Tematy projektu: Determinanty zaufania wśród brytyjczyków - regresja logistyczna

Autor: Sebastian Boruch

Spis treści

[1. Cel badania i opis zbioru 3](#_Toc504939131)

[2. Statystyki opisowe zmiennych 5](#_Toc504939132)

[3. Model uporządkowanej regresji logistycznej 13](#_Toc504939133)

[3.1. Wyniki modelu uporządkowanej regresji logistycznej 14](#_Toc504939134)

[3.2. Ocena jakości modelu regresji uporządkowanej 20](#_Toc504939135)

[4. Model regresji binarnej typu forward 22](#_Toc504939136)

[4.1 Wyniki modelu regresji logistycznej binarnej typu forward 22](#_Toc504939137)

[4.2. Ocena jakości modelu regresji binarnej typu forward 25](#_Toc504939138)

[4.3 Ocena jakości modelu regresji binarnej typu backward 27](#_Toc504939139)

[5. Bibliografia 28](#_Toc504939140)

[6. Spis tabel i rysunków 28](#_Toc504939141)

[7. Kody SAS 29](#_Toc504939142)

# 1. Cel badania i opis zbioru

Prezentowany projekt stanowi analizę zaufania wśród brytyjskiego społeczeństwa. W tym projekcie do analizy badanego problemu wykorzystano regresję uporządkowaną i binarną (forward i backward).

Celem przeprowadzonej analizy było zidentyfikowanie w jakim stopniu czynniki społeczno - demograficzne wpływają na osobiste odczucie dotyczące tego, czy ludzie chcą badanego wykorzystać, czy też są wobec niego uczciwi. Analiza umożliwi również zbadanie jak bardzo brytyjskie społeczeństwo ufa sobie nawzajem. W budowie modeli regresji logistycznej wykorzystano dane pochodzące z European Social Survey. Jest to platforma gromadząca dane dotyczące wzorców zachowań, nastawień oraz ocen kluczowych sfer życia społecznego wielu krajów Europy.

W projekcie wykorzystano dane z badania przeprowadzonego w Wielkiej Brytanii w 2016 roku. Oryginalny zbiór zebrany na podstawie ankiet przeprowadzanych osobiście liczy 1959 obserwacji i opisuje odpowiedzi na 499 pytań. Autor do przeprowadzenia tego raportu wybrał 11 zmiennych. Te pytania wraz z nazwami zmiennych do nich przypisanych znajdują się w tabeli poniżej:

*Tabela 1. Charakterystyki zmiennych*

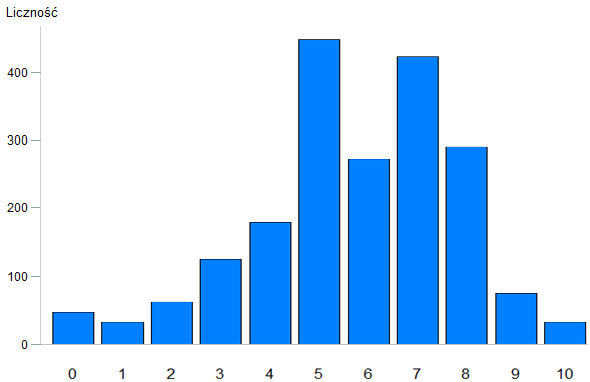
|  |  |  |  |
| --- | --- | --- | --- |
| Pytanie | Nazwa  zmiennej | Typ | Wartości |
| Większość ludzi  chce cię wykorzystać  czy są wobec ciebie  szczerzy? | PPLFAIR | Kategoryczna  porządkowa –  zmienna celu | 1 – nieufny  2 – średnioufny  3 – ufny |
| Ile masz lat? | AGEA | Numeryczna | 15-94 |
| Ile ukończyłaś/łeś lat  edukacji? | EDUYRS | Numeryczna | 0-54 |
| W podanej skali, jak  bardzo zgadzasz się ze stwierdzeniem, że  integracja europejska  powinna się pogłębiać? | EUFTF | Nominalna  (liczby całkowite od 0 do 10) | 0 – integracja  postąpiła za daleko  10 – należy  pogłębiać integrację |
| W podanej skali, jak  bardzo jesteś  szczęśliwa/wy? | HAPPY | Nominalna  (liczby całkowite  od 0 do 10) | 0 – bardzo  nieszczęśliwa/wy  10 – bardzo  szczęśliwa/wy |
| Jak oceniasz swoje zdrowie? | HEALTH | Nominalna (liczby całkowite od 1 do 5) | 1 – bardzo dobrze  5 – bardzo źle |
| W podanej skali, jak  ważne jest dla ciebie  odnoszenie sukcesów i  bycie docenionym? | IPSUCES | Nominalna  (liczby całkowite  od 1 do 6) | 1 – jest dla mnie  bardzo ważne  6 – nie jest dla mnie  ważne |
| W podanej skali, jak  bardzo jesteś  zadowolona/ny z  funkcjonowania  demokracji w Twoim  kraju? | STFDEM | Nominalna  (liczby całkowite  od 0 do 10) | 0 – bardzo  niezadowolona/ny  10 – bardzo  zadowolona/ny |
| Jaki jest ogólny przychód  netto gospodarstwa  domowego? | HINCTNTA | Nominalna (liczby całkowite od 1 do 10 oznaczające decyle w społeczeństwie) | 1 – pierwszy decyl  10 – dziesiąty decyl |
| Jak bardzo ufasz  parlamentowi w twoim  kraju? | TRSTPRL | Nominalna  (liczby całkowite  od 0 do 10) | 0 – nie ufam  10 – ufam całkowicie |
| Ilu masz podwładnych w pracy? | NJBSPV | Numeryczna | 0-50 |

*Źródło: Opracowanie własne na podstawie danych z European Social Survey 2016*

# 2. Statystyki opisowe zmiennych

Analizowana zmienna celu PPLFAIR wyjściowo miała 10 kategorii - od 0 do 9, gdzie 0 określało największy poziom nieufności i poczucia wykorzystania przez społeczeństwo. Poniżej na Wykresie 1, zilustrowano początkowy rozkład zmiennej. Najpopularniejszą kategorią wśród ankietowanych z UK okazał się średni poziom równy 6.

*Rysunek 1. Rozkład zmiennej PPLFAIR - 11 poziomów*



pplfair

*Źródło: opracowanie własne*

Ten rozkład zmiennej okazuje się być trudny do zinterpretowania poprzez regresję logistyczną. Sam zbiór cechował się również dużą ilością braków danych ( ponad 9% na całym zbiorze). Dlatego autor raportu zdecydował, aby brakujące dane lub odpowiedzi oznaczone jako ,,brak odpowiedzi”, ,,nie dotyczy”, ,,nie wiem” lub ,,odmowa odpowiedzi” zastąpić danymi wygenerowanymi przez metodę średniej ruchomej (PROC MI w SAS). Metoda średniej ruchomej jest metodą imputacji dostępną m.in. dla zmiennych ciągłych. Jest ona podobna do metody regresji, z tą różnicą, że dla każdej brakującej wartości przypisuje wartość losowo z zestawu obserwowanych wartości, których przewidywane wartości są najbliższe przewidywanej wartości dla brakującej wartości z symulowanego modelu regresji. Metoda średniej ruchomej zapewnia, że wartości imputowane są wiarygodne i mogą być bardziej odpowiednie niż metoda regresji, jeśli założenie normalności zostanie naruszone. Zastosowanie tej metody umożliwiło użycie wszystkich obserwacji do każdego z modeli regresji.

Na potrzeby omawianej regresji uporządkowanej skategoryzowano zmienną PPLFAIR na 3 kategorie:

1 – nieufny

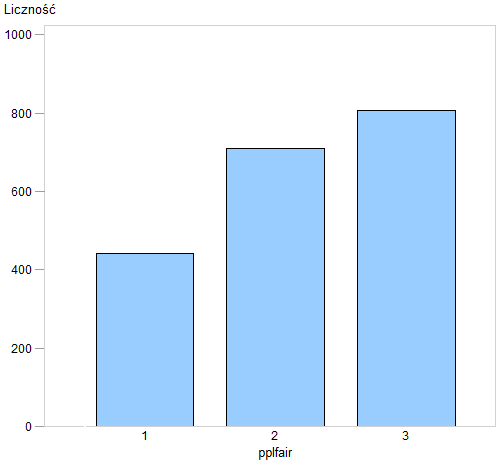
2 – średnioufny

3 – ufny.

W pierwszej kategorii znalazły się osoby, które oceniając uczciwość polskiego społeczeństwa wybrały ze skali od 0 do 4. Kolejna kategoria skala 5 i 6, zaś ostatnia zawierała odpowiedzi ze skali od 7 do 10.

Zmienna objaśniana jest zmienną porządkową rosnącą, poziom 3 określa najwyższy poziom ufności z zaprezentowanych. Na Rysunku 2 przedstawiono histogram zmiennej PPLFAIR po modyfikacji.

*Rysunek 2. Rozkład zmiennej PPLFAIR - 3 poziomy*



*Źródło: opracowanie własne*

Histogram nie prezentuje zbilansowanej liczby dla każdej kategorii. Jednak, dzięki zastąpieniu braków danych przewidzianymi wartościami, liczność każdej z kategorii przekracza 400, co umożliwia przeprowadzenie wiarygodnej analizy zbioru danych. W celu dokładnego przeanalizowania zmiennej celu, przygotowano tabelę z liczebnością oraz procentem występowania każdej z kategorii.

*Tabela 2. Rozkład zmiennej PPLFAIR*

| **pplfair** | **Liczebność** | **Procent** | **Liczebność skumulowana** | **Procent skumulowany** |
| --- | --- | --- | --- | --- |
| 1 | 441 | 22.51 | 441 | 22.51 |
| 2 | 710 | 36.24 | 1151 | 58.75 |
| 3 | 808 | 41.25 | 1959 | 100.00 |

*Źródło: opracowanie własne*

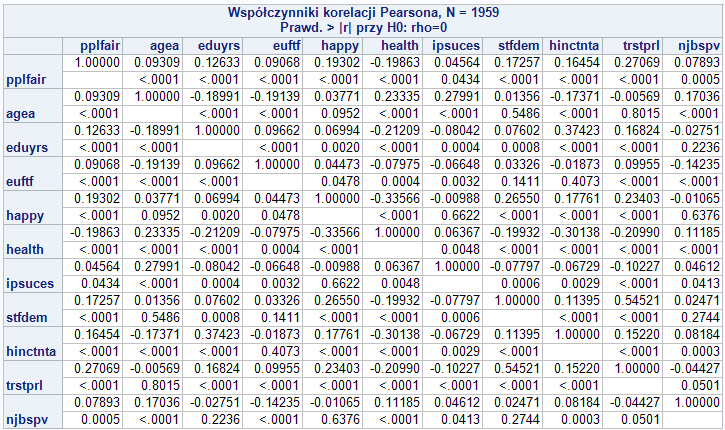
Kategoria 1 stanowi 22,5% odpowiedzi, kategoria 2 stanowi 36% a kategoria 3 około 42%.

Statystyki zmiennych objaśniających

W dalszych rozdziałach raportu ukazane będzie, iż większość ze zmiennych jest istotna statystycznie. Jednym z założeń analizy regresji jest brak występowania współliniowości zmiennych objaśniających). Wprowadzając do modelu regresji silnie skorelowane ze sobą zmienne wprowadzamy do modelu (przy każdej zmiennej) małą bądź zerową unikalną "część wyjaśnienia" zmiennej zależnej. W zależności od sposobu liczenia jeden z predyktorów silnie powiązanych straci swoją "moc" przewidywania.

Współczynnik korelacji Pearsona służy do sprawdzenia czy dwie zmienne ilościowe są powiązane ze sobą związkiem liniowym. Wynik Pearsona może wahać się od -1 do 1. Wartości skrajne czyli -1 i 1 oznaczają idealną, totalną korelację między zmienną A i zmienną B. Wynik równy ‚zero’ oznacza brak współwystępowania wartości tych dwóch zmiennych w naturze (brak korelacji).

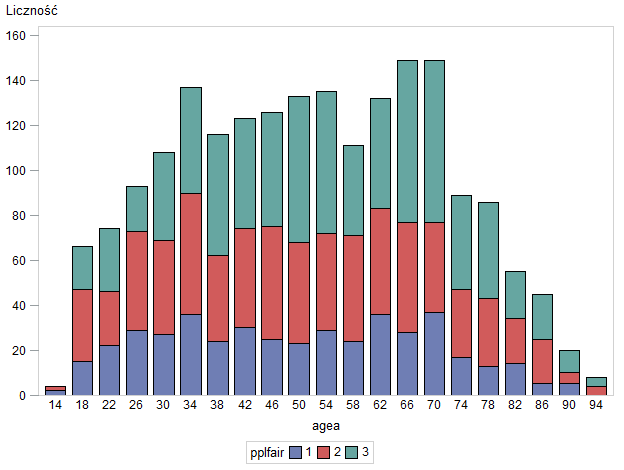
*Tabela 3. Współczynniki korelacji Pearsona*



*Źródło: opracowanie własne*

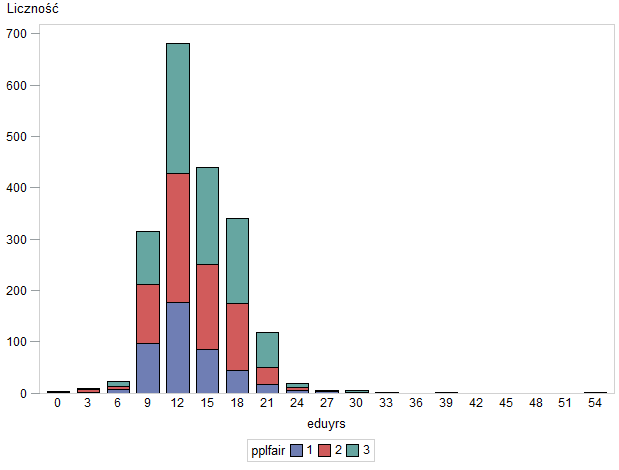
Jak widać po wynikach tabeli, prawie wszystkie zmienne są skorelowane ze sobą w stopniu słabym bądź bardzo słabym, co pozwala na przeprowadzenie analizy regresji. Korelacją silną jest tylko ta pomiędzy stfdem a trstprl, ale zmienna stfdem nie została uwzględniona w wynikach analizy.

*Rysunek 3.* *Rozkład zmiennej AGEA z uwzględnieniem PPLFAIR*



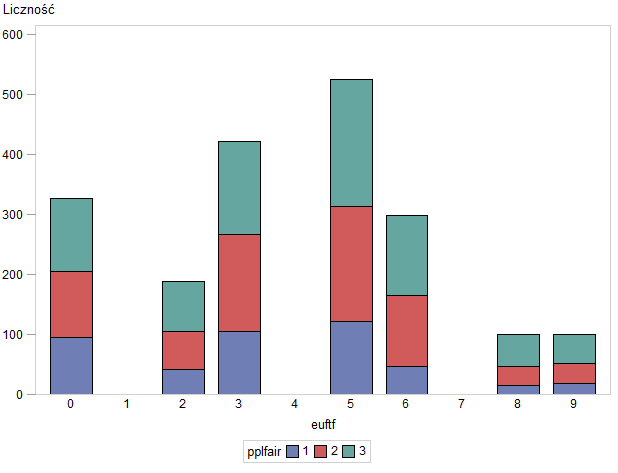
*Źródło: opracowanie własne*

*Rysunek 4. Rozkład zmiennej EDUYRS z uwzględnieniem PPLFAIR*



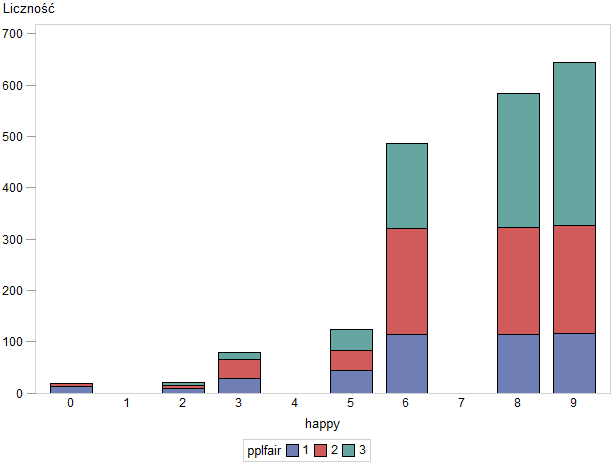
*Źródło: opracowanie własne*

*Rysunek 5. Rozkład zmiennej EUFTF z uwzględnieniem PPLFAIR*



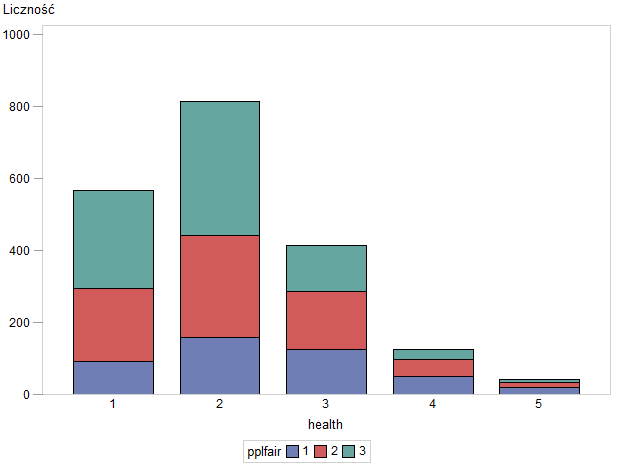
*Źródło: opracowanie własne*

*Rysunek 6. Rozkład zmiennej HAPPY z uwzględnieniem PPLFAIR*



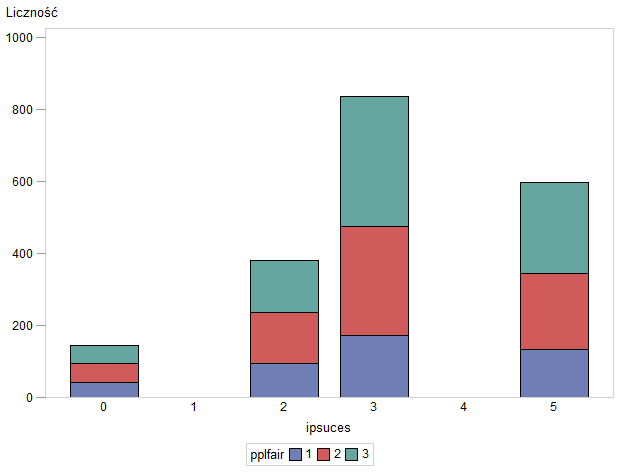
*Źródło: opracowanie własne*

*Rysunek 7. Rozkład zmiennej HEALTH z uwzględnieniem PPLFAIR*



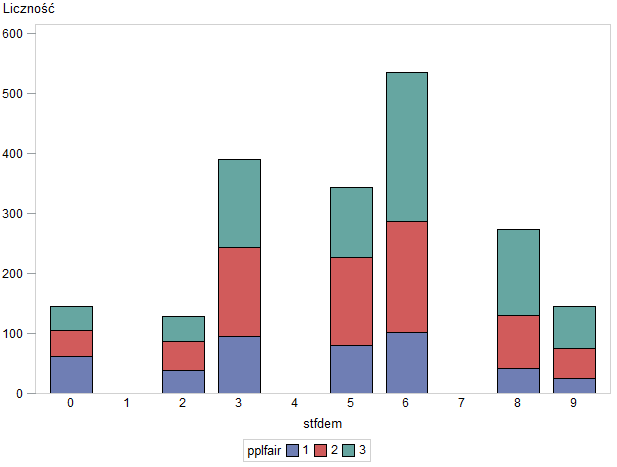
*Źródło: opracowanie własne*

*Rysunek 8. Rozkład zmiennej IPSUCES z uwzględnieniem PPLFAIR*



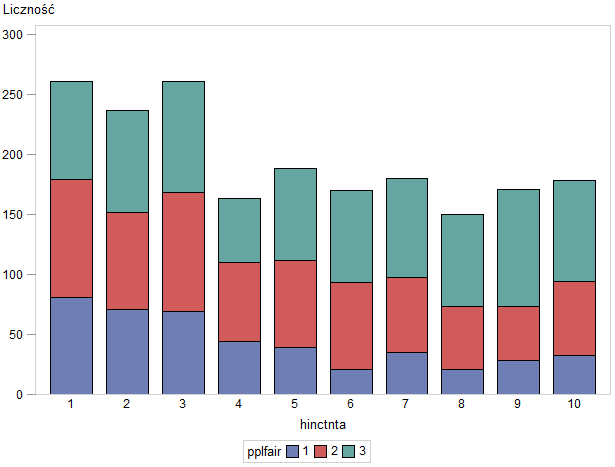
*Źródło: opracowanie własne*

*Rysunek 9. Rozkład zmiennej STFDEM z uwzględnieniem PPLFAIR*



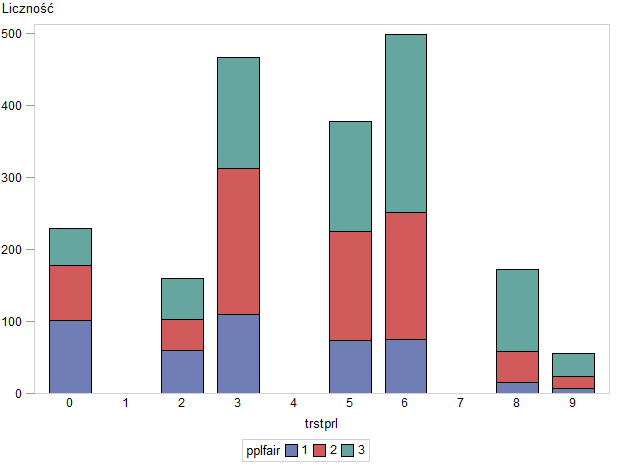
*Źródło: opracowanie własne*

*Rysunek 10. Rozkład zmiennej HINCTNTA z uwzględnieniem PPLFAIR*



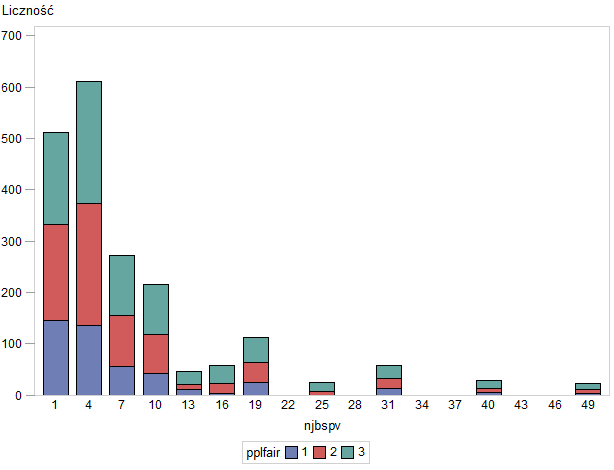
*Źródło: opracowanie własne*

*Rysunek 11. Rozkład zmiennej TRSTPRL z uwzględnieniem PPLFAIR*



*Źródło: opracowanie własne*

*Rysunek 12. Rozkład zmiennej NJBSPV*



*Źródło: opracowanie własne*

# 3. Model uporządkowanej regresji logistycznej

W przedstawionym projekcie zaprezentowano model uporządkowanej regresji logistycznej, gdzie zmienna objaśniania jest zmienną jakościową Y przyjmującą 3 kategorie, których wartość uzależniona jest od poziomu zmiennych niezależnych X1,X2,...,Xk (jakościowych bądź ilościowych). Podstawowym założeniem modelu uporządkowanej regresji logistycznej jest założenie o proporcjonalności odds. Przy spełnionym założeniu proporcjonalności odds, iloraz szans (odds ratio) mierzący efekt wpływu wystawienia na ryzyko dla szacowanych wariantów porównań będzie równy, niezależnie od tego, w jaki sposób zostaną pogrupowane kategorie uporządkowania. W modelu regresji porządkowej nie interesuje nas, do którego z porównań dany parametr beta się odnosi- fakt ten jest konsekwencją założenia proporcjonalności odds. Gdyby założenie to nie zostało spełnione, alternatywnym wyjściem byłaby estymacja modelu o postaci wielomianu.

W przypadku zmiennych czynnikowych zastosowano kodowanie porządkowe. W tym celu dokonano odpowiednich zmian w kodzie programu. Dodatkowo do budowy modelu włączono interakcje zmiennych jakościowych, ilościowych i jakościowo-ilościowych. Żadna z interakcji nie okazała się statystycznie istotna.

Dla modelu za funkcję linkująca przyjęto LOGIT. Dla prezentowanego modelu z trzema kategoriami zmiennej celu możliwe są dwa umiejscowienia cut-point, czyli uzyskać można maksymalnie 2 różne ilorazy szans. Natomiast po przeanalizowaniu liczebności kategorii, jedynym sensownym rozwiązaniem okazuje się kategoryzacja zmiennej w następujący sposób:

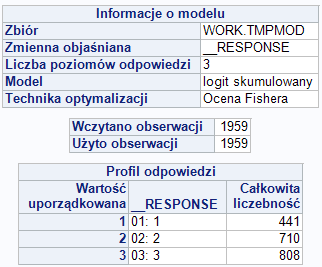
PPLFAIR = <1,2> vs PPLFAIR = 3

Model został zbudowany w oparciu o krokową selekcję zmiennych. Jest to modyfikacja wyboru następnych zmiennych i różni się od metody forward i backward tym, że zmienne, które zostaną raz zaklasyfikowane do modelu nie muszą w nim pozostać. Proces selekcji krokowej kończy się, kiedy żadna ze zmiennych niezaklasyfikowanych do modelu nie jest istotna na poziomie wystarczającym do wstawienia do modelu a wszystkie zaklasyfikowane charakteryzują się poziomem istotności wystarczającym do pozostawienia ich w modelu. Dla prezentowanego modelu poziom istotności określono jako 0,05.

## 3.1. Wyniki modelu uporządkowanej regresji logistycznej

W tej części projektu zaprezentowano wyniki modelu uporządkowanej regresji logistycznej.

*Tabela 4.* *Podstawowe wyniki modelu*



*Źródło: Opracowanie własne*

Każda z 1959 obserwacji została użyta w modelu dzięki wykorzystaniu PPM (patrz str. 5).

*Tabela 5. Status zbieżności modelu*



*Źródło: Opracowanie własne*

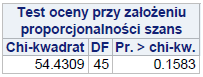
Kryterium zbieżności zostało spełnione, a więc znaleziono parametry największej wiarygodności.

Kolejno zweryfikowano test oceny przy założeniu proporcjonalności szans. Spełnienie założenia o niezmienności odds ratio jest podstawowym warunkiem estymacji uporządkowanego modelu regresji liniowej. Do weryfikacji tego założenia wykorzystywany jest test punktowy z poniższymi hipotezami:

H0: założenie o proporcjonalności odds jest w mocy

H1: założenie o proporcjonalności odds nie jest w mocy.

*Tabela 6. Test proporcjonalności szans*



*Źródło: Opracowanie własne*

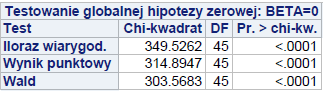
Wartośc p- value dla testu oceny przy założeniu proporcjonalności szans przy ostatnim kroku wyniósł 0,1583, a więc brak podstaw po odrzucenia hipotezy zerowej mówiącej o proporcjonalności szans- założenie zostało spełnione. Szanse w różnych grupach są proporcjonalne, co oznacza, że bez względu na podział kategorii zmiennej objaśnianej (czyli wybór, które kategorie są traktowane jako referencyjne) oszacowania ilorazów szans pozostają takie same.

Następnym etapem analizy była weryfikacja istotności parametrów modelu testując globalną hipotezę zerową:

H0:βj=0 (wszystkie parametry są równe zero- badany zestaw zmiennych objaśniających jest nieodpowiedni), przy hipotezie alternatywnej:

H1:βj ≠ 0 (przynajmniej jeden parametr jest istotny statystycznie).

*Tabela 7. Test Beta=0*



*Źródło: Opracowanie własne*

Wyniki testów, zarówno Walda jak i Score, przyjmują wartości p-value na poziomie niższym niż 0,001, co sugeruje odrzucenie hipotezy zerowej na rzecz alternatywnej, czyli przynajmniej jeden ze współczynników modelu jest istotnie różny od zera.

Dalej zbadano, czy każdy estymowany parametr pojedynczo istotnie różni się od zera. Metoda krokowej selekcji, pozwoliła pozostawić w modelu tylko zmienne statystycznie istotne. Kolejne kroki selekcji ukazuje Tabela 10.

*Tabela 8. Wyniki selekcji krokowej*



*Źródło: Opracowanie własne*

Ostateczny model zawiera 7 zmiennych istotnych statystycznie. Zmienne STFDEM (ocena funkcjonowania demokracji), IPSUCES, EDUYRS okazały się nieistotna.

*Tabela 9. Wyniki efektów typu 3*

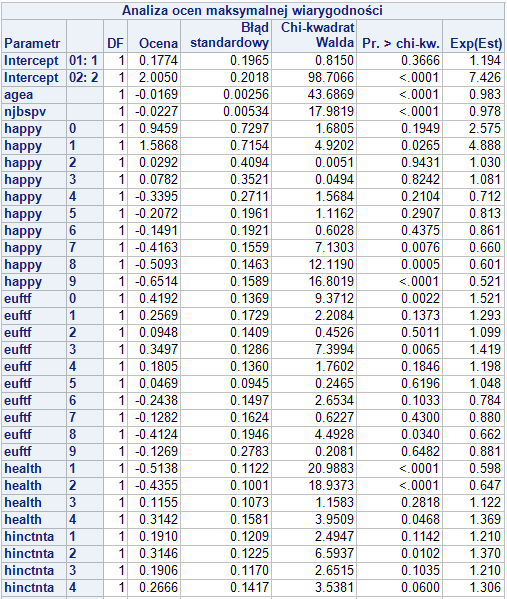
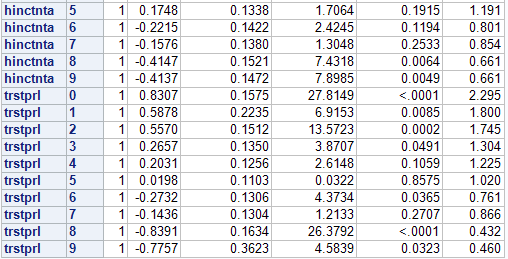


*Źródło: Opracowanie własne*

Wszystkie wartości p-value dla zmiennych przedstawionych w Tabeli 10 sugerują, że należy odrzucić hipotezę zerową na rzecz hipotezy alternatywnej, mówiącej o tym, że oceny parametrów istotnie różnią się od zera, co pozwala na ich uzasadnioną interpretację.

Na podstawie poniższych tabeli można dokonać interpretacji poszczególnych oszacowań parametrów. Ze względu na spełnienie założenia proporcjonalności szans interpretacja szans zarówno dla sytuacji bycia nieufnym w porównaniu do bycia średnioufnym bądź ufnym jest taka sama jak dla sytuacji bycia ufnym w porównaniu do bycia średnioufnym bądź nieufnym.

*Tabela 10. Analiza ocen maksymalnej wiarygodności*

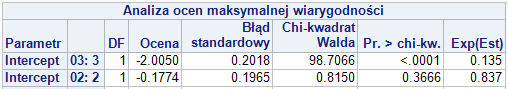
 

*Źródło: Opracowanie własne*

Interpretacja - stała:

Gdyby stwierdzenie ,,wszystkie inne zmienne przyjmują wartość 0” miało sens, interpretacja stałej byłaby następująca: dla kategorii 1 szanse, że respondent będzie nieufny, a nie średnioufny i ufny są jak 12:10. Dla kategorii 2, szanse, że respondent będzie średnioufny, a nie ufny są jak 7,5:1.

*Tabela 11. Analiza ocen maksymalnej wiarygodności przy dopasowaniu modelu do poziomu 3*



*Źródło: Opracowanie własne*

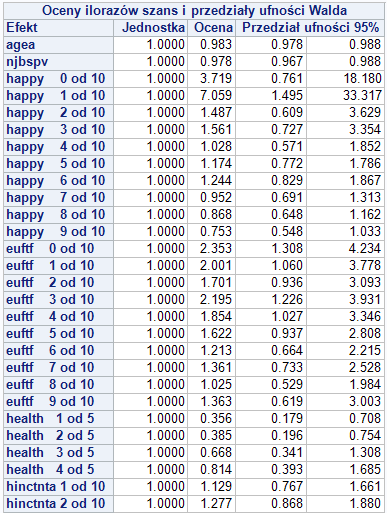
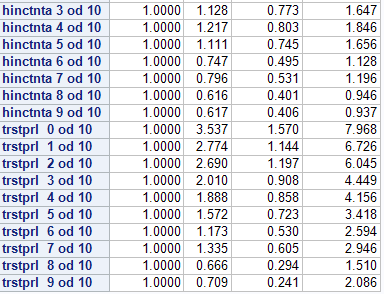
Przy dopasowaniu modelu do kategorii 3 interpretacja wygląda następująco: dla kategorii 3 szansa, że respondent będzie średnioufny lub nie ufny w relacji do ufnego to 13:100. Dla kategorii 2 szasa, że respondent będzie nieufny, a nie średnioufny lub ufny to 83:100.

Interpretacja- zmienne ciągłe:

Jeśli osoby będą różniły się tylko pod względem wieku, to osoba starsza o rok będzie miała o 2% mniejsze szanse na bycie ufnym niż osoba młodsza.

Jeśli osoby będą się różnić pod względem liczby podwładnych, to osoba z 1 więcej podwładnym będzie miała o 2% mniejsze szanse na bycie ufnym.

*Tabela 12.* *Ocena ilorazów szans*

*Źródło: Opracowanie własne*

Interpretacja - zmienne kategoryczna o kodowaniu typu efekt:

Kodowanie typu efekt nie umożliwia interpretacji wartości funkcji wykładniczej parametrów w kategoriach ilorazu szans. Wartość przydatną w zakresie interpretacji można znaleźć w tabeli powyżej. Żadna z kategorii nie jest istotna statystycznie.

## 3.2. Ocena jakości modelu regresji uporządkowanej

Do oceny modelu uporządkowanej regresji logistycznej wykorzystano:

• Współczynnik determinacji R2

• Procent zgodnych, niezgodnych obserwacji

• Statystyki D Somers, Gamma

• Statystka c, określająca pole pod krzywą ROC

• Kryterium Akaike, Bayesa-Schwartza oraz -2 log L

Współczynnik determinacji R2 to jedna z podstawowych miar jakości dopasowania modelu. Informuje o tym, jaka część zmienności zmiennej objaśnianej została wyjaśniona przez model. Jest miarą stopnia, w jakim model wyjaśnia kształtowanie się zmiennej objaśnianej.

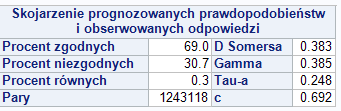
*Tabela 13. Współczynnik R-kwadrat*



*Źródło: Opracowanie własne*

Współczynnik determinacji R2 wyniósł 0,1634. Jakość dopasowania modelu do danych jest w omawianym modelu dość dobra (jak na model regresji logistycznej), Należy mieć na uwadze fakt, że niskie wartości współczynnika determinacji spowodowane są modelowaniem zjawiska na poziomie indywidualnym (mikroekonomicznym). Według literatury modele oparte na tego typu danych z reguły osiągają wartości z przedziału 5%-40%.

*Tabela 14. Miary skojarzenia*



*Źródło: Opracowanie własne*

Miary skojarzenia badające zdolności predykcyjne modelu wskazują na 69% poprawnie zaklasyfikowanych obserwacji, 30,7% źle zaklasyfikowanych i 0,3% przypadków bez decyzji. Statystyki D Somersa, Gamma oraz Tau-a na podstawie tablic kontyngencji testując niezależność zmiennych objaśnianej i objaśniających. Statystyka Gamma mówi o proporcjonalnej redukcji błędu, czyli nadwyżce zgodnych par ponad niezgodne wyrażana jako procent w stosunku do wszystkich wyodrębnionych par i wynosi 38%. Statystyka D Sommera uwzględnia też pary związane, których liczba w omawianym modelu jest bardzo niska, dlatego wartość tej statystyki jest zbliżona do statystyki Gamma. Statystyka c określa pole pod krzywą ROC. Statystyka c powstała jako zależność pomiędzy zgodnymi skojarzeniami par a popełnionymi przez model niezgodnościami przyporządkowań. Im większa wartość c tym model lepiej przewiduje wartości zmiennej zależnej. Dla tego modelu wartość pola pod krzywą ROC wyniosła 0,69.

Kolejnym etapem była analiza kryteriów informacyjnych: Akaike, Bayesa-Schwartza oraz -2 log L. Stosując kryteria informacyjne do wyboru modelu spośród zbioru modeli - kandydatów wybiera się ten model, któremu odpowiada minimalna wartość danego kryterium informacyjnego.

*Tabela 15.* *Statystyki dopasowania*



*Źródło: Opracowanie własne*

Na podstawie kryteriów informacyjnych nie da się wybrać modelu z optymalną liczbą zmiennych, da się jedynie porównać jakość przynajmniej dwóch modeli. Warto przyznać, że wartości AIC i SC powinny być jak najniższe.

# 4. Model regresji binarnej typu forward

Następnym modelem regresji opracowanym w tym referacie jest model regresji binarnej typu forward. Jest to metoda, która polega na stopniowym dołączaniu do modelu kolejnych zmiennych. W pierwszym kroku tworzony jest model bez zmiennych przyczynowych. Zmienne są kolejno dodawane do modelu według malejącej wartości statystyki resztowej (i rosnącego p-value dla tej statystyki). W każdym kroku do modelu dodawana jest ta zmienna, której statystyka resztowa ma najmniejsze p-value. Zmienne są dodawane dopóty, dopóki p-value jest mniejsze od określonej wartości maksymalnej (kryterium wejścia). Algorytm kończy się, kiedy dodanie dowolnej z pozostałych zmiennych powoduje, że p-value jest większe niż kryterium wejścia.

Metoda ta ma dwa ograniczenia. Niektóre ze zmiennych nigdy nie trafiają do modelu, przez co ich istotność nie jest nigdy określona. Drugim ograniczeniem jest to, że zmienna, która została dodana do modelu, pozostaje w nim już do końca algorytmu, nawet jeśli utraciła istotność po dodaniu innych zmiennych.

Zmienna PPLFAIR została przekonwertowana na zmienną binarną z wartościami:

* 0 dla kategorii 1 i 2 (nieufny i średnioufny). Można tą wartość potraktować jako ,,raczej nieufny”
* 1 dla kategorii 3 (ufny).

Charakterystyka zbioru została przedstawiona w rozdziale 2. Dla prezentowanego modelu poziom istotności określono jako 0,05.

W przedstawionym modelu, każda ze zmiennych posiada tylko 1 stopień swobody. Autor uznał, że dla potrzeb łatwiejszej interpretacji wyników jeden z modeli przedstawionych w referacie powinien posiadać właśnie taki rodzaj kodowania zmiennych.

## 4.1 Wyniki modelu regresji logistycznej binarnej typu forward

*Tabela 16. Podstawowe wyniki modelu forward*

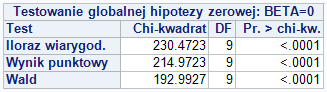


*Źródło: Opracowanie własne*

Charakterystyka zbioru została wspomniana w poprzednich rozdziałach. Różnicą jest tu wartość zmiennej zależnej: są tylko 2 wartości uporządkowanej. Liczebność pierwszej wynosi 1151, drugiej 808. Nie są to ilości bliźniacze, ale autor uznał je za wystarczająco zbliżone i optymalne przy tworzeniu analizy.

Status zbieżności- kryterium zbieżności został spełnione.

*Tabela 17. Test Beta=0 dla forward*



*Źródło: Opracowanie własne*

Weryfikacja istotności parametrów modelu testując globalną hipotezę zerową. Wyniki testów, zarówno Walda jak i Score, przyjmują wartości p-value na poziomie niższym niż 0,001, co sugeruje odrzucenie hipotezy zerowej na rzecz alternatywnej.

*Tabela 18. Wyniki selekcji postępującej (forward)*



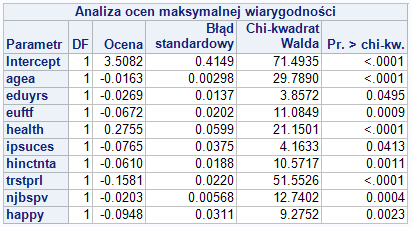
*Źródło: Opracowanie własne*

Model ostateczny zawiera 8 zmiennych istotnych statystycznie. Każda w tym modelu posiada tylko 1 stopień swobody, co ułatwia interpretację. Iteracja nr 9 zakończyła się na poziomie p-value 0.048, to jest poniżej ustalonego progu 0.05. Zmienne najbardziej statystycznie istotne to TRSTPRL, HEALTH i HINCTNTA gdzie p-value jest poniżej 0.001.

W poniższej tabeli uwzględniono analizę ocen maksymalnej wiarygodności. Po wynikach modelu można dojść do następujących wniosków:

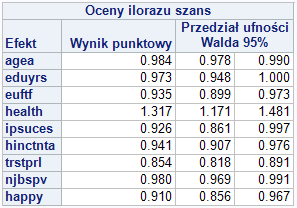
* wyraz wolny dla tego modelu wynosi 3,5. Oznacz to, że gdyby wszystkie pozostałe zmienne miały wartość 0, taki byłby poziom zaufania respondenta
* respondent starszy o 1 rok, będzie miał o 2% mniejsze zaufanie, niż respondent rok młodszy
* respondent uczęszczający do szkoły o 1 rok dłużej, będzie miał o 3% mniejsze zaufanie, w porównaniu z respondentem, który uczęszczał o 1 rok krócej
* respondent, który wierzy w integrację europejską o 1/10 bardziej, będzie miał o 7% mniejsze zaufanie do ludzi
* respondent, który będzie miał o 1/5 gorsze zdrowie, będzie ufał ludziom o 28% bardziej
* respondent, który uważa że sukcesy są ważne o 1/6 bardziej, będzie ufał ludziom 8% bardziej
* respondent, który wierzy w parlament o 1/10 bardziej, będzie ufał ludziom o 15% mniej
* respondent, który ma o 1 podwładnego więcej, będzie ufał ludziom o 2% mniej
* respondent, który jest o 1/10 bardziej szczęśliwy, będzie ufał ludziom o 9% mniej

*Tabela 19. Analiza ocen maksymalnej wiarygodności dla forward*



*Źródło: Opracowanie własne*

*Tabela 20. Oceny ilorazu szans dla forward*



*Źródło: Opracowanie własne*

*Rysunek 13. Ilorazy szans*



*Źródło: Opracowanie własne*

## 4.2. Ocena jakości modelu regresji binarnej typu forward

Do oceny jakości modelu użyto tych samych miar, co do regresji uporządkowanej.

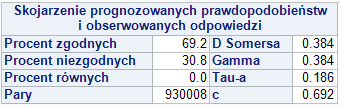
*Tabela 21. Współczynnik R-kwadrat dla forward*



*Źródło: Opracowanie własne*

Współczynnik determinacji R2 wyniósł 0,1495. Jakość dopasowania modelu do danych jest w omawianym modelu dobra.

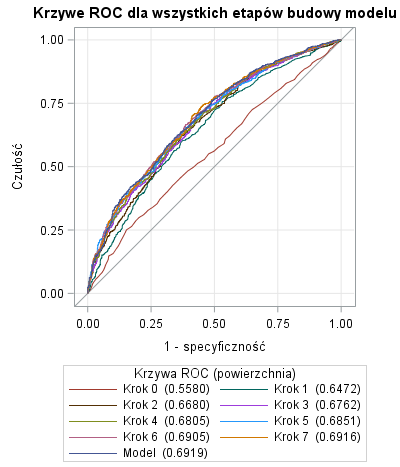
*Tabela 22. Miary skojarzenia dla forward*



*Źródło: Opracowanie własne*

Miary skojarzenia badające zdolności predykcyjne modelu wskazują na 69,2% poprawnie zaklasyfikowanych obserwacji, 30,8% źle zaklasyfikowanych i 0,4% przypadków bez decyzji. Statystyki D Somersa, Gamma oraz Tau-a na podstawie tablic kontyngencji testując niezależność zmiennych objaśnianej i objaśniających. Statystyka Gamma mówi o proporcjonalnej redukcji błędu, czyli nadwyżce zgodnych par ponad niezgodne wyrażana jako procent w stosunku do wszystkich wyodrębnionych par i wynosi 38 %. Statystyka D Sommera uwzględnia też pary związane, których liczba w omawianym modelu jest bardzo niska, dlatego wartość tej statystyki jest zbliżona do statystyki Gamma. Statystyka c określa pole pod krzywą ROC. Statystyka c powstała jako zależność pomiędzy zgodnymi skojarzeniami par a popełnionymi przez model niezgodnościami przyporządkowań. Im większa wartość c tym model lepiej przewiduje wartości zmiennej zależnej. Dla modelu forward po 8 iteracjach wartość pola pod krzywą ROC wyniosła 0,692.

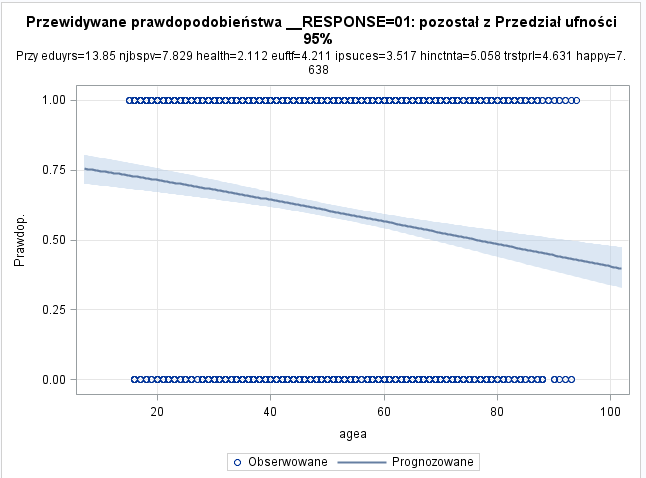
*Rysunek 14. Krzywe ROC dla każdej z iteracji*



*Źródło: Opracowanie własne*

Wykres krzywej ROC ukazuje jak jakość dopasowania modelu poprawiała się wraz z każdą iteracją. Można również zauważyć, ze już od kroku 2. do ostatniego jakość modelu była zbliżona.

*Rysunek 15. Przewidywane prawdopodobieńśtwa*



*Źródło: Opracowanie własne*

Na wykresie można zobaczyć, jak funkcja prawdopodobieństwa maleje wraz z wiekiem. Oznacza to, że w wraz z wiekiem, respondenci stają się mniej ufni.

*Tabela 23. Statystyki dopasowania*



*Źródło: Opracowanie własne*

## 4.3 Ocena jakości modelu regresji binarnej typu backward

Dla potrzeb referatu autor przeprowadził również analizę z użyciem modelu regresji logistycznej typu backward. Metoda backward oznacza, że usuwane są najmniej istotne zmienne z modelu zawierającego wszystkie zmienne objaśniające dopóki wszystkie zmienne w modelu będą istotne.

Wyniki modelu okazały się bliźniaczo podobne jak w regresji typu forward. Jedyną róźnicą jest liczba kroków- w metodzie backward przeprowadzono tylko jedna iterację- usunięcie zmiennej STFDEM.

# 5. Bibliografia

1. Allison P.D., “Logistic Regression Using SAS Theory and Application”, SAS Press

2. ,,Zaawansowane metody analiz statystycznych”, red. Naukowa Ewa Frątczak, Oficyna Wydawnicza SGH, Warszawa 2013

3. „Ekonometria i badania operacyjne”, red. naukowa m. Gruszczyński, T. Kuszewski, M. Podgórska, Wydawnictwo Naukowe PWN, Warszawa 2013

# 6. Spis tabel i rysunków

Tabela 1. Charakterystyki zmiennych…………………………………………………………………..3

Tabela 2. Rozkład zmiennej PPLFAIR…………………………………………………………………7

Tabela 3. Współczynniki korelacji Pearsona……………………………………………………………7

Tabela 4. Podstawowe wyniki modelu………………………………………………………………...14

Tabela 5. Status zbieżności modelu……………………………………………………………………14

Tabela 6. Test proporcjonalności szans………………………………………………………………..15

Tabela 7. Test Beta=0………………………………………………………………………………….15

Tabela 8. Wyniki selekcji krokowej…………………………………………………………………...16

Tabela 9. Wyniki efektów typu 3………………………………………………………………………16

Tabela 10. Analiza ocen maksymalnej wiarygodności………………………………………………...17

Tabela 11. Analiza ocen maksymalnej wiarygodności przy dopasowaniu modelu do poziomu 3……18

Tabela 12. Ocena ilorazów szans………………………………………………………………………19

Tabela 13. Współczynnik R-kwadrat…………………………………………………………………..20

Tabela 14. Miary skojarzenia…………………………………………………………………………..20

Tabela 15. Statystyki dopasowania…………………………………………………………………….21

Tabela 16. Podstawowe wyniki modelu forward………………………………………………………22

Tabela 17. Test Beta=0 dla forward……………………………………………………………………23

Tabela 18. Wyniki selekcji postępującej (forward) …………………………………………………...23

Tabela 19. Analiza ocen maksymalnej wiarygodności dla forward…………………………………...24

Tabela 20. Oceny ilorazu szans dla forward…………………………………………………………...24

Tabela 21. Współczynnik R-kwadrat dla forward……………………………………………………..25

Tabela 22. Miary skojarzenia dla forward……………………………………………………………..25

Tabela 23. Statystyki dopasowania…………………………………………………………………….27

Rysunek 1. Rozkład zmiennej PPLFAIR - 11 poziomów………………………………………………5

Rysunek 2. Rozkład zmiennej PPLFAIR - 3 poziomy………………………………………………….6

Rysunek 3. Rozkład zmiennej AGEA z uwzględnieniem PPLFAIR…………………………………...8

Rysunek 4. Rozkład zmiennej EDUYRS z uwzględnieniem PPLFAIR………………………………..8

Rysunek 5. Rozkład zmiennej EUFTF z uwzględnieniem PPLFAIR…………………………………..9

Rysunek 6. Rozkład zmiennej HAPPY z uwzględnieniem PPLFAIR………………………………….9

Rysunek 7. Rozkład zmiennej HEALTH z uwzględnieniem PPLFAIR………………………………10

Rysunek 8. Rozkład zmiennej IPSUCES z uwzględnieniem PPLFAIR………………………………10

Rysunek 9. Rozkład zmiennej STFDEM z uwzględnieniem PPLFAIR………………………………11

Rysunek 10. Rozkład zmiennej HINCTNTA z uwzględnieniem PPLFAIR………………………….11

Rysunek 11. Rozkład zmiennej TRSTPRL z uwzględnieniem PPLFAIR…………………………….12

Rysunek 12. Rozkład zmiennej NJBSPV z uwzględnieniem PPLFAIR………………………………12

Rysunek 13. Ilorazy szans…………………………………...…………………………………...........25

Rysunek 14. Krzywe ROC dla każdej z iteracji…………………………………...…………………..26

Rysunek 15. Przewidywane prawdopodobieńśtwa………………………………….............................27

# 7. Kody SAS

**Regresja uporządkowana**

/\* -------------------------------------------------------------------

Kod wygenerowany przez zadanie SAS-a

Wygenerowany dnia: niedziela, 28 stycznia 2018 o godz. 18:33:10

Przez zadanie: Regresja logistyczna

Dane wejściowe: Local:WORK.RDATA

Serwer: Local

------------------------------------------------------------------- \*/

ODS GRAPHICS ON;

%***\_eg\_conditional\_dropds***(WORK.SORTTempTableSorted,

WORK.TMPMod);

/\* -------------------------------------------------------------------

Sortowanie zbioru Local:WORK.RDATA

------------------------------------------------------------------- \*/

**PROC** **SQL**;

CREATE VIEW WORK.SORTTempTableSorted AS

SELECT T.pplfair, T.agea, T.eduyrs, T.stfdem, T.njbspv, T.happy, T.hinctnta, T.euftf, T.trstprl, T.ipsuces, T.health

FROM WORK.RDATA as T

;

**QUIT**;

**DATA** WORK.TMPMod;

SET WORK.SORTTempTableSorted;

length \_\_RESPONSE $ **10**;

IF pplfair=**1** THEN \_\_RESPONSE="01: 1";

IF pplfair=**2** THEN \_\_RESPONSE="02: 2";

IF pplfair=**3** THEN \_\_RESPONSE="03: 3";

**RUN**;

TITLE;

TITLE1 "Rezultaty regresji logistycznej";

FOOTNOTE;

FOOTNOTE1 "Wygenerowane przez System SAS (&\_SASSERVERNAME, &SYSSCPL) dnia %TRIM(%QSYSFUNC(DATE(), NLDATE20.)) o godz. %TRIM(%SYSFUNC(TIME(), NLTIMAP20.))";

**PROC** **LOGISTIC** DATA=WORK.TMPMod

PLOTS(ONLY)=ALL

;

CLASS happy (PARAM=EFFECT) hinctnta (PARAM=EFFECT) euftf (PARAM=EFFECT) trstprl (PARAM=EFFECT) ipsuces (PARAM=EFFECT) health (PARAM=EFFECT);

MODEL \_\_RESPONSE (DESCENDING)=agea eduyrs stfdem njbspv happy euftf health hinctnta trstprl ipsuces / expb

SELECTION=STEPWISE

SLE=**0.05**

SLS=**0.05**

INCLUDE=**0**

RSQUARE

LINK=LOGIT

CLPARM=WALD

CLODDS=WALD

ALPHA=**0.05**

;

**RUN**;

**QUIT**;

/\* -------------------------------------------------------------------

Koniec kodu zadania

------------------------------------------------------------------- \*/

**RUN**; **QUIT**;

%***\_eg\_conditional\_dropds***(WORK.SORTTempTableSorted,

WORK.TMPMod);

TITLE; FOOTNOTE;

ODS GRAPHICS OFF;

**Regresja binarna typu forward:**

ODS GRAPHICS ON;

%***\_eg\_conditional\_dropds***(WORK.SORTTempTableSorted,

WORK.TMPMod);

**PROC** **SQL**;

CREATE VIEW WORK.SORTTempTableSorted AS

SELECT T.pplfair, T.agea, T.eduyrs, T.euftf, T.happy, T.health, T.ipsuces, T.stfdem, T.hinctnta, T.trstprl, T.njbspv

FROM WORK.RDATA as T

;

**QUIT**;

**DATA** WORK.TMPMod;

SET WORK.SORTTempTableSorted;

length \_\_RESPONSE $ **20**;

IF strip(put(pplfair,MOJA.))="inne" THEN \_\_RESPONSE="01: inne";

IF strip(put(pplfair,MOJA.))="trojki" THEN \_\_RESPONSE="02: trojki";

**RUN**;

TITLE;

TITLE1 "Rezultaty regresji logistycznej";

FOOTNOTE;

FOOTNOTE1 "Wygenerowane przez System SAS (&\_SASSERVERNAME, &SYSSCPL) dnia %TRIM(%QSYSFUNC(DATE(), NLDATE20.)) o godz. %TRIM(%SYSFUNC(TIME(), NLTIMAP20.))";

**PROC** **LOGISTIC** DATA=WORK.TMPMod

PLOTS(ONLY)=ALL

;

MODEL \_\_RESPONSE=agea eduyrs euftf health ipsuces stfdem hinctnta trstprl njbspv happy /

SELECTION=FORWARD

SLE=**0.05**

INCLUDE=**1**

RSQUARE

LINK=LOGIT

CLPARM=WALD

ALPHA=**0.05**

;

FORMAT pplfair MOJA.;

**RUN**;

**QUIT**;

/\* -------------------------------------------------------------------

Koniec kodu zadania

------------------------------------------------------------------- \*/

**RUN**; **QUIT**;

%***\_eg\_conditional\_dropds***(WORK.SORTTempTableSorted,

WORK.TMPMod);

TITLE; FOOTNOTE;

ODS GRAPHICS OFF;

**Tworzenie formatu dla zmiennej PPLFAIR:**

/\* -------------------------------------------------------------------

Kod wygenerowany przez zadanie SAS-a

Wygenerowany dnia: czwartek, 25 stycznia 2018 o godz. 01:12:29

Przez zadanie: Tworzenie formatu (moja - Local)

Dane wejściowe: .''n

Serwer:

------------------------------------------------------------------- \*/

%***\_eg\_conditional\_dropds***(WORK.CFMT\_0000);

**PROC** **FORMAT**

LIB=WORK

;

VALUE moja ( DEFAULT=**8** )

**3** = "trzy"

OTHER = "pozostałe";

**RUN**;

**RUN**; **QUIT**;

TITLE; FOOTNOTE;

**Inputowanie zbioru danych:**

**DATA** rdata ;

/\* -------------------------------------------------------------------

Proszę zmienić ścieżkę pliku na właściwą

------------------------------------------------------------------- \*/

INFILE "C:/Users/Sebastian/Desktop/Analiza danych- Big data 2015-18/4. Semestr/Regresja logistyczna/ECLIB000/mydata.txt"

DSD

LRECL= **31** ;

INPUT

pplfair

agea

eduyrs

euftf

happy

health

ipsuces

stfdem

hinctnta

trstprl

njbspv

;

**RUN**;

**proc** **print** data= rdata;

**run**;

**Korelacje:**

/\* -------------------------------------------------------------------

Kod wygenerowany przez zadanie SAS-a

Wygenerowany dnia: niedziela, 28 stycznia 2018 o godz. 18:41:13

Przez zadanie: Korelacje

Dane wejściowe: Local:WORK.RDATA

Serwer: Local

------------------------------------------------------------------- \*/

ODS GRAPHICS ON;

%***\_eg\_conditional\_dropds***(WORK.SORTTempTableSorted);

/\* -------------------------------------------------------------------

Sortowanie zbioru Local:WORK.RDATA

------------------------------------------------------------------- \*/

**PROC** **SQL**;

CREATE VIEW WORK.SORTTempTableSorted AS

SELECT T.pplfair, T.agea, T.eduyrs, T.euftf, T.happy, T.health, T.ipsuces, T.stfdem, T.hinctnta, T.trstprl, T.njbspv

FROM WORK.RDATA as T

;

**QUIT**;

TITLE;

TITLE1 "Analiza korelacji";

FOOTNOTE;

FOOTNOTE1 "Wygenerowane przez System SAS (&\_SASSERVERNAME, &SYSSCPL) dnia %TRIM(%QSYSFUNC(DATE(), NLDATE20.)) o godz. %TRIM(%SYSFUNC(TIME(), NLTIMAP20.))";

**PROC** **CORR** DATA=WORK.SORTTempTableSorted

PLOTS=NONE

COV

PEARSON

KENDALL

VARDEF=DF

;

VAR pplfair agea eduyrs euftf happy health ipsuces stfdem hinctnta trstprl njbspv;

**RUN**;

/\* -------------------------------------------------------------------

Koniec kodu zadania

------------------------------------------------------------------- \*/

**RUN**; **QUIT**;

%***\_eg\_conditional\_dropds***(WORK.SORTTempTableSorted);

TITLE; FOOTNOTE;

ODS GRAPHICS OFF;