | Baltimore | 613217 | 2027.01 |
| Tennessee | Memphis | 652765 | 2003.32 |
| Missouri | Kansas City | 484948 | 1724.31 |
| Wisconsin | Milwaukee | 595168 | 1597.36 |
| Ohio | Cleveland | 385351 | 1556.76 |
| California | Stockton4 | 309566 | 1414.56 |

*Tabela 1: Wybrane miasta i ilość popełnianych przestępstw per 100000*

# Cel projektu

Celem tego projektu będzie zbadanie wpływu czynników społeczno-geograficzno-gospodarczych na ilość przestępstw w każdym z 50 stanów USA. Podjęta zostanie próba wyestymowania modelu ekonometrycznego postaci:

**Y** *=* **α0 + α1\* x1 + α2\*x2 + ... + αn\*xn + ε**

oraz jego interpretacji.

# Opis danych

Dane pochodzą z roku 2017 i odnoszą się do 50 stanów USA mamy zatem taką samą liczbę obserwacji

Obliczenia oraz budowa modelu przeprowadzone zostały przy wykorzystaniu programu *Gretl.*

Model będzie przedstawiał zależność między zmienną objaśnianą Y, a zmiennymi objaśniającymi Xn(obok wytłuszczonej nazwy zmiennej podano jej nazwę alternatywną).W moich badaniach będę używał danych ze zmiennymi binarnymi procentowymi lub liczbowymi(per 100tys.). Kolejne zmienne będą oznaczać:

Zmienna objaśniana

**Y** Crime- ilość popełnianych przestępstw

Źródło: <https://ucr.fbi.gov/crime-in-the-u.s/2017/crime-in-the-u.s.-2017/topic-pages/tables/table-4>

Zmienne objaśniające

**X1** Homicide\_rate -ilość zabójstw

Źródło: <https://deathpenaltyinfo.org/murder-rates-nationally-and-state>

**X2** State\_tax\_rate-Stałe stawki podatku od sprzedaży od 1 stycznia 2017 r.

Źródło: <https://taxfoundation.org/state-and-local-sales-tax-rates-in-2017/>

**X3**  License\_to\_owngun-zmienna binaran 1 jeśli w stanie tym trzeba posiadać licencję by mieć broń

Źródło: <https://lawcenter.giffords.org/search-gun-law-by-state/>

**X4** Border\_on\_Mexico-zmienna binarna 1 jeśli stan graniczy z Meksykiem

Źródło: <https://pl.wikipedia.org/wiki/Podzia%C5%82_terytorialny_Stan%C3%B3w_Zjednoczonych>

**X5** States\_with\_the\_death\_penalty- zmienna binarana 1 jeśli w danym stanie jest stosowana kara śmierci

Źródło: <https://deathpenaltyinfo.org/states-and-without-death-penalty>

**X6** Number\_of\_deaths- liczba zgonów

Źródło: <https://www.kff.org/other/state-indicator/death-rate-per-100000/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D>

**X7** Personal\_income-roczny dochód na osobę

Źródło: <https://www.statista.com/statistics/303555/us-per-capita-personal-income/>

**X8** Population\_of\_white\_people-procent populacji jaki stanowią biali

**X9** Population\_of\_black\_people- procent populacji jaki stanowią czarni

Źródło: <https://www.kff.org/other/state-indicator/distribution-by-raceethnicity/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D>

# Podstawowe statystyki opisowe

Statystyki opisowe wstępnie wybranego modelu mają się następująco:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Średnia | Mediana | Minimalna | Maksymalna |
| Y | 5655,4 | 5848,1 | 3386,7 | 9256,1 |
| X1 | 5,0100 | 5,0000 | 1,0000 | 12,400 |
| X2 | 5,1010 | 5,9750 | 0,00000 | 7,2500 |
| X3 | 0,080000 | 0,00000 | 0,00000 | 1,0000 |
| X4 | 0,080000 | 0,00000 | 0,00000 | 1,0000 |
| X5 | 0,68000 | 1,0000 | 0,00000 | 1,0000 |
| X6 | 756,52 | 736,15 | 584,90 | 957,10 |
| X7 | 49,791 | 47,910 | 36,636 | 71,823 |
| X8 | 68,960 | 72,000 | 21,000 | 93,000 |
| X9 | 10,220 | 7,0000 | 1,0000 | 38,000 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Percentyl 5% | Percentyl 95% | Zakres Q3-Q1 | Brakujące obs. |
| Y | 3497,8 | 8182,4 | 2060,2 | 0 |
| X1 | 1,5200 | 9,4150 | 4,1250 | 0 |
| X2 | 0,00000 | 7,0000 | 1,8438 | 0 |
| X3 | 0,00000 | 1,0000 | 0,00000 | 0 |
| X4 | 0,00000 | 1,0000 | 0,00000 | 0 |
| X5 | 0,00000 | 1,0000 | 1,0000 | 0 |
| X6 | 621,39 | 939,53 | 102,05 | 0 |
| X7 | 39,212 | 65,930 | 10,228 | 0 |
| X8 | 37,000 | 91,900 | 22,000 | 0 |
| X9 | 1,0000 | 31,450 | 11,250 | 0 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Odch.stand. | Wsp. zmienności | Skośność | Kurtoza |
| Y | 1407,8 | 0,24894 | 0,27236 | -0,47086 |
| X1 | 2,6266 | 0,52428 | 0,53579 | -0,36018 |
| X2 | 1,9809 | 0,38833 | -1,6051 | 1,7960 |
| X3 | 0,27405 | 3,4256 | 3,0963 | 7,5870 |
| X4 | 0,27405 | 3,4256 | 3,0963 | 7,5870 |
| X5 | 0,47121 | 0,69296 | -0,77174 | -1,4044 |
| X6 | 89,701 | 0,11857 | 0,60770 | -0,28445 |
| X7 | 7,8542 | 0,15774 | 0,77017 | 0,14515 |
| X8 | 15,855 | 0,22992 | -0,73379 | 0,33933 |
| X9 | 9,5111 | 0,93063 | 1,1871 | 0,60670 |

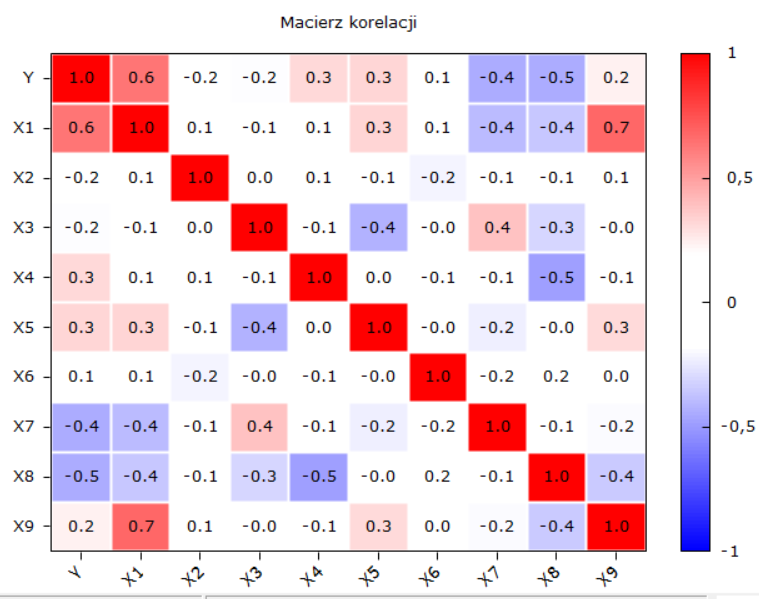
*Tabela 2: Statystyki opisowe*

Średnia zmiennej objaśnianej wynosi 5655,4 współczynnik zmienności jest równy 0,24894 (świadczy o to małej/umiarkowanej zmienności), jej wartości mieszczą się w przedziale (3386,7; 9256,1). Odchylenie standardowe wskazuje ,iż wartości odchylają średnio o 1407,8.

Współczynniki zmienności wszystkich zmiennych objaśniających są wyższe niż 10%, największą wartość przyjmuje dla zmiennej X9(Population\_of\_black\_people) najmniejszy natomiast dla X6(Number\_of\_deaths) należy mieć na uwadze późniejszy jej wpływ lub jego brak na zmienną objaśnianą.

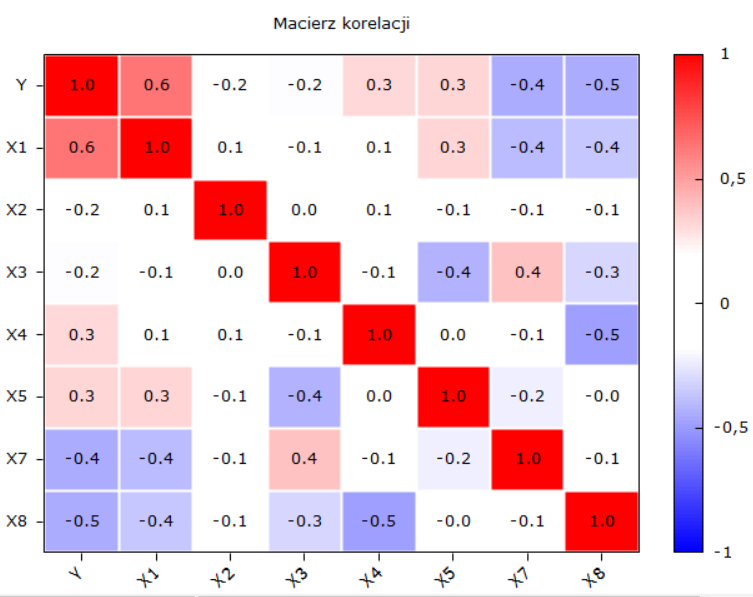
# Model podstawowy

## Korelacja



*Rysunek 1: Macierz korelacji*

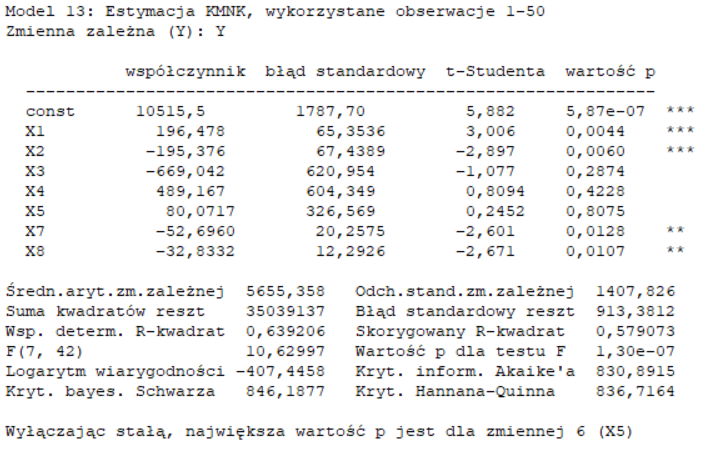
Analizując macierz korelacji można zauważyć iż zmienne objaśniające raczej nie są miedzy sobą skorelowane co świadczy o dobrym dopasowaniu zmiennych do modelu. Można by natomiast oczekiwać większej korelacji miedzy zmienną objaśnianą a zmiennymi objaśniającymi dlatego już na tym etapie zdecydowano o usunięciu zmiennej X6(Number\_of\_deaths) i jej pominięciu w dalszym procesie badawczym. Na tej samej zasadzie zdecydowano o usunięciu zmiennej X9(Population\_of\_black\_people) dodatkowym powodem była wysoka korelacja ze zmienną X1(Homicide\_rate), która ma wysoką korelacje ze zmienną objaśnianą co jest korzystne dla modelu.



*Rysunek 2: Macierz korelacji’*

## Model KMNK

Po analizie korelacji zmiennych zadecydowano o estymacji pierwszego modelu wstępnego. Model został wyestymowany za pomocą metody najmniejszych kwadratów. Jest to model podstawowy zawierający siedem zmiennych objaśniających.

**

*Rysunek 3: Model podstawowy*

Przedstawiony model obrazuje rzeczywistość w 64% procentach (R-kwadrat=0,639) jego skorygowana wartość nie jest z nim rozbieżna, odrzucono zatem stwierdzenie o przeparametryzowaniu modelu.

## Metoda Hellwiga

Wyniki działania tej metody dla analizowanych danych są następujące:



*Rysunek 4:Hellwig*

Metoda wykazała jako najlepsze dla analizowanego modelu są zmienne:

X1(Homicide\_rate), X4(Border\_on\_Mexico), X5(States\_with\_the\_death\_penalty), X7 (Personal\_income), X8(Population\_of\_white\_people)

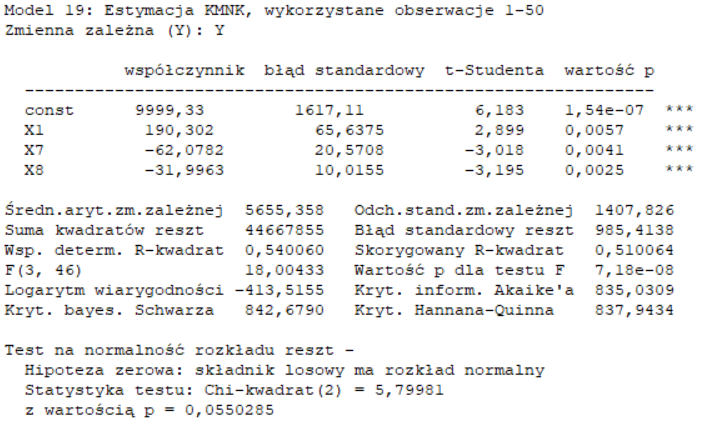
## Metoda krokowo wsteczna

Drugą metodą jaka została zastosowana by wybrać najoptymalniejszy podzbiór zmiennych była metoda krokowo wsteczna. Do badania jak bardzo istotne są zmienne został zastosowany test T-studenta o następujących hipotezach:

***H0: Zmienna jest nie istotna***

***H1: Zmienna jest istotna***

W tej sytuacji pożądanym wynikiem, jest jak najniższa wartość p-value, świadcząca o istotności badanej zmiennej.



*Rysunek 5: Model uzyskany po przeprowadzeniu metody krokowo wstecznej*

Badanie normalności rozkładu reszt wyszło pozytywnie. Zdecydowano o usunięciu 4 zmiennych. Częścią wspólną obu modeli okazały się zmienne X1(Homicide\_rate), X7 (Personal\_income), X8(Population\_of\_white\_people). Wyestymowany zostanie model zawierający właśnie te zmienne gdyż jak sugerują obie z przeprowadzanych metod są one najbardziej istotne dla naszego modelu.

# Weryfikacja statystyczna modelu

## Test istotności zmiennych jako całości(Współczynnik R2)

***H0: Wszystkie współczynniki równe 0***

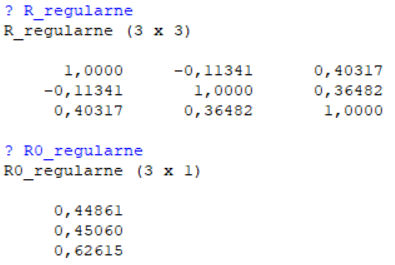
***H1: Przynajmniej jeden współczynnik różny od 0***



*Rysunek 6: Wartość p dla test F*

Wywnioskowano odrzucenie hipotezy zerowej oraz przyjęto istotność współczynnika R-kwadrat

## Efekt Katalizy

Wartość R-kwadrat może być nieprawidłowa jeśli któraś ze zmiennych okazała się być katalitorem.W tym celu przeprowadzono badani na efekt katalizy: *Rysunek 7: R-Regularne*

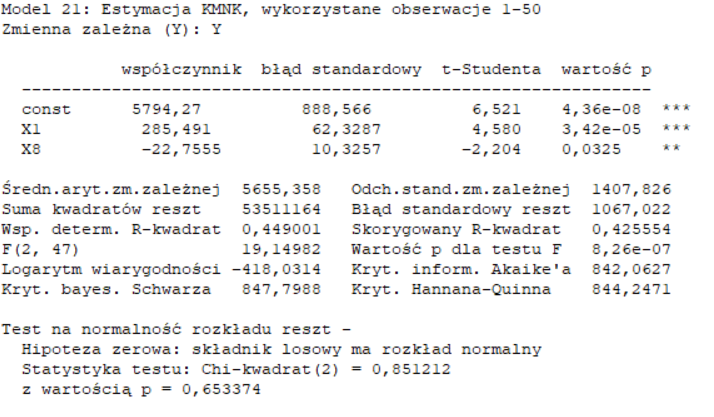
Dla regularnej pary korelacyjnej zmienna jest katliztorem gdy:

𝑟𝑖𝑗 < 0 𝑙𝑢𝑏 𝑟𝑖𝑗 > 𝑟𝑖𝑟𝑗

Wartość -0,11341 jest mniejsza niż 0. Stwierdzono o obecności katalizatora.

Zmienna X7(Personal\_income) okazała się być katalizatorem dlatego też zostanie usunięta z modelu i będzie pominięta w dalszych badaniach.

## Model po usunięciu katlizatora



*Rysunek 8: Model uzyskany po usunięciu katalizatora*

|  |  |  |  |
| --- | --- | --- | --- |
|  | R0 |  | R |
|  | Y | Homicide\_rate | Population\_of\_white\_people |
| Homicide\_rate | 0,4506 | 1 | 0,36482 |
| Population\_of\_white\_people | 0,62615 | 0,36482 | 1 |

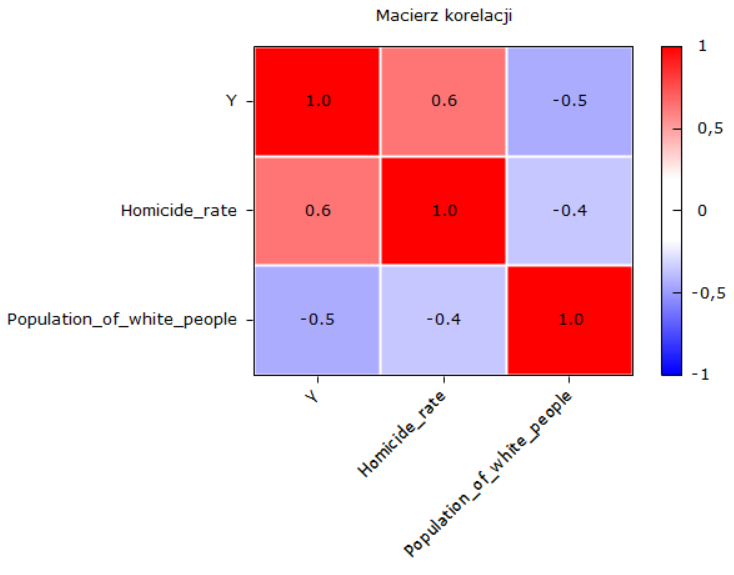
*Tabela 3: Kataliza(Regularna para korelacyjna)*

W badanym przypadku:

0,36482<0 oraz 0,36482>0.7196358

Stwierdzono zatem o braku katalizatora który byłby niekorzystny dla naszego modelu.

## Koincydencja



Model jest koincydentny ponieważ wartości w macierzy korelacji mają taki sam znak jak współczynniki przy tej zmiennej objaśniającej

*Rysunek 9: Macierz korelacji w ostatecznym modelu*

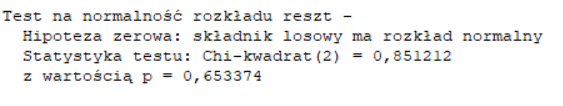
## Normalność rozkładu reszt

Jednym z warunków poprawności modelu ekonometrycznego i stosowania go w praktyce jest rozkład normalny reszt, przy czym istnieje wiele testów statystycznych pozwalających na weryfikacje spełnienia tego warunku.

***H0: Reszty mają rozkład normalny***

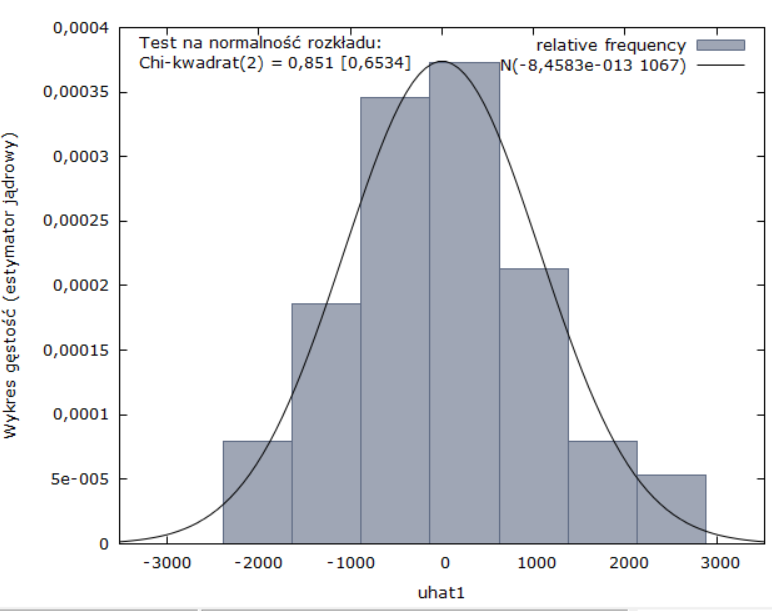
***H1: Reszty nie mają rozkładu normalnego***

Już wcześniej wykazano ,iż reszty mają rozkład normalny jednak przeprowadzono ten test jeszcze raz po usunięciu katlaizatora.



*Rysunek 10: Wynik testu na rozkłady reszt*

Na mocy Testu Doornika-Hansena przyjęto hipotezę zerową, świadczy to o normalnym rozkładzie reszt.

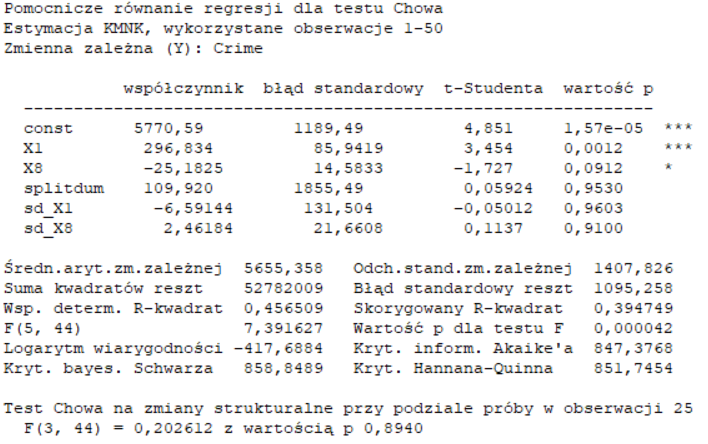
 *Wykres1: Rozkład reszt*

## Stabilność parametrów

Przeprowadzony zostanie test Chowa.

***H0 :parametry są takie same w podpróbach***

***H1 :parametry różnią się w podpróbach***

****** *Rysunek 11: Test Chowa*

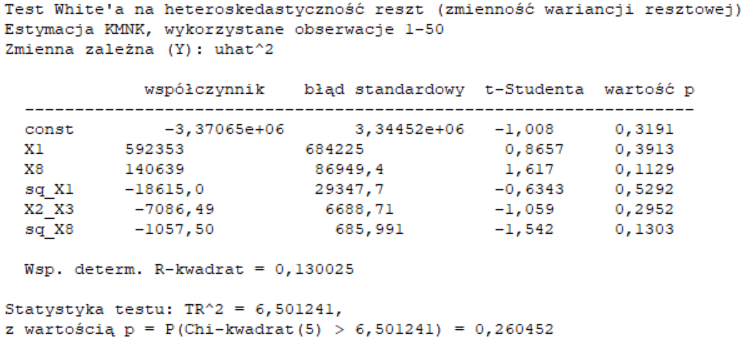
P-value=0,894 jest to wartość większa niż 0,05 przyjmujemy zatem hipotezę zerową, znaczy to że parametry są stabilne.

## Heteroskedastyczność składnika losowego

W celu zbadania heteroskedastyczności przeprowadzono test White’a o następujących hipotezach:

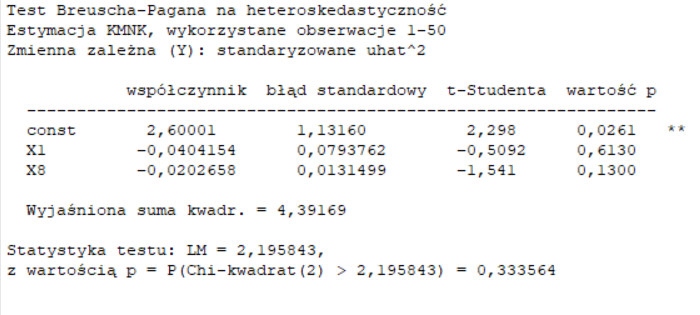
***H0 : homoskedastyczność składnika losowego***

***H1 : heteroskedastyczność składnika losowego***

******

*Rysunek 12: Test White’a*

Stwierdzono ,że test Whita’a nie wykazuję na negatywne zjawiko jakim jest heteroskedastyczność. Zdecydowano o przeprowadzeniu jeszcze jednego testu o takich samych hipotezach.



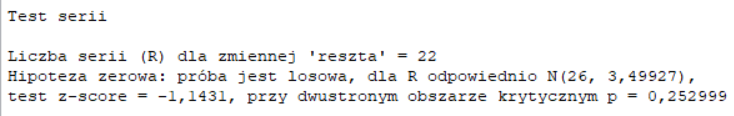
*Rysunek 13: Test Breuscha-Pagana*

Drugi z przeprowadzonych testów także miał pozytywny dla nas wynik. Przyjęto ,iż nie występuje zjawisko heteroskedastyczności.

## Losowość próby

Zostało przeprowadzone badanie losowości próby (dane zostały posortowane od najmniejszych do największych) w tym celu dodano zmienną reszta a także przeprowadzono test liczby serii o następujących hipotezach:

***H0 : próba jest losowa***

***H1 : brak losowości próby***

*Rysunek 12: Test serii*

Wartość p=0,252999 jest to więcej niż założony poziom istotności wywnioskowano więc że badana próba jest losowa.

## Liniowość modelu

Następną rzeczą jaka została sprawdzona była liniowość modelu.

***H0: dobór jednostek do próby jest losowy; model jest liniowy.***

***H1: dobór jednostek do próby nie jest losowy; model jest nieliniowy.***

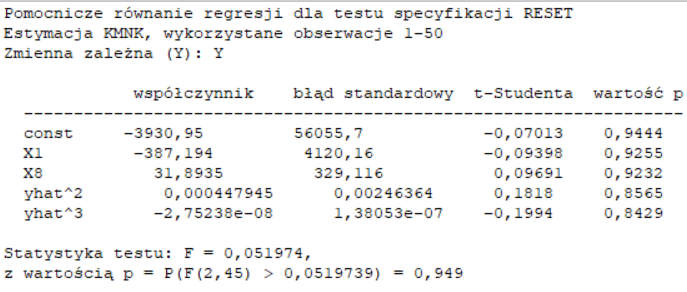
Na mocy testu serii omawianego powyżej stwierdzono o przyjęciu hipotezy zerowej.

## Test RESET-Ramseya

Test poprawności specyfikacji dla modeli regresji liniowej. Innymi słowy stosowany jest w celu sprawdzenia, czy to liniowa postać modelu (względem funkcji kwadratowej, lub sześciennej) jest najlepszym możliwym do wybrania modelem.

***H0: Model jest poprawny funkcyjnie***

***H1: Model nie jest poprawny funkcyjnie***

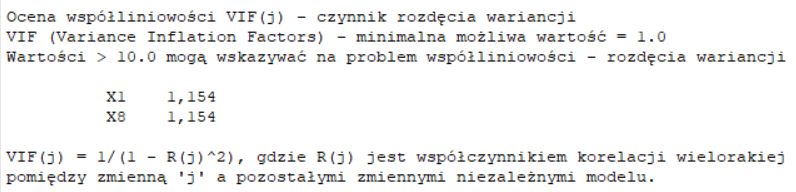
******

*Rysunek 13: Test RESET*

Wartość p dla testu RESET-Ramseya jest znacznie większa niż poziom istotności dlatego przyjęto ,iż model jest poprawny funkcyjnie.

## Współliniowość

Zdecydowano o sprawdzeniu współliniowości za pomocą polecenia „vif” badanie dało następujące rezultaty:

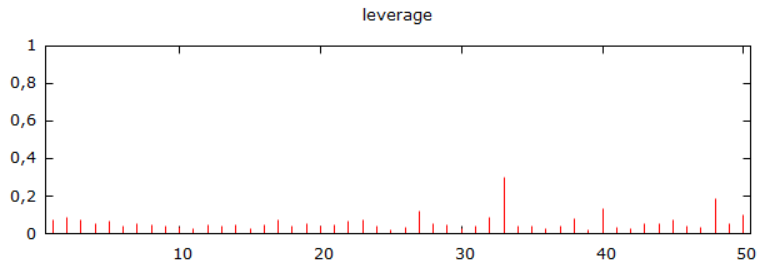


*Rysunek 14: Ocena współliniowości*

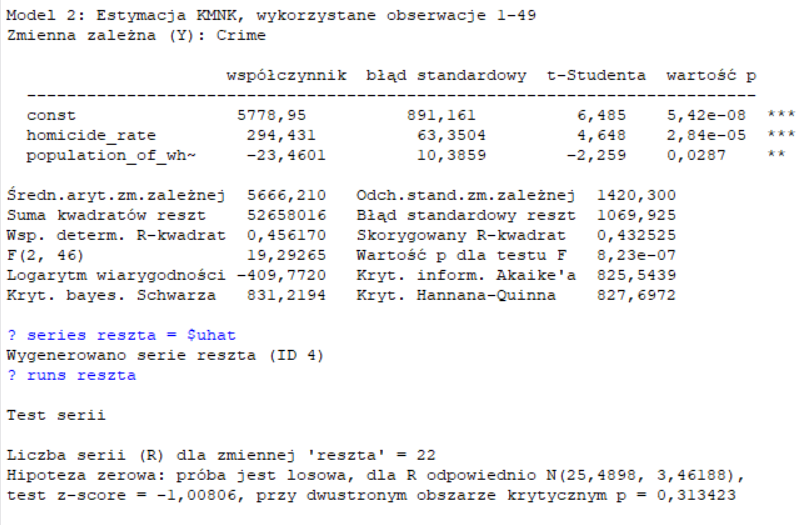
Czynnik rozdęcia wariancji jest w obu przypadkach mniejszy niż 10 na tej podstawie stwierdzono o braku współliniowości pomiędzy zmiennymi.

## Obserwacja odstająca

Relatywnie odległa od pozostałych elementów próby. Innymi słowy, posiadająca nietypową wartość zmiennej niezależnej (objaśniającej) lub nietypowe wartości obydwu zmiennych – zależnej (objaśnianej) i objaśniającej (objaśniających w analizie regresji wielokrotnej). Oznacza to, że związek między Xi a Yi dla danej obserwacji może być inny niż dla reszty obserwacji w zbiorze danych.



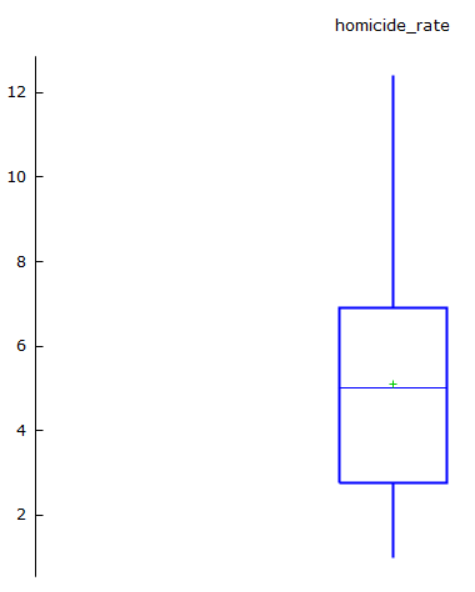
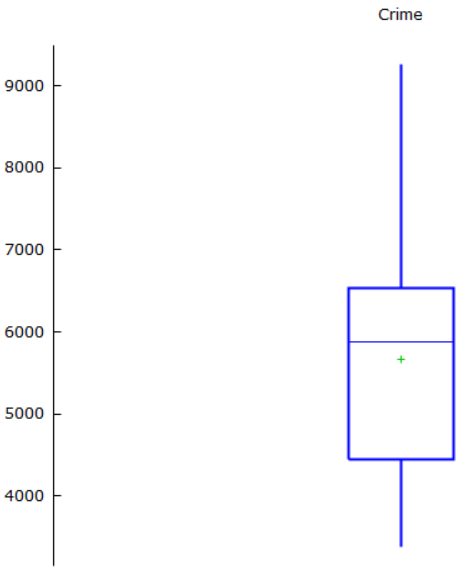
*Wykres 2:Obserwacje odstające*

Największą wartość zaobserwowano dla 33 obserwacji jest to stan North Dakota. Zdecydowano o usunięciu tej obserwacji.

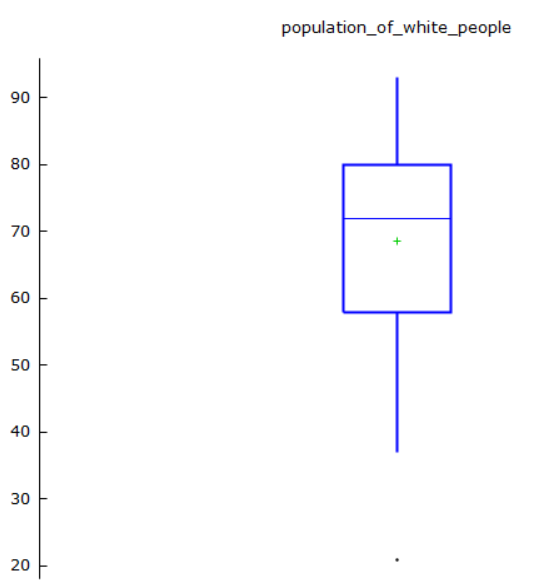
*Rysunek 15: Ols dla 49 obserwacji oraz test serii dla modelu finalnego*

Przeprowadzone jeszcze raz wszystkie testy aby sprawdzić czy model jest tak samo poprawny w tej wersji, uzyskano pozytywne wyniki.

## Wykresy pudełkowe



*Wykres 3:Wykres pudełkowy zmiennej Crime Wykres 4:Wykres pudełkowy zmiennej homicide\_rate*

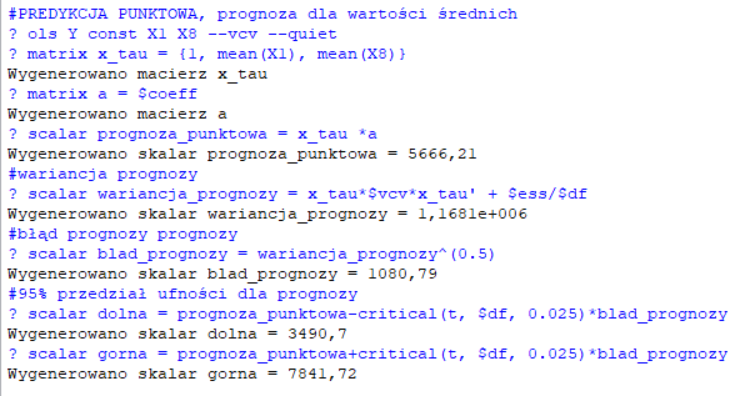


Analizując wykresy pudełkowe każdej ze zminnych wysnuto wniosek o braku obserwacji odstających w zmiennej Y oraz X1 mamy za to jedną obserwacje odstająca przy zmiennej population\_of\_whit\_people. Jest to stan Hawaje najmłodszy ze stanów mieszczący się tylko i wyłącznie na wyspach.

*Wykres 5:Wykres pudełkowy zmiennej population\_of\_white\_people*

# Prognoza

Etapem wieńczącym budowę modelu ekonometrycznego jest praktyczne wykorzystanie modelu, najczęściej w procesie predykcji. *Predykcją ekonometryczną* nazywamy proces wnioskowania w przyszłość na podstawie modelu ekonometrycznego. Efektem predykcji jest oszacowanie nieznanej wartości zmiennej objaśnianej w okresie prognozowanym, zwane często prognozą. W tych badaniach użyto dwóch rodzajów predykcji przedziałową i punktową.

*Rysunek 16: Prognoza*

* Prognoza przedziałowa wykazała iż z prawdopodobieństwem 95% zmienna przyjmie wartości z przedziału (3490,7;7841,72)
* Prognozowanie w puncie została przeprowadzona na zasadzie wyboru średnich zmiennych objaśniających i wykazała że ilość przestępstw będzie kształtować się na poziomie 5666

# Podsumowanie

## Postać liniowa modelu

Po przeprowadzeniu całościowej weryfikacji statystycznej otrzymano model ostateczny

**Crime** *=𝛽1+ 𝛽2∗* Homicide\_rate *+ 𝛽3∗* Population\_of\_white\_people *+ 𝜀*

Gdzie:

*𝛽1 = 5778,95*

*𝛽2 = 294,431*

*𝛽3 = -23,4601*

Model który otrzymano odzwierciedla zmiany ilość przestępstw w 46% ,dowiedziono że zmienne które mają istotny wpływ na ilość przestępstw to zmienne **X1(Homicide\_rate)** i **X8(Population\_of\_white\_people)** dzięki współczynnikom które otrzymano jesteśmy w stanie stwierdzić jaki te zmienne mają wpływ na zmienną objaśnianą **Y(Crime)**

## Wnioski

* Przy założeniu ceteris paribus wywnioskowano że zwieszenie o 1 zmiennej **X1(Homicide\_rate)** sprawi zwiększenie liczby przestępstw o 294
* Drugi wniosek jaki wysnuto to zależność miedzy **X8(Population\_of\_white\_people)** a zmienną objaśnianą zwiększenie tej zmiennej objaśniającej o 1% spowoduje spadek liczby przestępstw o 23
* Rozlegała analiza ekonometryczna spośród 9 zmiennych wyłoniła 2 które mogą wpłynąć na ilość przestępstw popełnianych w każdym ze stanów

# Bibliografia

*1.Uniform Crime Reporting (UCR) Program*

*(*<https://www.fbi.gov/services/cjis/ucr>)

2. *Crime in the United States - statistics & facts* Liisa Jaaskelainen 2017

(<https://www.statista.com/topics/2153/crime-in-the-united-states/>)

3.*Crime in the United States 2019* - Bernan Press

4.*Ekonometria*, Sobczyk Mieczysław Warszawa 2013

5.*Ameryka Kraina więzień*-Stanisław Krawczyk 7.05.2018

( <http://magazynkontakt.pl/ameryka-kraina-wiezien.html>)