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, -, -

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: |

Spis treści:

1. Cel i zakres analizy
2. Prezentacja zbioru i przygotowanie danych
3. Model regresji logistycznej – binarnej
4. **Cel i zakres analizy**

Celem niniejszej pracy jest analiza czynników wpływających na subiektywną ocenę stanu zdrowia wśród obywateli Niemczech. Społeczeństwo niemieckie jest świetnym przykładem reprezentującym kraj wysoko rozwinięty. 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 obserwujemy również wzrost otyłości wśród obywateli krajów wysokorozwiniętych, ze względu na powszechny dostęp do żywności wysoko przetworzonej, towarów luksusowych oraz używek – takich jak alkohol.

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.

***Tabela 1. Opis zmiennych wykorzystywanych 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 |

*Opracowanie własne na podstawie :* <http://nesstar.ess.nsd.uib.no/webview/>

Zmienną celu przyjętą *health* czyli poziom subiektywnego stanu zdrowia. Jest to zmienna porządkowa przyjmująca 5 kategorii, zaprezentowana w poniższej tabeli.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | |  | | --- | |  | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | ***Tabela 2 Rozkład zmiennej 'health'***   | **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 |   *Opracowanie własne na podstawie oprogramowania SAS* | | | |

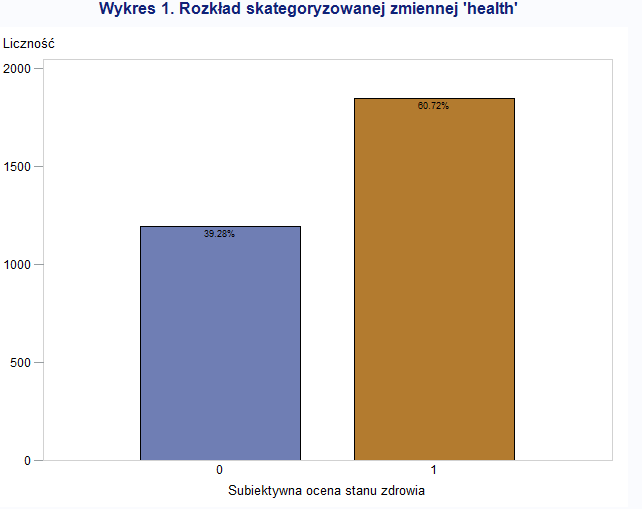
2.2 Przygotowanie zbioru danych

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.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Tabela 3 Rozkład skategoryzowanej zmiennej 'health'***   | **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 | |

*Opracowanie własne na podstawie oprogramowania SAS*

|  |
| --- |
|  |

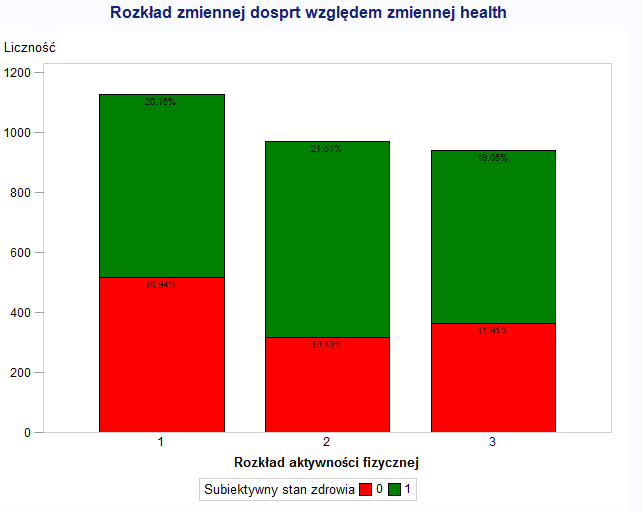
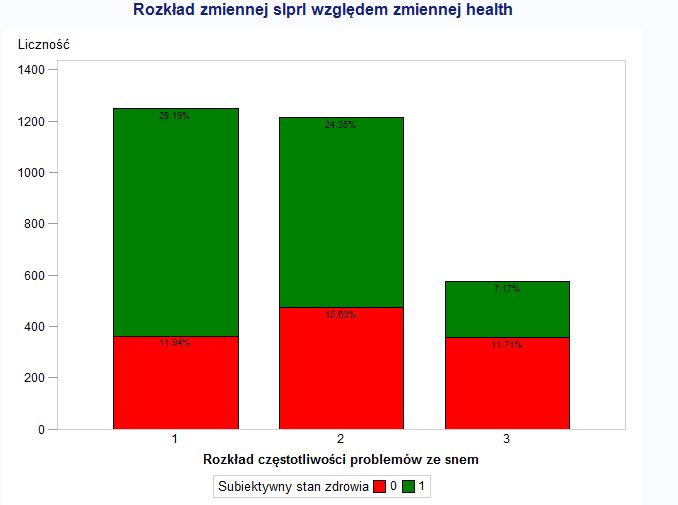


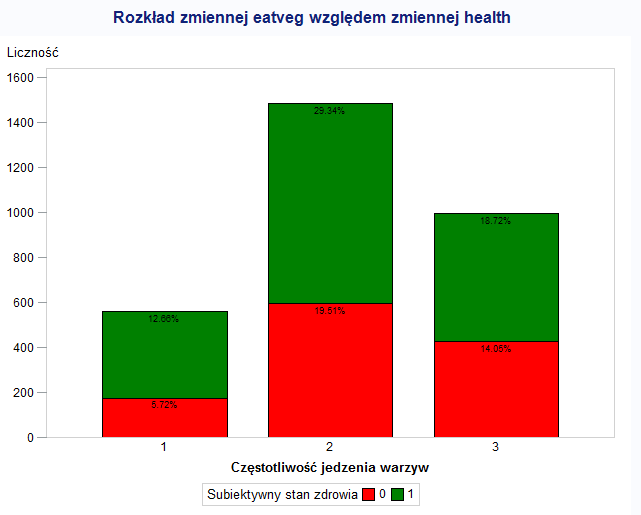
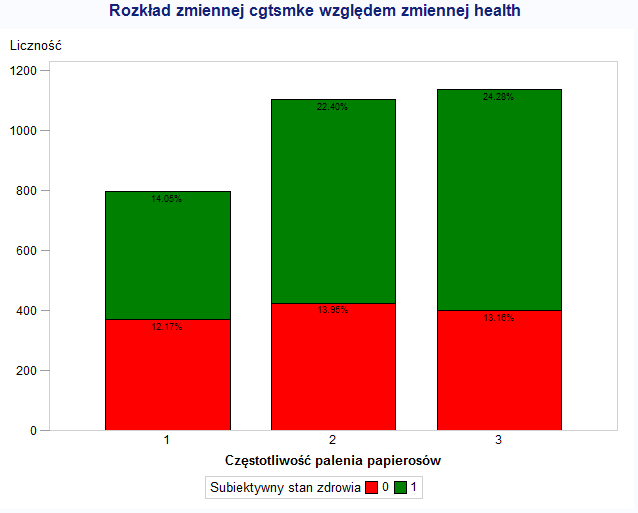
Kategoryzacji poddano również zmienne objaśniające, głównie ze względu na niewielkie liczności niektórych kategorii oraz możliwość merytorycznego pogrupowania większości z nich. Zmienne porządkowe zostały podzielone na 3 kategorie, według schematu prezentowanego w Tabeli 4. Rozkłady poszczególnych zmiennych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

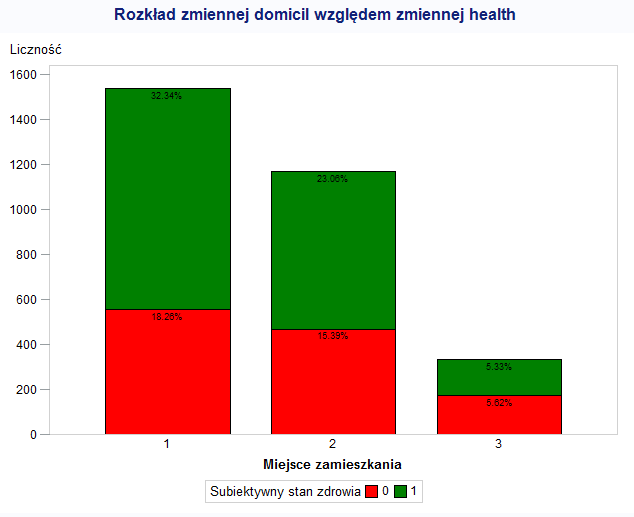
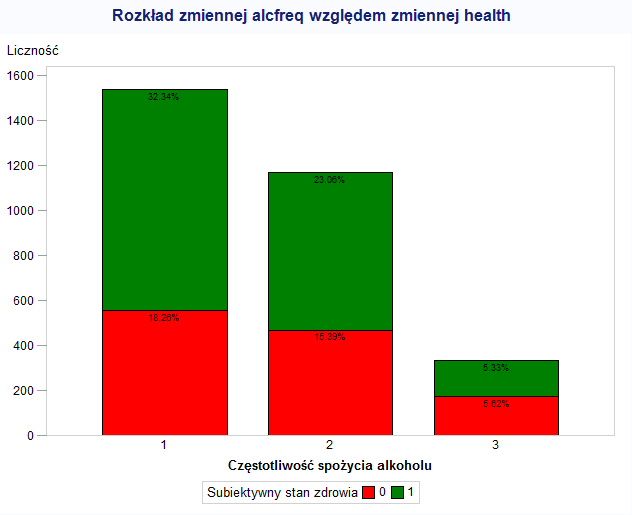
***Tabela 4. Podział zmiennych porządkowych***

|  |  |  |
| --- | --- | --- |
| **Zmienna** | **Typ i początkowe kategorie** | **Podział** |
| *health* | Porządkowa (5 kategorii) | {1,2} – pozytywna, {3,4,5} - negatywna |
| *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 |

*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.

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.

***Tabela 5. Schemat kodowania zmiennych***

|  | **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 |
| 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 |

*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*.

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.

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

| **Testowanie globalnej hipotezy zerowej: BETA=0** | | | |
| --- | --- | --- | --- |
| **Test** | **Chi-kwadrat** | **DF** | **Pr. > chi-kw.** |
| Iloraz wiarygod. | 450.5911 | 12 | <.0001 |
| Wynik punktowy | 424.1886 | 12 | <.0001 |
| Wald | 364.2368 | 12 | <.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 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.

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

| **Analiza ocen maksymalnej wiarygodności** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Parametr** | **Kat.** | **DF** | **Ocena** | **Błąd standardowy** | **Chi-kwadrat Walda** | **Pr. > chi-kw.** |
| Intercept |  | 1 | 2.9727 | 0.1945 | 233.6917 | <.0001 |
| agea |  | 1 | -0.0322 | 0.00241 | 178.2069 | <.0001 |
| gndr | 2 | 1 | -0.1017 | 0.0855 | 1.4143 | 0.2343 |
| slprl | 2 | 1 | -0.4654 | 0.0907 | 26.3361 | <.0001 |
| slprl | 3 | 1 | -1.3378 | 0.1153 | 134.6303 | <.0001 |
| alcfreq | 2 | 1 | -0.2966 | 0.0912 | 10.5746 | 0.0011 |
| alcfreq | 3 | 1 | -0.6445 | 0.1372 | 22.0533 | <.0001 |
| cgtsmke | 1 | 1 | -0.8121 | 0.1064 | 58.2765 | <.0001 |
| cgtsmke | 2 | 1 | -0.1948 | 0.0974 | 4.0047 | 0.0454 |
| dosprt | 2 | 1 | 0.4505 | 0.0980 | 21.1491 | <.0001 |
| dosprt | 3 | 1 | 0.3157 | 0.0981 | 10.3634 | 0.0013 |
| domicil | 2 | 1 | -0.0830 | 0.1017 | 0.6658 | 0.4145 |
| domicil | 3 | 1 | -0.1519 | 0.0965 | 2.4773 | 0.1155 |

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 tych dwóch modeli.

***Tabela 8. Porównanie kryterium dopasowania dla 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 |

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.

***Tabela 9. Testy łącznej istotności parametrów 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 |

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%.

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

| **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 |

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.

3.3 Interpretacja parametrów modelu