KADD - projekt

Analiza powodów rozwodów w Polsce

Kacper Skelnik

16 czerwca 2020

1 Wstęp

Projekt miał sprawdzić czy istnieje zależność pomiędzy powodami rozwodów w Polsce na przestrzeni lat 2003-2019. Powody które udało się sklasyfikować (w bazie danych) to:

- niedochowanie wierności małżeńskiej
- naganny stosunek do członków rodziny
- nadużywanie alkoholu
- trudności mieszkaniowe
- nieporozumienia na tle finansowym
- niezgodność charakterów
- niedobór seksualny
- dłuższa nieobecność
- różnice światopoglądowe
- narkotyki
- hazard

oraz ”inne” których nie brałem pod uwagę. Z racji na dużą ilość różnych rozpatrywanych powodów zdecydowałem się wyłonić grupy które zachowują się podobnie.

2 Dane

Dane których użyłem pochodzą z bazy danych Głównego urzędu statystycznego pod linkiem. Znajdujące się tam dane zawierają liczbę rozwodów których powodem jest wyłącznie jeden z wylistowanych powyżej. Zapewni to klarowność wyników i brak ”szumu” wynikającego z mieszania się grup. Dane można pobrać w wygodnych formatach do obróbki, które zapewnią minimum pracy przed właściwą analizą. Pobrane dane przekopiowałem do plików tekstowych tak, że otrzymałem 11 różnych plików (jeden dla każdego powodu) z kolumnami:

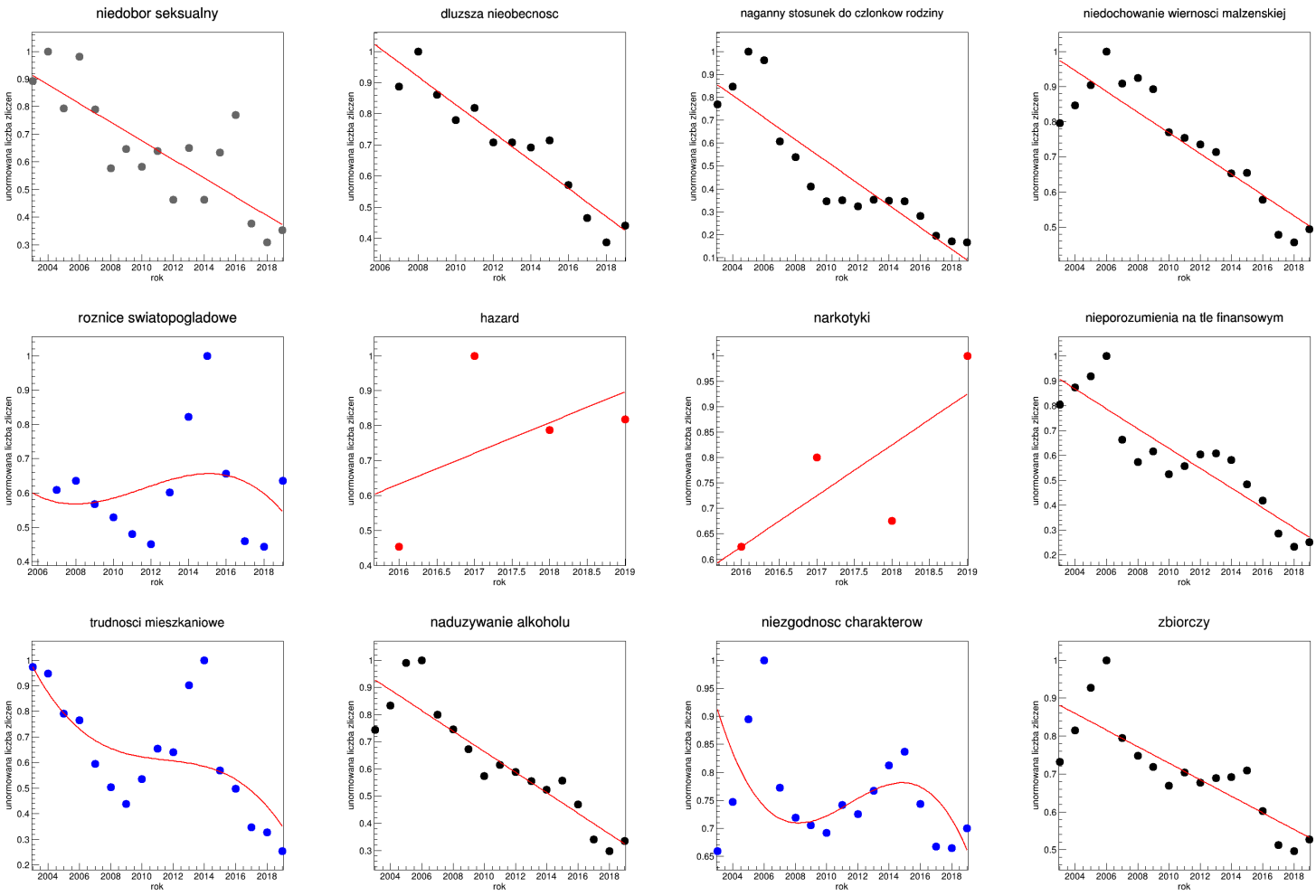
- rok
- liczba zliczeń rozwodów

Tak przygotowane dane można już łatwo wczytać używając języka C++. Przykładowy zestaw danych:

Rok	**Liczba zliczeń**		
2003	6230	2011	5906
2004	6627	2012	5754
2005	7085	2013	5589
2006	7834	2014	5115
2007	7113	2015	5124
2008	7239	2016	4528
2009	6996	2017	3750
2010	6030	2018	3581
2011	5906	2019	3864

Tablica 1: Liczba rozwodów w latach 2003-2019 z powodu niedochowania wierności małżeńskiej

3 Wyniki



Rysunek 1: Wykres zbiorczy wszystkich unormowanych rozkładów wraz z dopasowanymi krzywymi

Powyżej znajdują się wszystkie rozpatrywane rozkłady (unormowane) wraz z dopasowaną krzywą. Punkty każdej grupy są narysowane innym kolorem. Rozkład zbiorczy został stworzony poprzez dodanie do siebie zliczeń z wszystkich rozkładów w danym roku.

4 Analiza wyników

Do każdego rozkładu dopasowaną funkcję za pomocą metody Fit wchodzącą w skład środowiska ROOT. Parametry funkcji prostej oraz wartość testu dopasowania do funkcji dopasowanej χ^2 i R^2 znajdują się poniżej.

Rozkład	**p0**	**p1**	**p2**	**p3**	χ^2	R^2
niedobór seksualny	68.56 (12.52)	-0.034 (6)	-	-	0.24	0.66
dłuższa nieobecność	90.94 (9.0)	-0.045 (5)	-	-	0.04	0.90
naganny stosunek do członków rodziny	96.71 (12.42)	-0.048 (6)	-	-	0.23	0.80
niedochowanie wierności małżeńskiej	60.01 (7.8)	-0.029 (4)	-	-	0.092	0.79
różnice światopoglądowe	$4 \cdot 10^6$	-6014.43	2.99	-0.0005	0.26	0.05
hazard	-176.53 (217.62)	0.088 (100)	-	-	0.12	0.25
narkotyki	-200.98 (117.2)	0.1 (1)	-	-	0.034	0.60
nieporozumienia na tle finansowym	80.48 (9.72)	-0.040 (5)	-	-	0.14	0.82
trudności mieszkaniowe	$4 \cdot 10^6$	-6014.43	2.99	-0.0005	0.39	0.43
nadużywanie alkoholu	76.56 (8.75)	-0.038 (4)	-	-	0.12	0.83
niezgodność charakterów	$4 \cdot 10^6$	-6014.43	2.99	-0.0005	0.18	0.40
rozkład zbiorczy	44.57 (7.71)	-0.022 (4)	-	-	0.09	0.68

Tablica 2: Parametry funkcji dopasowanych do danych

W przypadku braku podanej niepewności parametru oznacza to, że ta była zbyt mała żeby zmieścić się w przedziale dwóch znaczących cyfr po przecinku. Może zastanawiać to, że wszystkie krzywe trzeciego stopnia mają te same współczynniki mimo różnych kształtów. Tu również różnice nie zmieściły się w przedziale liczb znaczących lub po prostu w tabeli. Z racji tego poniżej znajdują się ich dokładniej podane wartości:

różnice światopoglądowe:

- **p0** = 4.0327 $\cdot 10^6$ (0.649467)
- **p1** = -6014.43 (0.000323006)
- **p2** = 2.99 (1.60459 $\cdot 10^-7$)
- **p3** = -0.000495477 (7.96186 $\cdot 10^-11$)

trudności mieszkaniowe:

- **p0** = 4.03272 $\cdot 10^6$ (0.577021)
- **p1** = -6014.43 (0.000287496)
- **p2** = 2.98999 (1.4296 $\cdot 10^-7$)
- **p3** = -0.000495479 (7.09473 $\cdot 10^-11$)

niezgodność charakterów:

- **p0** = 4.0327 $\cdot 10^6$ (0.385018)
- **p1** = -6014.43 (0.000191832)
- **p2** = 2.99 (9.53899 $\cdot 10^-8$)
- **p3** = -0.000495478 (4.73397 $\cdot 10^-11$)

Oprócz tego zostały wykonane również testy R^2 dopasowania rozkładów do rozkładu zbiorczego. W tym przypadku wyniki są przedstawione w dwóch wariantach: przy sprawdzeniu dopasowania do punktów rozkładu zbiorczego i przy testowaniu dopasowania do funkcji dopasowanej do rozkładu zbiorczego.

Rozkład	R^2 **do punktów**	R^2 **do funkcji**
niedobór seksualny	0.50	0.37
dłuższa nieobecność	0.30	0.25
naganny stosunek do członków rodziny	1.05	0.97
niedochowanie wierności małżeńskiej	0.68	0.47
różnice światopoglądowe	0.45	0.42
hazard	1.41	1.05
narkotyki	2.83	2.14
nieporozumienia na tle finansowym	0.66	0.55
trudności mieszkaniowe	0.44	0.33
nadużywanie alkoholu	0.57	0.44
niezgodność charakterów	2.57	1.85

Tablica 3: Wyniki testów R^2 dopasowania do rozkładu zbiorczego

Warto zwrócić na uwagę, że niektóre wartości nie mieszczą się w przedziale $0 \leq R^2 \leq 1$. W tym przypadku nie jest to błąd ponieważ wartość tesu musi zawierać się w tym przedziale dla szczególnego przypadku, gdy dopasowana jest prosta $y = ax + b$ metodą najmniejszych kwadratów. W innym przypadku nie ma większej przesłanki, żeby wartość testu musiała znajdować się zawsze w tym przedziale. Ciągle jednak $R^2 = 1$ jest wynikiem idealnym.

Następnym krokiem było wykonanie testu χ^2 pomiędzy każdym rozkładem, tak aby móc stwierdzić które z rozkładów są ze sobą zgodne. Wyniki testu są przedstawione w formie macierzy:

	niedobór seksualny	dłuższa nieobecność	naganny stosunek do członków rodziny	niedochowanie wierności małżeńskiej	różnice światopoglądowe	hazard	narkotyki	nieporozumienia na tle finansowym	trudności mieszkaniowe	nadużywanie alkoholu	niezgodność charakterów
niedobór seksualny	**1**	0	0	0	0	0	0	**1**	0	**1**	0
dłuższa nieobecność	0	**1**	0	**1**	0	0	0	0	0	0	0
naganny stosunek do członków rodziny	0	0	**1**	0	0	0	0	0	0	0	0
niedochowanie wierności małżeńskiej	0	0	0	**1**	0	0	0	0	0	0	0
różnice światopoglądowe	0	0	0	0	**1**	0	0	0	0	0	0
hazard	0	0	0	0	0	**1**	**1**	0	0	0	0
narkotyki	0	0	0	0	0	**1**	**1**	0	0	0	0
nieporozumienia na tle finansowym	**1**	0	**1**	0	0	0	0	**1**	0	**1**	0
trudności mieszkaniowe	0	0	0	0	0	0	0	0	**1**	0	0
nadużywanie alkoholu	**1**	0	0	0	0	0	0	**1**	0	**1**	0
niezgodność charakterów	0	0	0	0	0	0	0	0	0	0	**1**

Tablica 4: Macierz testów χ^2

Poziom istotności testu został ustalony na:

$$\alpha = 0.01$$

Jak widać taka metoda ma pewną wadę. Zgodnie ze wzorem na statystykę χ^2

$$\chi^2 = \sum_{i=o}^{n} \left(\frac{O_i - E_i}{\sigma_i} \right)^2$$

gdzie O_i jest to wartość mierzona, E_i wartość oczekiwana a σ_i odchylenie standardowe wartości mierzonej. Fakt, że macierz w Tablicy 4 nie jest dokładnie symetryczna wynika właśnie z tego, że w zależności jaki rozkład przyjmiemy za mierzony otrzymamy różne odchylenia standardowe. Ten problem rozwiązuje użycie testu Kołmogorowa-Smirnowa, który idealnie sprawdza się do porównywania dwóch rozkładów. Poniżej znajduję się analogiczna macierz dla wyników testu Kołmogorowa-Smirnowa.

	niedobór seksualny	dłuższa nieobecność	naganny stosunek do członków rodziny	niedochowanie wierności małżeńskiej	różnice światopoglądowe	hazard	narkotyki	nieporozumienia na tle finansowym	trudności mieszkaniowe	nadużywanie alkoholu	niezgodność charakterów
niedobór seksualny	1	0	0	**1**	0	0	0	1	0	1	0
dłuższa nieobecność	0	1	**1**	1	0	0	0	**1**	0	**1**	0
naganny stosunek do członków rodziny	0	**1**	1	0	0	0	0	0	0	0	0
niedochowanie wierności małżeńskiej	**1**	**1**	0	1	0	0	0	0	0	**1**	**1**
różnice światopoglądowe	0	0	0	0	1	0	0	0	0	0	0
hazard	0	0	0	0	0	1	1	0	0	0	0
narkotyki	0	0	0	0	0	1	1	0	0	0	0
nieporozumienia na tle finansowym	1	**1**	0	0	0	0	0	1	0	1	0
trudności mieszkaniowe	0	0	0	0	0	0	0	0	1	0	0
nadużywanie alkoholu	1	**1**	0	**1**	0	0	0	1	0	1	0
niezgodność charakterów	0	0	0	**1**	0	0	0	0	0	0	1

Tablica 5: Macierz testów Kołmogorowa-Smirnowa

Jak widać w tym przypadku macierz jest w pełni symetryczna. Zielonym kolorem zaznaczone są punkty które zgadzają się z wynikiem testu χ^2, czerwonym te które się nie zgadzają, czarnym nowe punkty. W testach przyjęto:

$$\lambda = 0.5$$

Czyli test był uznany za zdany gdy jego wartość < 0.5.

Na podstawie powyższych wyników można było podjąć próbę pogrupowania rozkładów. Na Rysunku 1 widać trzy różne grupy:

- Czerwona, która zawiera rozkłady, trudne do sklasyfikowania ze względu na małą ilość danych, ale które są ze sobą zgodne (zarówno w teście χ^2 i Kołmogorowa-Smirnowa)
- Niebieska, która zawiera rozkłady które trudno było przypisać do reszty ze względu na ich chaotyczny przebieg (wymagały dopasowania wielomianu trzeciego stopnia). Mimo bardzo podobnych dopasowanych wielomianów, żaden test nie potwierdził ich zgodności.
- Czarna, która zawiera rozkłady które można było uznać za quazi-liniowe a które nieźle zgadzają się z rozkładem zbiorczym (biorąc pod uwagę test R^2 gdy > 0.5). Oprócz tego każdy z nich jest zgodny z którymś z pozostałych rozważając wyniki testów Kołmogorowa-Smirnowa.

Oprócz tego jest jeszcze zaznaczony na szaro rozkład "Niedobór seksualny" który uzyskał najgorszy wynik (biorąc pod uwagę rozkłady czarne) z testu R^2, ale który wyglądem przypomina pozostałe rozkłady (i jest do nich zgodny ze względu na test χ^2 i Kołmogorowa-Smirnowa).

5 Wnioski

Do pełnego wyciągnięcia wniosków należałoby mieć kompetencję z dziedziny socjologi, psychologi i historii współczesnej. Nawet wówczas gdybym takowe posiadał, analiza byłby żmudna i bardzo subiektywna. W związku z tym zrezygnuję z próby szukania głębszego dna w przedstawionych wynikach a skupię się na stronie statystyczno-programistycznej.

Po pierwsze pierwotny pomysł na zastosowanie testu χ^2 do zbadania zgodności rozkładów okazał się nietrafiony. Test ten dobrze sprawdza się gdy mamy rozkład teoretyczny który porównujemy z danymi, ale w przypadku dwóch rozkładów empirycznych powstaje problem z odchyleniem standardowym opisany powyżej. O wiele lepiej sprawdza się test Kołmogorowa-Smirnowa, który nawet w dość rygorystycznej wersji pokazał wiele rozkładów które są ze sobą zgodne.

Kolejną sprawą jest duża ilość rozkładów które należy opracować. Okazało się to dość kłopotliwe i pracochłonne. Szczególnie trudna stała się analiza Tablicy 5 która zawiera wyniki testów Kołmogorowa-Smirnowa. Problem stał się na tyle wielowymiarowy, że wyznaczenie grup bazując tylko na tym okazało się niemożliwe. Dlatego pomocne okazało się policzenie na początku testu R^2 zgodności z rozkładem zbiorczym. W tym miejscu warto zwrócić uwagę na interesujący fakt, że wartości testu przy porównaniu z krzywą dopasowaną do rozkładu zbiorczego są zawsze niższe niż przy porównaniu do punktów rozkładu.

Ostatecznie można stwierdzić, że omawiany zbiór da się pogrupować i jest o wiele bardziej jednorodny niż można by przypuszczać.