Przedmiot: Regresja Logistyczna z wykorzystaniem narzędzi SAS

Projekt nr 1 Tematyka: Analiza determinant oceny subiektywnego poziomu zdrowia w Niemczech

Osoba oceniająca: dr hab. prof. SGH Ewa Frątczak

Imię, nazwisko studenta: Karol Szerszeń, Magda Waśkiel, Szymon Ptaszyński

Nr albumu: 71273, 71361, 71039

Data oddania projektu: 18.12.2016

| Obszar oceny | Uzyskana liczba punktów | Uwagi |
| --- | --- | --- |
| **wartość merytoryczna**  (0-10 pkt) |  |  |
| **programowanie SAS**  (0-4 pkt) |  |  |
| **estetyka projektu**  (0-2 pkt) |  |  |
| **wkład pracy własnej**  (0-4 pkt) |  |  |
|  |  | Ocena końcowa: |

Contents

**No table of contents entries found.**

1. **Cel i zakres analizy**

Według Światowej Organizacji Zdrowia „Zdrowie to nie tylko całkowity brak choroby, czy kalectwa, ale także stan pełnego, fizycznego, umysłowego i społecznego dobrostanu (dobrego samopoczucia)”. Ocena stanu zdrowia jest jednym z kluczowych czynników wykorzystywanych do oceny jakości życia. Współcześnie do badań stanu zdrowia wykorzystuje się informacje z dwóch źródeł:

1. Obiektywne – potwierdzone badaniami wykonywanymi przez profesjonalistów z dziedziny medycyny. W tej kategorii znajdą się wskaźniki takie jak współczynnik umieralności noworodków, zmniejszenie śmiertelności, czy poprawa warunków klinicznych.
2. Subiektywne – samoocena, czyli klasyfikacja dokonywana osobiście przez samych badanych.

Metody obiektywne są fundamentalne, jednak obecnie ich znaczenie jest zdecydowanie większe w państwach słabo rozwiniętych. Wynika to z faktu, iż kraje wysoko rozwinięte posiadają stabilny i w większości wypadków obiektywnie efektywny system opieki zdrowotnej. W tym przypadku, np. zbyt niska liczba lekarzy na 100 mieszkańców, czy wysoka śmiertelność wśród noworodków będą problemami krajów słabo rozwiniętych. Dla porównania w 2015 Niemczech współczynnik umieralności noworodków wyniósł 4 promile, a w Angoli 96 promili.[[1]](#footnote-1) Wskaźniki obiektywne mają, więc różne zastosowania w zależności od stopnia rozwoju kraju. W krajach wysoko rozwiniętych mają one przede wszystkich funkcje informacyjne, monitorujące, natomiast w krajach słabo rozwiniętych są one wskaźnikami kontrolnymi i są wykorzystywane w procesie rozwiązywania podstawowych problemów zdrowotnych. Z zachowaniem powyższych funkcji, metody obiektywne znajdują zastosowanie w, m.in. epidemiologii.

W przypadku krajów wysokorozwiniętych o niealarmujących wskaźnikach obiektywnych podkreśla się znaczenie czynników subiektywnych. Pomagają one przede wszystkim w identyfikacji grup ryzyka, co umożliwia działania prewencyjne. Wykorzystywane są również w procesie budowania strategii zdrowotnych prowadzonych przez instytucje państwowe, np. w kampaniach społecznych. Dodatkowo, podkreśla się ich wagę w oszacowaniu potencjalnych korzyści i strat wynikających z podjętych działań medycznych, szczególnie obecnie, gdy wysoki poziom techniczny umożliwia uratowanie lub przedłużenie życia kosztem obniżenia jego jakości.

W niniejszej pracy skupiono się na analizie czynników wpływających na subiektywną ocenę zdrowia obywateli Niemiec. Społeczeństwo niemieckie jest świetnym przykładem reprezentującym kraj wysokorozwinięty. Interesującym zagadnieniem, więc jest postrzeganie swojego stanu zdrowia wśród osób, które mają dostęp do nowoczesnej służby zdrowia, a ich świadomość społeczna na tematy związane z ochroną zdrowia jest kształtowana przez najnowsze trendy w tym zakresie. Z drugiej strony obserwuje się zmianę problemów zdrowotnych wynikających ze stylu życia, np. wzrost otyłości wśród obywateli krajów wysoko rozwiniętych, ze względu na powszechny dostęp do żywności wysoko przetworzonej, towarów luksusowych oraz używek – takich jak alkohol.

Hipotezy badawcze niniejszej pracy to:

1. Największy wpływ na wysoką ocenę stanu zdrowia mają czynniki pozytywne (aktywność fizyczna, jedzenie warzyw, sen). Zdaje się, iż ćwiczenia fizyczne, odpowiednia dieta i odpoczynek są czynnikami, które najbardziej powinny poprawiać ludzkie samopoczucie.
2. Osoby o wysokiej aktywności fizycznej najlepiej oceniają swój stan zdrowia. Jak wiadomo, ćwiczenia fizyczne są zalecane przez lekarzy dla ludzi w każdym wieku, czy posiadają one realny wpływ na wartości samooceny zdrowia?
3. Czy interakcja częstotliwości spożycia alkoholu i problemów ze snem ma istotny wpływ na negatywną ocenę stanu zdrowia?
4. Czy ludzie młodzi oceniają swój stan zdrowia lepiej niż ludzie starsi? Czy widać wyraźną zmianę oceny stanu zdrowia między poszczególnymi grupami wiekowymi?
5. Osoby mieszkające na wsi oceniają swój stan zdrowia lepiej niż osoby zamieszkujące miasta. Współcześnie zauważa się przepływ ludności ze wsi do miast, szczególnie ze względów ekonomicznych, czy jednak pozostanie na wsi wpływa znacząco na ocenę stanu zdrowia?

Zmienne zostały dobrane dwuetapowo: pierwotnie wybrano kilkanaście cech na podstawie intuicyjnego ich wpływu na stan zdrowia. W zależności od rodzaju zmiennych dokonano ich odpowiedniej kategoryzacji. Następnie za pomocą wybranych metod zweryfikowano ich istotność statystyczną w modelu.

Projekt składa się z rozdziałów obejmujących przedstawienie wybranych zmiennych, omówienie modelu regresji logistycznej binarnej oraz podsumowanie wyników analizy.

1. **Prezentacja zbioru i przygotowanie danych**

2.1 Prezentacja zbioru danych

Analiza została przeprowadzona na podstawie zbioru danych z siódmej edycji badania ESS (European Social Survey) przeprowadzonego w roku 2014 i udostępnionego przez Norwegian Social Science Data Services. Surowy zbiór danych składa się z 601 zmiennych oraz 40185 obserwacji. Z niego zostały wyodrębnione obserwacje odpowiadające respondentom z interesującego nas kraju czyli Niemiec (zmienna country = ‘DE’). Następnie odrzucono wszystkie obserwacje, które nie niosły ze sobą żadnej informacji o opinii respondenta na temat jego stanu zdrowia – czyli włączona tylko te obserwacje, dla których zmienna *health* przyjmuje wartości z przedziału 1 do 5. W ostatecznym zbiorze znalazło się 3040 obserwacji oraz 10 zmiennych – w tym zmienna celu oraz zmienna wagowa. W wyniku selekcji otrzymano zbiór zawierający jedną zmienną binarną, trzy zmienne ilościowe oraz sześć porządkowych. Poniższa tabela prezentuje wszystkie zmienne użyte w analizie.

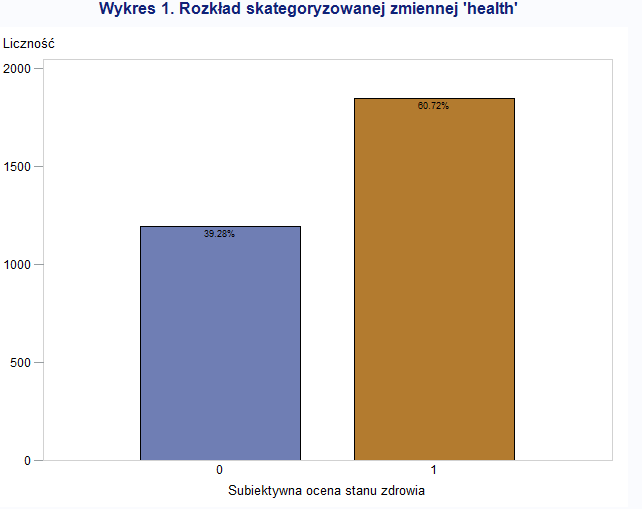
|  |  |  |
| --- | --- | --- |
| Zmienna | Etykieta | Typ |
| health | Subiektywna ocena stanu zdrowia | Porządkowa (5 kategorii) |
| agea | Wiek respondenta | Ilościowa (15 -102) |
| cgtsmoke | Częstotliwość palenia papierosów | Porządkowa (5 kategorii) |
| dosprt | Aktywność fizyczna w ciągu ostatnich 7 dni | Ilościowa(0 - 7) |
| eatveg | Częstotliwość jedzenia warzyw | Porządkowa (6 kategorii) |
| alcfreq | Częstotliwość spożycia alkoholu | Porządkowa (7 kategorii) |
| slprl | Częstotliwość problemów ze snem w ciągu ostatnich 7 dni | Porządkowa (6 kategorii) |
| gndr | Płeć | Binarna {1,2} |
| domicil | Miejsce zamieszkania | Porządkowa (5 kategorii) |
| dweight | Zmienna wagowa | Ilościowa |

***Tabela 1. Opis zmiennych wykorzystywanych w analizie*** *Opracowanie własne na podstawie:* <http://nesstar.ess.nsd.uib.no/webview/>

Jako zmienną celu przyjętą *health,* czyli poziom subiektywnego stanu zdrowia. Jest to zmienna porządkowa przyjmująca 5 kategorii, zaprezentowana w poniższej tabeli. Rozkład zmiennej jest asymetryczny prawostronnie. Przed przystąpieniem do modelowania zmienną objaśnianą należy zbinaryzować.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | |  | | --- | |  | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | |  |  |  |  |  | | --- | --- | --- | --- | --- | | Subiektywna ocena stanu zdrowia | | | | | | health | **Liczebność** | **Procent** | **Liczebność skumulowana** | **Procent skumulowany** | | 1 | 518.15 | 17.03 | 518.15 | 17.03 | | 2 | 1370.23 | 45.03 | 1888.38 | 62.06 | | 3 | 884.82 | 29.08 | 2773.2 | 91.14 | | 4 | 228.17 | 7.50 | 3001.37 | 98.64 | | 5 | 41.33 | 1.36 | 3042.7 | 100.00 |   ***Tabela 2. Rozkład zmiennej health*** *Opracowanie własne na podstawie oprogramowania SAS*  ***Rysunek 1. Rozkład liczebności zmiennej health*** *Opracowanie własne na podstawie oprogramowania SAS* | | | |

2.2 Przygotowanie zbioru danych

Ze względu na występowanie kategorii mało licznych, a także w celu uproszczenia modelu dokonano grupowania kategorii. Zdecydowano się na skategoryzowanie zmiennej *health* na dwie kategorie, co umożliwi budowę modelu regresji logistycznej binarnej z wykorzystaniem tejże zmiennej, jako obiektu modelowania. Podziału dokonano ze względu na liczebności skumulowane, które prowadzą do rozkładu zmiennej w przybliżeniu 50%. Najbliżej tej wartości okazał się podział przypisujący do kategorii ‘0’ obserwacje z pierwotnych kategorii 1 oraz 2. Tak, więc w zbiorze wynikowym zmienna *health* o wartości 0 będzie oznaczać negatywną ocenę stanu zdrowia, natomiast 1 – pozytywną. Taki podział ma również sens ze względów merytorycznych – osoby oceniające swój stan zdrowia jak bardzo dobry oraz dobry odnoszą się pozytywnie do swojego stanu, natomiast pozostałe osoby raczej postrzegają go negatywnie.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | | Subiektywna ocena stanu zdrowia | | | | | | health | **Liczebność** | **Procent** | **Liczebność skumulowana** | **Procent skumulowany** | | 0 | 1154.32 | 37.94 | 1154.32 | 37.94 | | 1 | 1888.38 | 62.06 | 3042.7 | 100.00 | |

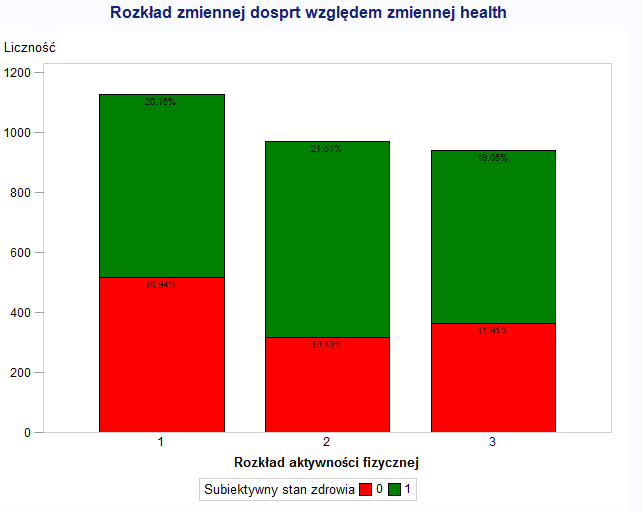
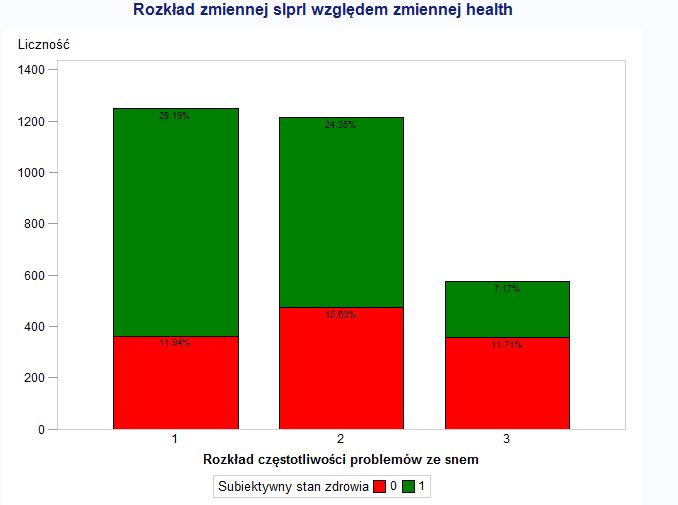
***Tabela 3. Rozkład skategoryzowanej zmiennej health*** *Opracowanie własne na podstawie oprogramowania SAS*

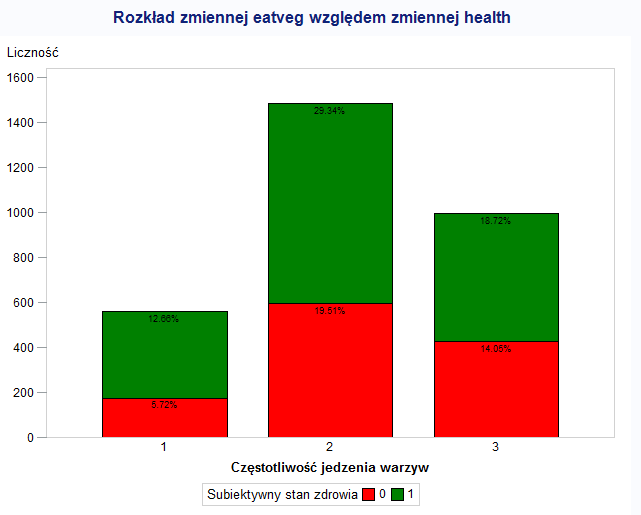
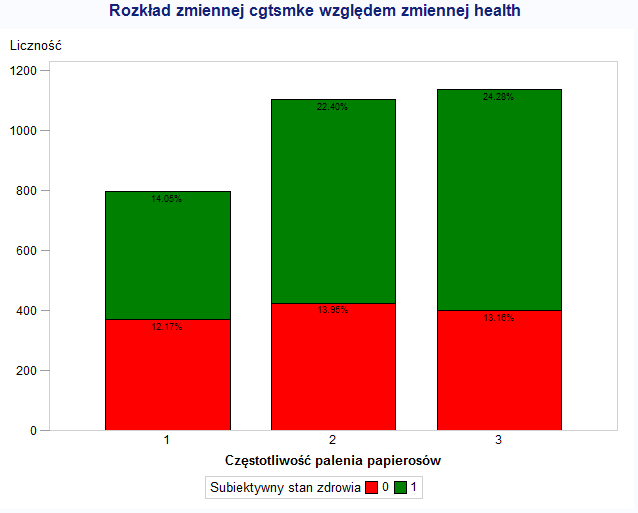
|  |
| --- |
|  |

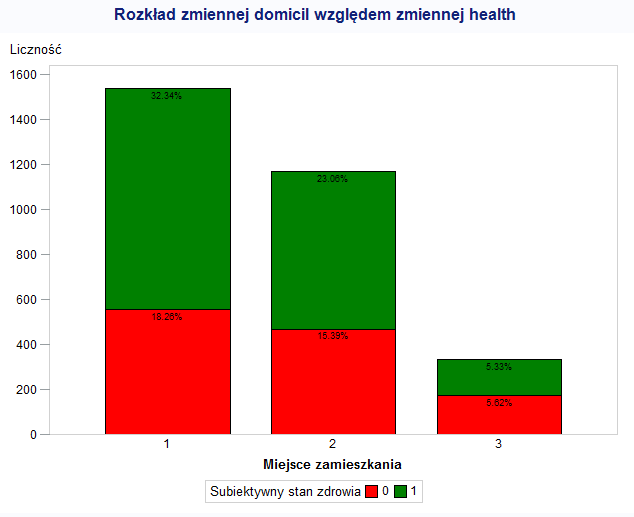
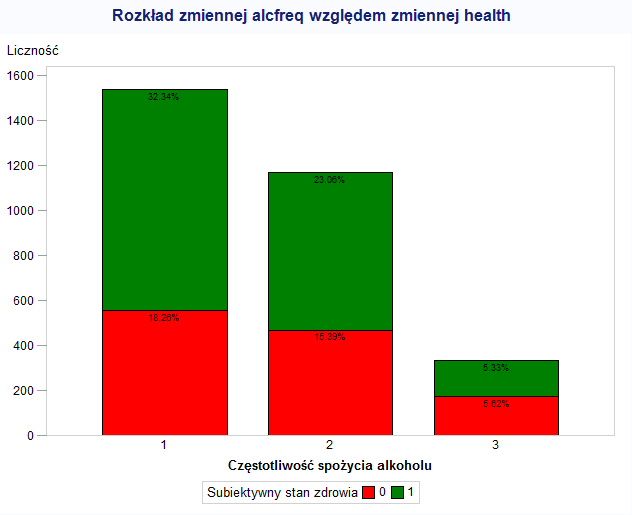
Kategoryzacji poddano również zmienne objaśniające, głównie ze względu na niewielkie liczności niektórych kategorii oraz możliwość merytorycznego pogrupowania większości z nich. Zmienne porządkowe zostały podzielone na 3 kategorie, według schematu prezentowanego w Tabeli 4. Rozkłady poszczególnych zmiennych względem zmiennej ‘*health*’ prezentują poniższe wykresy.

|  |  |  |
| --- | --- | --- |
| Zmienna | Typ i początkowe kategorie | Podział |
| health | Porządkowa (5 kategorii) | {1,2} – negatywna, {3,4,5} - pozytywna |
| cgtsmoke | Porządkowa (5 kategorii) | {1,2} – palacz, {3,4} – były palacz, {5} – osoba niepaląca |
| dosprt | Ilościowa(0 - 7) | {1, 2} – mało aktywna, {3,4,5} – aktywna, {6, 7} – bardzo aktywna |
| eatveg | Porządkowa (6 kategorii) | {1,2} – dwa razy dziennie, {3} – raz dziennie, {4,5,6} – mniej niż raz dziennie |
| alcfreq | Porządkowa (7 kategorii) | {1,2,3} – często, {4,5,6} – czasami, {7} – nigdy |
| slprl | Porządkowa (6 kategorii) | {1} – spokojny, {2} – raczej spokojny, {3, 4, 5, 6} - niespokojny |
| domicil | Porządkowa (5 kategorii) | {1,2} – duże miasto, {3} – małe miasto, {4,5} - wieś |

***Tabela 4. Podział zmiennych porządkowych*** *Opracowanie własne*







Jak można dostrzec na wykresach niestety nie we wszystkich kategoriach udało się uzyskać równomierny rozkład kategorii, głównie ze względu na liczności poszczególnych kategorii połączonych z brakiem merytorycznego uzasadnienie ich łączenia. Niemniej udało się uzyskać wystarczająco liczne kombinacje kategorii zmiennych objaśniających ze zmienną zależną.

Ze względu na skategoryzowanie większości zmiennych zbiór nie zawiera obserwacji skrajnych, które miałyby wpływ na model.

Do modelu włączono interakcję pomiędzy zmiennymi slprl i alcfreq czyli zależność pomiędzy problemami ze snem i częstotliwością spożycia alkoholu. Częstotliwość problemów ze snem może mieć wpływ na subiektywną ocenę stanu zdrowia w zależności od częstotliwości spożywanego alkoholu.

2.3 Kodowanie zmiennych

Kodowanie zmiennych niezależnych na potrzeby modelu regresji logistycznej ma decydujący wpływ na końcową interpretację otrzymanych wyników analizy. W niniejszej pracy zdecydowano się w dużej mierze na referencyjne kodowanie zmiennych, jedynie w przypadku jednej zmiennej zastosowano kodowanie porządkowe. Tabela 5 zawiera zestawienie informacji dotyczących kodowania zmiennych wraz z wyróżnionymi kategoriami referencyjnymi. Poziomy referencyjne zostały przyporządkowane kategoriom, które były dominantami.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Informacje o poziomach klasyfikacji** | | | |
| **Klasa** | **Etykieta** | **Wartość** | **Zmienne planu** | |
| gndr | Płeć | 1 | 0 |  |
|  | 2 | 1 |  |
| slprl | Częstotliwość problemów ze snem | 1 | 0 | 0 |
|  | 2 | 1 | 0 |
|  | 3 | 0 | 1 |
| eatveg | Częstotliwość jedzenia warzyw | 1 | 1 | 0 |
| 2 | 0 | 0 |
| 3 | 0 | 1 |
| alcfreq | Częstotliwość spożycia  alkoholu | 1 | 0 | 0 |
|  | 2 | 1 | 0 |
|  | 3 | 0 | 1 |
| cgtsmke | Częstotliwość palenia  papierosów | 1 | 1 | 0 |
|  | 2 | 0 | 1 |
|  | 3 | 0 | 0 |
| dosprt | Aktywność fizyczna w ciągu tygodnia | 1 | 0 | 0 |
|  | 2 | 1 | 0 |
|  | 3 | 0 | 1 |
| domicil | Miejsce  zamieszkania | 1 | 0 | 0 |
|  | 2 | 1 | 0 |
|  | 3 | 1 | 1 |

***Tabela 5. Schemat kodowania zmiennych*** *Opracowanie własne na podstawie oprogramowania SAS*

1. **Budowa modelu regresji logistycznej binarnej**

Poniżej zaprezentowana zostanie budowa modelu regresji logistycznej opisującego subiektywną ocenę stanu zdrowia obywateli Niemczech. Ponieważ zmienna przyjmuje dwie wartości został użyty model regresji binarnej. Model ten wyjaśnia jak kształtuje się pozytywna ocena stanu zdrowia w porównaniu do oceny negatywnej.

3.1 Budowa modelu

W pierwszym etapie budowy modelu przystąpiono do estymacji współczynników modelu. Aby otrzymać najwłaściwszą postać finalnego zestawu, wygenerowano parę modeli pośrednich oraz porównano otrzymane na ich podstawie wyniki. Najpierw poddano estymacji model ze wszystkimi zmiennymi, a następnie przy użyciu automatycznej metody selekcji zmiennych: selekcji krokowej *–* *stepwise selection*.

W modelu z krokową selekcją zmiennych zostały usunięte zmienne *domicil* (miejsce zamieszkania)*,* *eatveg* (jedzenie warzyw) oraz zmienna interakcji *slprl\*alcfreq*.

Tabela 6 zawiera statystyki opisujące model z wszystkimi zmiennymi. Wartości testów ilorazu wiarygodności, punktowego oraz testu Walda przy dwunastu stopniach swobody wyraźnie wskazują na odrzucenie hipotezy zerowej mówiącej o uznaniu wszystkich zmiennych modelu jako nieistotnych statystycznie przy 5% poziomie istotności. Natomiast kryteria dopasowania mniejsze dla modelu z współzmiennymi wskazują na lepszą dobroć dopasowania modelu do danych.

|  |  |  |  |
| --- | --- | --- | --- |
| Testowanie globalnej hipotezy zerowej: BETA=0 | | | |
| **Test** | **Chi-kwadrat** | **DF** | **Pr. > chi-kw.** |
| Iloraz wiarygod. | 466.2928 | 18 | <.0001 |
| Wynik punktowy | 437.5258 | 18 | <.0001 |
| Wald | 371.8052 | 18 | <.0001 |
| **Statystyki dopasowania** | | | |
| **Kryterium** | **Tylko wyraz wolny** | **Wyraz wolny i współzmienne** | |
| AIC | 4025.121 | 3598.530 | |
| SC | 4031.137 | 3676.738 | |
| -2 log L | 4023.121 | 3572.530 | |

***Tabela 6. Statystyki dla modelu pełnego***

Tabela 7. przedstawia oceny metody maksymalnej wiarygodności dla wszystkich kategorii poszczególnych parametrów. Większość z nich jest statystycznie istotna, gdyż przy 1 stopniu swobody wartość p jest mniejsza od 0.05. Nieistotne statystycznie zmienne to płeć (*gndr*) oraz miejsce zamieszkania (*domicyl*) dla obu kategorii. Na poziomie istotności 5% nie mamy podstaw do odrzucenia hipotezy zerowej mówiącej o nieistotności statystycznej tych zmiennych, tak wiec nie powinny się one znaleźć w ostatecznym modelu. Dla drugiej kategorii zmiennej cgtsmke, czyli częstotliwości palenia papierosów możemy stwierdzić, że na poziomie 5% zmienna jest istotna, natomiast jest nieistotna na poziomie istotności 1%. Jednak mimo to włączamy ją do modelu, gdyż kategoria 1 jest istotna przy wartości p mniejszej od 0.0001.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Analiza ocen maksymalnej wiarygodności | | | | | | | |
| **Parametr** |  |  | **DF** | **Ocena** | **Błąd standardowy** | **Chi-kwadrat Walda** | **Pr. > chi-kw.** |
| Intercept |  |  | 1 | 2.9815 | 0.2007 | 220.6841 | <.0001 |
| agea |  |  | 1 | -0.0323 | 0.00242 | 178.5620 | <.0001 |
| gndr | 2 |  | 1 | -0.1009 | 0.0855 | 1.3910 | 0.2382 |
| slprl | 2 |  | 1 | -0.4848 | 0.1254 | 14.9424 | 0.0001 |
| slprl | 3 |  | 1 | -1.3046 | 0.1741 | 56.1564 | <.0001 |
| eatveg | 1 |  | 1 | 0.2175 | 0.1176 | 3.4219 | 0.0643 |
| eatveg | 3 |  | 1 | -0.2230 | 0.0919 | 5.8906 | 0.0152 |
| alcfreq | 2 |  | 1 | -0.3671 | 0.1450 | 6.4066 | 0.0114 |
| alcfreq | 3 |  | 1 | -0.4664 | 0.2130 | 4.7943 | 0.0286 |
| cgtsmke | 1 |  | 1 | -0.8076 | 0.1065 | 57.5114 | <.0001 |
| cgtsmke | 2 |  | 1 | -0.1932 | 0.0974 | 3.9294 | 0.0474 |
| dosprt | 2 |  | 1 | 0.4451 | 0.0981 | 20.5960 | <.0001 |
| dosprt | 3 |  | 1 | 0.3070 | 0.0983 | 9.7467 | 0.0018 |
| domicil | 2 |  | 1 | -0.0788 | 0.1019 | 0.5985 | 0.4391 |
| domicil | 3 |  | 1 | -0.1533 | 0.0966 | 2.5190 | 0.1125 |
| slprl\*alcfreq | 2 | 2 | 1 | 0.1271 | 0.1933 | 0.4324 | 0.5108 |
| slprl\*alcfreq | 2 | 3 | 1 | -0.2456 | 0.3057 | 0.6451 | 0.4219 |
| slprl\*alcfreq | 3 | 2 | 1 | 0.0576 | 0.2446 | 0.0555 | 0.8137 |
| slprl\*alcfreq | 3 | 3 | 1 | -0.4361 | 0.3752 | 1.3505 | 0.2452 |

***Tabela 7. Analiza ocen maksymalnej wiarygodności dla modelu z wszystkimi zmiennymi***

Ze względu na nieistotność niektórych cech w modelu z wszystkimi zmiennymi zdecydowano się na stworzenie modelu z automatycznym doborem zmiennych za pomocą selekcji krokowej. Poniżej zaprezentowano porównanie obu modeli.

|  |  |  |
| --- | --- | --- |
| Statystyki dopasowania (model selekcji krokowej) | | |
| **Kryterium** | **Tylko wyraz wolny** | **Wyraz wolny i współzmienne** |
| AIC | 4025.121 | 3599.355 |
| SC | 4031.137 | 3659.515 |
| -2 log L | 4023.121 | 3579.355 |
| **Statystyki dopasowania (model z wszystkimi zmiennymi)** | | |
| **Kryterium** | **Tylko wyraz wolny** | **Wyraz wolny i współzmienne** |
| AIC | 4025.121 | 3598.530 |
| SC | 4031.137 | 3676.738 |
| -2 log L | 4023.121 | 3572.530 |

***Tabela 8. Porównanie kryterium dopasowania dla modeli***

Model, w którym zastosowano selekcję krokową nie jest znacznie lepszy od modelu z wszystkimi zmiennymi. Kryterium Akaike w obu przypadkach jest zbliżone, podobnie jak i kryterium podwojonego ilorazu funkcji wiarygodności. Natomiast kryterium Schwarza-Bayesa jest lepsze w przypadku modelu z selekcją, jednak tylko o 17.

3.2 Ocena jakości modelu

Kolejnym wykonanym krokiem jest ocena jakości modelu binarnego. Weryfikacja modelu obejmuje zbadanie istotności całego zestawu zmiennych objaśniających w modelu.

|  |  |  |  |
| --- | --- | --- | --- |
| Testowanie globalnej hipotezy zerowej: BETA=0 | | | |
| **Test** | **Chi-kwadrat** | **DF** | **Pr. > chi-kw.** |
| Iloraz wiarygod. | 443.7657 | 9 | <.0001 |
| Wynik punktowy | 418.4031 | 9 | <.0001 |
| Wald | 360.2354 | 9 | <.0001 |

***Tabela 9. Testy łącznej istotności parametrów modelu***

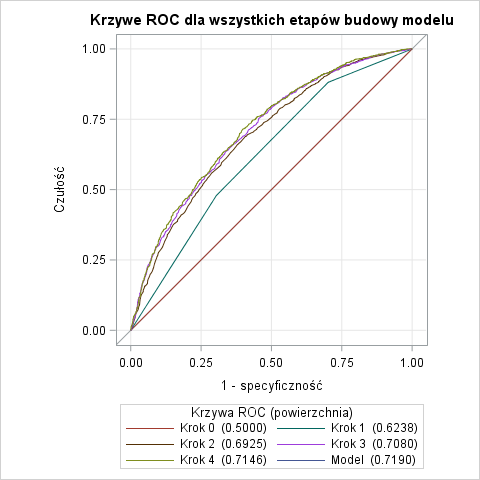
Na podstawie Tabeli 9. jesteśmy w stanie stwierdzić, że przy każdym poziomie istotności odrzucana jest hipoteza zerowa na korzyść hipotezy alternatywnej, mówiącej o istotności przynajmniej jednego z parametrów w modelu, co świadczy o poprawnej specyfikacji modelu.

Dla tego modelu współczynnik pseudo Coxa-Snella wynosi 0.1363, natomiast Negelkerke przyjmuje wartość 0.1854. W związku z tym, że wartości tej miary należy maksymalizować, aby otrzymać jak najlepszy model, możemy stwierdzić, iż dopasowanie modelu do danych jest raczej przeciętne.

Tabela 10 przedstawia statystyki obrazujące możliwości predykcyjne modelu. Wśród przeanalizowanych 3029 obserwacji poprawnie zakwalifikowanych zostało 71.9%, natomiast niepoprawnie 28.1%. Udało się zaklasyfikować wszystkie obserwacje. Wartość statystyki gamma dla modelu równa 0,438 oznacza, że znając zmienne niezależne redukujemy błąd oszacowania rangowania par o około 43,8%.

|  |  |  |  |
| --- | --- | --- | --- |
| Skojarzenie prognozowanych prawdopodobieństw i obserwowanych odpowiedzi | | | |
| Procent zgodnych | 71.9 | D Somersa | 0.438 |
| Procent niezgodnych | 28.1 | Gamma | 0.438 |
| Procent równych | 0.0 | Tau-a | 0.209 |
| Pary | 2187108 | c | 0.719 |

***Tabela 10. Statystyki mocy predykcyjnej modelu***

Statystyka D-Sommersa informuje nas o nadwyżce par poprawnie zakwalifikowanych w stosunku do błędnie sklasyfikowanych i w omawianym modelu wynosi ona 0.438. Statystyka C odpowiada polu pod krzywą ROC czyli opisuje jak dobrze model jest dopasowany do danych. W tym przypadku wynosi 71.9% więc możemy stwierdzić, iż w przypadku zagadnienia socjologicznego jakim jest subiektywna ocena zdrowia model jest dobrej jakości. Hipoteza zerowa testu Hosmera i Lemeshowa mówi, iż model jest dobrze dopasowany do danych. Na żadnym typowym poziomie istotności nie na podstaw do odrzucenia hipotezy zerowej dla obydwu modeli.

|  |  |  |
| --- | --- | --- |
| Test zgodności Hosmera i Lemeshowa | | |
| **Chi-kwadrat** | **DF** | **Pr. > chi-kw.** |
| 6.6788 | 8 | 0.5717 |

3.3. Interpretacja parametrów modelu

Poniżej przedstawione są oszacowania parametrów dla modelu oszacowanego dla Niemiec. Przykładowa interpretacja parametrów to:

* Wraz ze wzrostem wieku o 1 rok prawdopodobieństwo pozytywnej oceny stanu zdrowia spada o 3% *ceteris paribus*.
* Osoby, które w ciągu ostatniego tygodnia doświadczyły niespokojnego snu mają 133% większą szansę na negatywną ocenę stanu zdrowia w porównaniu do osób, które miały spokojny sen
* Palacze papierosów mają o 78% większą szansę na negatywną ocenę stanu zdrowia w porównaniu do osób niepalących
* Osoby aktywne fizycznie mają o 45% większą szansę na pozytywną ocenę stanu zdrowia w stosunku do osób nieaktywnych fizycznie.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Analiza ocen maksymalnej wiarygodności | | | | | | |
| **Parametr** |  | **DF** | **Ocena** | **Błąd standardowy** | **Chi-kwadrat Walda** | **Pr. > chi-kw.** |
| Intercept |  | 1 | 2.7967 | 0.1768 | 250.1598 | <.0001 |
| agea |  | 1 | -0.0321 | 0.00241 | 177.5549 | <.0001 |
| slprl | 2 | 1 | -0.4678 | 0.0903 | 26.8163 | <.0001 |
| slprl | 3 | 1 | -1.3378 | 0.1141 | 137.3859 | <.0001 |
| eatveg | 1 | 1 | 0.2044 | 0.1176 | 3.4219 | 0.0803 |
| eatveg | 3 | 1 | -0.2009 | 0.0910 | 5.8906 | 0.0272 |
| alcfreq | 2 | 1 | -0.3177 | 0.0885 | 12.8929 | 0.0003 |
| alcfreq | 3 | 1 | -0.6511 | 0.1359 | 22.9533 | <.0001 |
| cgtsmke | 1 | 1 | -0.7856 | 0.1053 | 55.6847 | <.0001 |
| cgtsmke | 2 | 1 | -0.1742 | 0.0963 | 3.2735 | 0.0704 |
| dosprt | 2 | 1 | 0.4593 | 0.0978 | 22.0680 | <.0001 |
| dosprt | 3 | 1 | 0.3116 | 0.0979 | 10.1236 | 0.0015 |

1. **Podsumowanie**

W części pierwszej wyszczególniliśmy postawione przed projektem hipotezy badawcze. Posiadając wyniki modelu regresji binarnej, może stwierdzić iź:

1. Jest prawdą, iż na wysoką ocenę stanu zdrowia mają czynniki pozytywne. Osoba bardzo aktywna fizycznie ocenia swój stan zdrowia o 31,16% lepiej niż osoba mało aktywna. Już niewielkie zmiany w aktywności fizycznej, czyli przejście z grupy mało aktywnej do aktywnej powoduje wzrost samooceny zdrowia o 45,93%. Aktywność fizyczna jest zmienną, która najlepiej wpływa na ogólną ocenę zdrowia.
2. Największy wpływ na wysoką ocenę stanu zdrowia mają czynniki pozytywne (aktywność fizyczna, jedzenie warzyw, sen). Zdaje się, iż ćwiczenia fizyczne, odpowiednia dieta i odpoczynek są czynnikami, które najbardziej powinny poprawiać ludzkie samopoczucie.

|  |  |  |
| --- | --- | --- |
| eatveg | Porządkowa (6 kategorii) | {1,2} – dwa razy dziennie, {3} – raz dziennie, {4,5,6} – mniej niż raz dziennie |

1. Osoby o wysokiej aktywności fizycznej najlepiej oceniają swój stan zdrowia. Jak wiadomo, ćwiczenia fizyczne są zalecane przez lekarzy dla ludzi w każdym wieku, czy posiadają one realny wpływ na wartości samooceny zdrowia?
2. Czy interakcja częstotliwości spożycia alkoholu i problemów ze snem ma istotny wpływ na negatywną ocenę stanu zdrowia?
3. Czy ludzie młodzi oceniają swój stan zdrowia lepiej niż ludzie starsi? Czy widać wyraźną zmianę oceny stanu zdrowia między poszczególnymi grupami wiekowymi?
4. Osoby mieszkające na wsi oceniają swój stan zdrowia lepiej niż osoby zamieszkujące miasta. Współcześnie zauważa się przepływ ludności ze wsi do miast, szczególnie ze względów ekonomicznych, czy jednak pozostanie na wsi wpływa znacząco na ocenę stanu zdrowia?

Załącznik 1. Kod SAS

libname essout 'C:\Users\rilld8\Documents\My SAS Files\9.4\RL';

libname out 'C:\Users\rilld8\Documents\My SAS Files\9.4\RL';

/\*Wczytanie zbioru danych dla Niemczech \*/

**data** Essout.DE\_Ess;

set Out.Ess7e02\_0\_f1(keep= idno --region dweight) ;

where cntry='DE';

**run**;

/\* Stworzenie interesującego nas zbioru danych z uwzględnieniem zmiennych objaśniających \*/

**data** Out.zbior;

set essout.de\_ess

(keep=health gndr agea domicil cgtsmke dosprt eatveg alcfreq slprl dweight);

where (health in (**1**:**5**)); /\* Wybór obserwacji dla których nie brakuje zmiennej celu - health \*/

**run**;

/\* Częstości dla zmiennej celu,

pomagają określić podział zbioru.

Otrzymujemy informacje, że 62% respondentów odpowiedziało,

że ich stan zdrowia jest dobry lub bardzo dobry \*/

**proc** **freq** data=out.zbior;

tables health;

weight dweight;

**run**;

/\* Przekodowanie zmiennej health na binarną,

interesujące nas zdarzenie to pozytywna ocena stanu zdrowia - 1 \*/

**data** out.zbior1;

set out.zbior;

if health<=**2** then health=**1**;

if health>=**3** then health=**0**;

**run**;

/\* Formatowanie zmiennej celu\*/

**proc** **format**;

value health\_f

**1**='Dobry'

**0**='Zły';

**run**;

**data** out.zbior1;

set out.zbior1;

format health health\_f.;

**run**;

**proc** **sort** data=out.zbior1;

by health;

**run**;

**proc** **freq** data=out.zbior1;

tables health;

weight dweight;

**run**;

/\* Przekodowanie zmiennych

Sprawdzenie częstości zmiennych, które posłużą do określenia podziału na poszczególne kategorie\*/

**proc** **freq** data=out.zbior1;

tables health gndr domicil cgtsmke dosprt eatveg alcfreq slprl;

by health;

weight dweight;

**run**;

**data** out.zbior2;

set out.zbior1;

if domicil <= **2** then domicil = **1**;

if domicil = **3** then domicil = **2**;

if domicil >= **4** then domicil = **3**;

if cgtsmke <=**2** then cgtsmke = **1** ;

if cgtsmke in (**3**:**4**) then cgtsmke = **2**;

if cgtsmke = **5** then cgtsmke = **3**;

if dosprt <= **2** then dosprt = **1**;

if dosprt in (**3**:**5**) then dosprt = **2**;

if dosprt >=**6** then dosprt = **3**;

if eatveg <= **2** then eatveg = **1**;

if eatveg = **3** then eatveg = **2**;

if eatveg >=**4** then eatveg=**3**;

if alcfreq <= **3** then alcfreq = **1**;

if alcfreq in (**4**:**6**) then alcfreq = **2**;

if alcfreq = **7** then alcfreq = **3**;

if slprl = **1** then slprl = **1**;

if slprl = **2** then slprl=**2**;

if slprl >= **3** then slprl=**3**;

**run**;

**proc** **format**;

value domicil\_f

**1**='Duże miasto'

**2**='Małe miasto'

**3**='Wieś';

value cgtsmke\_f

**1** = 'Palacz'

**2** = 'Były palacz'

**3** = 'Osoba niepaląca';

value dosprt\_f

**1** = 'Osoba mało aktywna'

**2** = 'Osoba aktywna'

**3** = 'Osoba bardzo aktywna';

value eatveg\_f

**1** = 'dwa razy dziennie'

**2** = 'raz dziennie'

**3** = 'mniej niż raz dziennie';

value alcfreq\_f

**1** = 'Często'

**2** = 'Czasami'

**3** = 'Nigdy';

value slprl\_f

**1** = 'Spokojny'

**2** = 'Raczej spokojny'

**3** = 'Niespokojny';

value gndr\_f

**1** = 'M'

**2** = 'K';

**run**;

**data** out.RL\_bin;

set out.zbior2;

format domicil domicil\_f.;

format cgtsmke cgtsmke\_f.;

format dosprt dosprt\_f.;

format eatveg eatveg\_f.;

format alcfreq alcfreq\_f.;

format slprl slprl\_f.;

format gndr gndr\_f.;

**run**;

**proc** **freq** data=out.rl\_bin;

table health;

weight dweight;

title "Rozkład zmiennej 'health'";

**run**;

/\* Kod dla wykresów rozkładu zmiennych objaśniających względem zmiennej 'health' \*/

PATTERN1 COLOR=RED;

PATTERN2 COLOR=GREEN;

Legend1

FRAME

POSITION = (BOTTOM CENTER OUTSIDE)

LABEL=( "Subiektywny stan zdrowia");

Axis1

STYLE=**1**

WIDTH=**1**

MINOR=NONE;

Axis2

STYLE=**1**

WIDTH=**1**

LABEL=( FONT='Arial /b' "Rozkład aktywności fizycznej");

TITLE;

TITLE1 "Rozkład zmiennej dosprt względem zmiennej health";

**PROC** **GCHART** DATA=out.RL\_BIN;

VBAR dosprt / SUBGROUP=health

CLIPREF

FRAME DISCRETE

TYPE=FREQ

INSIDE=PCT

LEGEND=LEGEND1

COUTLINE=BLACK

RAXIS=AXIS1

MAXIS=AXIS2;

**RUN**;

**QUIT**;

/\*Procedura PROC LOGISTIC użyta do stworzenia modelu regresji logistycznej binarnej dla wszystkich zmiennych\*/

**PROC** **LOGISTIC** DATA=out.rl\_bin

PLOTS(ONLY)=ODDSRATIO

PLOTS(ONLY)=ROC;

CLASS gndr (PARAM=REF REF='1') slprl (PARAM=REF REF='1') eatveg (PARAM=REF REF='2') alcfreq (PARAM=REF REF='1') cgtsmke (PARAM=REF REF='3') dosprt (PARAM=REF REF='1') domicil (PARAM=ORDINAL);

WEIGHT dweight;

MODEL health (Event = '1')=agea gndr slprl eatveg alcfreq cgtsmke dosprt domicil slprl\*alcfreq /

SELECTION=NONE

SLE=**0.05**

SLS=**0.05**

INCLUDE=**0**

CORRB

COVB

INFLUENCE

LACKFIT

AGGREGATE SCALE=NONE

RSQUARE

LINK=LOGIT

CLPARM=BOTH

CLODDS=BOTH

ALPHA=**0.05**

;

OUTPUT OUT=out.rl\_bin\_pred(LABEL="Statystyki i prognozy regresji logistycznej")

PREDPROBS=INDIVIDUAL

RESCHI=reschi\_health

RESDEV=resdev\_health

DIFCHISQ=difchisq\_health

DIFDEV=difdev\_health ;

**RUN**;

**QUIT**;

/\*Procedura PROC LOGISTIC użyta do stworzenia modelu regresji logistycznej binarnej z krokowym wyborem zmiennych\*/

**PROC** **LOGISTIC** DATA=out.rl\_bin

PLOTS(ONLY)=ODDSRATIO

PLOTS(ONLY)=ROC;

CLASS gndr (PARAM=REF REF='1') slprl (PARAM=REF REF='1') eatveg (PARAM=REF REF='2') alcfreq (PARAM=REF REF='1') cgtsmke (PARAM=REF REF='3') dosprt (PARAM=REF REF='1') domicil (PARAM=ORDINAL);

WEIGHT dweight;

MODEL health (Event = '1')=agea gndr slprl eatveg alcfreq cgtsmke dosprt domicil slprl\*alcfreq /

SELECTION=STEPWISE

SLE=**0.05**

SLS=**0.05**

INCLUDE=**0**

CORRB

COVB

INFLUENCE

LACKFIT

AGGREGATE SCALE=NONE

RSQUARE

LINK=LOGIT

CLPARM=BOTH

CLODDS=BOTH

ALPHA=**0.05**

;

OUTPUT OUT=out.rl\_bin\_pred(LABEL="Statystyki i prognozy regresji logistycznej")

PREDPROBS=INDIVIDUAL

RESCHI=reschi\_health

RESDEV=resdev\_health

DIFCHISQ=difchisq\_health

DIFDEV=difdev\_health ;

**RUN**;

**QUIT**;

1. http://data.worldbank.org/indicator/SP.DYN.IMRT.IN?order=wbapi\_data\_value\_2011+wbapi\_data\_value+wbapi\_data\_value-last&sort=asc [↑](#footnote-ref-1)