Uniwersytet Wrocławski
Wydział Matematyki i Informatyki
Instytut Matematyczny
Analiza danych ***Maciej Kaczyński***

Analiza danych medycznych w oparciu o wielomianową regresję logistyczną

Praca licencjacka
napisana pod kierunkiem
dr Michała Kosa

Wrocław 2022

Spis treści

1 Wstęp.

Niniejsza praca dotyczy tematyki wielomianowej regresji logistycznej, a w szczególności praktycznego zastosowania tego narzędzia statystycznego do analizy danych pochodzących z badania medycznego. Naczelnym celem jest zbudowanie modelu, który umożliwi predykcję stanu pacjenta po operacji na podstawie informacji tj. wiek, płeć pacjenta, występujące symptomy i tempo ich wystąpienia czy technika zastosowana przy zabiegu. Z pomocą tego modelu wysunięte zostaną wnioski, które mają służyć jako wsparcie specjalistów medycznych, a także pewne ukierunkowanie do dalszych badań i analiz statystycznych. Z postawionego celu wynika struktura pracy: pierwsza część opisuje teoretyczne podstawy wielomianowej regresji logistycznej i stosowanych metod statystycznych, druga zawiera właściwą analizę danych, czyli proces wybrania i zbadania optymalnego modelu, a na końcu pracy znajdują się wysunięte wnioski, w tym również tablice predykcji mające na celu wspomagać lekarzy i pacjentów przy podejmowaniu decyzji o operacji.

2 Część teoretyczna.

Pierwszy rozdział poświęcony zostanie oprawie teoretycznej zagadnienia, będącego tematem niniejszej pracy. Rozpoczniemy od ogólnego opisania pojęcia regresji i przybliżymy jej najbardziej podstawowe rodzaje, a następnie zbierzemy najważniejsze informacje na temat wielomianowej regresji logistycznej. Na koniec przedstawimy wybrane metody wyboru i oceny modeli regresji, które wykorzystane zostaną w następnym rozdziale, poświęconym analizie danych.

2.1 Wprowadzenie do regresji.

Regresja w sensie statystycznym odnosi się do opisywania zależności pomiędzy zmiennymi. W szczególności pozwala modelować tzw. zmienną wynikową (objaśnianą/zależną) z pomocą zmiennych niezależnych, nazywanych też zmiennymi objaśniającymi regresorami lub predyktorami. Regresja jest niezwykle użytecznym narzędziem statystycznym, ponieważ pozwala przewidywać przyszłe zjawiska na podstawie obecnie posiadanych danych (tzw. predykcja), a także powiększa nasze zrozumienie analizowanych zmiennych, poprzez wskazywanie regresorów, które mają na nie wpływ wraz z jego siłą. Zobrazować to możemy z pomocą tzw. Uogólnionych Modeli Liniowych (z ang. *Generalized Linear Models*; *GLM*), które stanowią grupę najprostszych i najpopularniejszych modeli regresji. Klasa GLM oparta jest na następujących założeniach:

1. Dysponujemy zbiorem n obserwacji $(y_i, x_{i1}, ..., x_{ip-1})_{i=1}^{n}$, przy czym zmienne $(x_{i1}, ..., x_{ip-1})$ to regresory, a y_i jest zmienną wynikową.

2. $y_1, ..., y_n$ to niezależne realizacje zmiennych losowych $Y_1, ..., Y_n$, pochodzących z tego samego rozkładu eksponencjalnego o gęstości/funkcji rozkładu prawdopodobieństwa postaci:

$$f(y_i; \theta_i, \phi_i) = c(y_i; \phi) \exp \frac{\theta_i y_i - a(\theta_i)}{\phi}$$

   gdzie θ_i oraz $\phi > 0$ są parametrami, a $c()$ i $a()$ są znanymi funkcjami i $a()$ ma ciągłą drugą pochodną.

3. Dla każdego i związek pomiędzy wartością oczekiwaną $E(Y_i)$ a $(x_{i1}, ..., x_{ip-1})$ ma postać:

$$g(E(Y_i)) = \eta_i(\beta) = \beta_0 + x_{i1}\beta_1 + ... + X_{ip-1}\beta_{p-1}$$

   gdzie:

   - $(\beta_0, \beta_1, ..., \beta_{p-1})' \in \mathbb{R}^p$ to wektor parametrów powiązanych z regresorami, przy czym parametr β_0 to tzw. *intercept* i nie stoi przy żadnym regresorze,
   - $\eta_i(\beta)$ to tzw. predyktor liniowy, będący kombinacją liniową regresorów oraz wektora β,
   - $g()$ jest tzw. funkcją linkującą/łączącą (z ang. *link function*).

Ostatnie założenie można przedstawić ogólniej w postaci macierzowej:

$$g(E(Y)) = \eta(\beta) = \mathbb{X}_i\beta$$

gdzie

- β jest wektorem parametrów,
- $\mathbb{X}$ jest macierzą planu o rozmiarze $n \times p$,
- Y jest wektorem losowym zmiennych wynikowych.

Do klasy GLM należy m. in. regresja logistyczna, która służy do modelowania zmiennych binarnych i w związku z tym jest niezwykle przydatna przy danych medycznych. Jej kanoniczną funkcją linkującą jest logit, czyli kwantyl rozkładu logistycznego. Stąd modele regresji logistycznej mają następującą postać[1]:

$$\log \frac{\pi_i}{1-\pi_i} = \mathbb{X}_i\beta$$

Symbolem π_i oznaczamy wartość oczekiwaną i-tego elementu wektora zmiennej objaśnianej, co jest popularną konwencją związaną z faktem, że zmienna ta ma rozkład zero-jedynkowy. Z tego wynika interpretacja wartości oczekiwanej zmiennej wynikowej jako prawdopodobieństwa zajścia zdarzenia, które jest kodowane tą zmienną ($EY_i = \pi_i = P(Y_i = 1)$). Widzimy zatem, że w regresji logistycznej predyktor liniowy jest równy logarytmowi szans (z ang. *log odds*) zmiennej Y. Oczywiście funkcję logit możemy odwrócić, otrzymując w ten sposób bezpośredni wzór na π_i:

$$\pi_i = P(Y_i = 1) = \frac{e^{\mathbb{X}_i\beta}}{e^{\mathbb{X}_i\beta}+1}.$$

2.2 Wielomianowa regresja logistyczna.

Poznaliśmy już regresję logistyczną, która pozwala nam modelować zmienne jakościowe o dwóch możliwych stanach. Co jednak ze zmienną jakościową, która ma więcej stanów? W celu radzenia sobie z takimi danymi stworzona została regresja wielomianowa logistyczna, którą możemy postrzegać jako rozszerzenie, a tak właściwie po prostu uogólnienie regresji logistycznej. Zauważmy bowiem, że modele regresji logistycznej możemy traktować jako układ dwóch równań:

- $P(Y_i = 1) = \frac{e^{\mathbb{X}_i\beta}}{e^{\mathbb{X}_i\beta}+1}$,
- $P(Y_i = 0) = 1 - P(Y_i = 1) = \frac{1}{e^{\mathbb{X}_i\beta}+1}$.

Istnienie dwóch możliwych stanów zmiennej wynikowej oznaczało, że wystarczyło nam jedno równanie (drugie wynikało bezpośrednio z pierwszego) i tylko jedna wartość oczekiwana π będąca parametrem rozkładu zero-jedynkowego. Modele wielomianowej regresji logistycznej są budowane analogicznie, ale przy m możliwych stanów potrzebujemy m-1 równań, każde modelujące inny parametr rozkładu wielomianowego. Od tego momentu będziemy rozważać najprostszy przykład, kiedy zmienna objaśniania ma 3 stany (oznaczane jako wartości 1, 2 lub 3), z czym spotkamy się w rozdziale poświęconym analizie danych (uogólnienie modelu nie jest trudne, ale przy większej liczbie stanów bardziej problematyczna staje się sama notacja). Stąd dostajemy teoretyczny model wielomianowej regresji logistycznej[2]:

- $\pi_3(\mathbb{X}_i) = P(Y_i = 3) = \frac{e^{\theta_i}}{e^{\eta_i}+e^{\theta_i}+1} = \frac{e^{\mathbb{X}_i\beta}}{e^{\mathbb{X}_i\alpha}+e^{\mathbb{X}_i\beta}+1}$,
- $\pi_2(\mathbb{X}_i) = P(Y_i = 2) = \frac{e^{\eta_i}}{e^{\eta_i}+e^{\theta_i}+1} = \frac{e^{\mathbb{X}_i\alpha}}{e^{\mathbb{X}_i\alpha}+e^{\mathbb{X}_i\beta}+1}$,

- $\pi_1(\mathbb{X}_i) = P(Y_i = 1) = \frac{1}{e^{\eta_i}+e^{\theta_i}+1} = \frac{1}{e^{\mathbb{X}_i\alpha}+e^{\mathbb{X}_i\beta}+1}$.

Widzimy, że nasz model posiada dwa wektory parametrów α i β, a co za tym idzie dwa predyktory liniowe η i θ. Każdy powiązany jest z innym równaniem. Podobnie jak przy regresji logistycznej, możemy pozbyć się jednego równania, które dostajemy z dwóch pozostałych. Wiąże się to z wybraniem bazowego stanu zmiennej objaśnianej, którym w opisywanym przez nas przypadku jest wartość 1. Możemy również przedstawić model w postaci podobnej do tej, która charakteryzuje modele klasy *GLM*:

- $\log \frac{P(Y_i=3)}{P(Y_i=1)} = \theta_i = \mathbb{X}_i\beta$,
- $\log \frac{P(Y_i=2)}{P(Y_i=1)} = \eta_i = \mathbb{X}_i\alpha$.

Znamy już schemat modelu regresji logistycznej wielomianowej i wiemy, jak możemy z jego pomocą szacować prawdopodobieństwa przyjęcia danego stanu przez zmienną wynikową. Występuje jednak pewien problem, bowiem potrzebujemy nie tylko wektora zmiennej objaśnianej Y i macierzy planu $\mathbb{X}$, ale również macierzy parametrów $\{\alpha\beta\}$ o rozmiarze $p \times 2$:

$$\begin{bmatrix} \alpha_0 & \beta_0 \\ \alpha_1 & \beta_1 \\ . & . \\ . & . \\ . & . \\ \alpha_{p-1} & \beta_{p-1} \end{bmatrix}$$

Stąd bierze się potrzeba ich estymacji, którą powszechnie wykonuje się metodą największej wiarogodności. Do tego przydatne jest utworzenie trzech pomocniczych zmiennych indykatorowych y_1, y_2, y_3, które są zakodowane w następujący sposób:

- $Y_i = 1 \Longrightarrow y_{1i} = 1$, $y_{2i} = 0$ i $y_{3i} = 0$,
- $Y_i = 2 \Longrightarrow y_{1i} = 0$, $y_{2i} = 1$ i $y_{3i} = 0$,
- $Y_i = 3 \Longrightarrow y_{1i} = 0$, $y_{2i} = 0$ i $y_{3i} = 1$.

Tym sposobem otrzymujemy następującą postać funkcji wiarogodności, dla której szukamy argumentu maksymalizującego:

$$L(a, b) = \prod_{i=1}^{n} [\pi_1(\mathbb{X}_i)^{y_{1i}} \cdot \pi_2(\mathbb{X}_i)^{y_{2i}} \cdot \pi_3(\mathbb{X}_i)^{y_{3i}}]$$

Łatwiej jest prowadzić obliczenia na funkcji log-wiarogodności:

$$l(a, b) = \sum_{i=1}^{n} [y_{2i} \cdot \mathbb{X}_i a + y_{3i} \cdot \mathbb{X}_i b - \log\left(1 + e^{\mathbb{X}_i a} + e^{\mathbb{X}_i b}\right)]$$

Do znalezienia maksimum potrzebujemy pochodnych cząstkowych funkcji log-wiarogodności po każdym spośród $2(p + 1)$ parametrów, a także wszystkich istniejących drugich pochodnych cząstkowych. Wynika to z faktu, że ekstremum funkcji znajdujemy poprzez odnalezienie miejsca zerowego gradientu funkcji log-wiarogodności, tymczasem jej hesjan musi być ujemnie określony, żeby było to maksimum. Liczymy te pochodne zgodnie z następującymi wzorami[3]:

- $\frac{\partial l(a,b)}{\partial a_k} = \sum_{i=1}^{n}[x_{ki} \cdot (y_{2i} - \pi_{2i})]$,
- $\frac{\partial l(a,b)}{\partial b_k} = \sum_{i=1}^{n}[x_{ki} \cdot (y_{3i} - \pi_{3i})]$,
- $\frac{\partial^2 l(a,b)}{\partial a_k \partial a_l} = -\sum_{i=1}^{n}[x_{li} \cdot x_{ki} \cdot \pi_{2i}(1 - \pi_{2i})]$,
- $\frac{\partial^2 l(a,b)}{\partial b_k \partial b_l} = -\sum_{i=1}^{n}[x_{li} \cdot x_{ki} \cdot \pi_{3i}(1 - \pi_{3i})]$,
- $\frac{\partial^2 l(a,b)}{\partial a_k \partial b_l} = \frac{\partial^2 l(a,b)}{\partial b_k \partial a_l} = \sum_{i=1}^{n}[x_{li} \cdot x_{ki} \cdot \pi_{2i} \cdot \pi_{3i}]$.

Niestety, nie jesteśmy w stanie wyprowadzić jawnego wzoru na estymatory parametrów modelu. Zamiast tego korzystamy z metod numerycznych, tj. metoda *Gradient Descent* czy metoda *Newtona*. Otrzymany z ich pomocą estymator, będący estymatorem największej wiarogodności, jest asymptotycznie normalny z następującymi parametrami rozkładu:

$$\{ab\} \ \rightarrow^n N\left(\{\alpha\beta\}, \begin{bmatrix} [\mathbb{X}'V_2\mathbb{X}] & -[\mathbb{X}'V_1\mathbb{X}] \\ -[\mathbb{X}'V_1\mathbb{X}] & [\mathbb{X}'V_3\mathbb{X}] \end{bmatrix}^{-1}\right)$$

gdzie V_1, V_2 i V_3 są macierzami diagonalnymi rozmiaru $n \times n$, których wyrazy stojące na głównej przekątnej to odpowiednio $\pi_{2i}\pi_{3i}$, $\pi_{2i}(1 - \pi_{2i})$ i $\pi_{3i}(1 - \pi_{3i})$ dla i=1, ... , n.

2.3 Metody wyboru regresorów.

Wiemy już w jaki sposób jest zbudowany model wielomianowej regresji logistycznej i rozumiemy, jak wygląda estymacja jego parametrów. Pozostaje jednak kwestia oceny naszego modelu. Jest to bowiem oczywiste, że do modelowania danej zmiennej wynikowej możemy dobrać wiele różnych kombinacji regresorów spośród zbioru posiadanych przez nas zmiennych, a każda taka kombinacja oznacza inny model do zbudowania. Wybór możliwych regresorów może następować już na poziomie planowania eksperymentu/badania. Alternatywnie, statystyk może otrzymać duży zbiór zmiennych, które badacze zaobserwowali i jego odpowiedzialnością będzie znalezienie tych, które mają istotny wpływ na zmienną objaśnianą (co już samoistnie prowadzi do lepszego zrozumienia badanego zjawiska).

2.3.1 Statystyka *Deviance* i kryteria informacyjne.

Pierwszym narzędziem oceny modelu, jakie opiszemy, jest tzw. statystyka *Deviance*. Pozwala ona oceniać dopasowanie modelu do danych, poprzez porównanie go z tzw. modelem nasyconym (z ang. *saturated model*). Jest to model wzorcowy, raczej nie spotykany w praktyce, w którym liczba parametrów jest równa liczbie obserwacji. Oznacza to, że dla modelu nasyconego zachodzi równość $EY = Y$, czyli jego predykcje są całkowicie prawidłowe. Statystyka *Deviance* ma następującą postać[4]:

$$Deviance(M) = 2 \cdot (l(a,b)^S - l(a,b)^M)$$

gdzie:

- M oznacza badany model,

- $l(a,b)^{\mathrm{S}}$ to maksimum funkcji log-wiarogodności modelu nasyconego,
- $l(a,b)^{\mathrm{M}}$ to maksimum funkcji log-wiarogodności badanego modelu.

Statystyka *Deviance* ma swoje zastosowania, ale jest ograniczona przez jedną poważną wadę: w naturalny sposób preferuje bardziej złożone modele, czyli dokładanie regresorów do modelu (nawet nieistotnych z perspektywy predykcji zmiennej wynikowej), powiększa wartość tej statystyki. Odpowiedzią na ten problem są kryteria informacyjne, tj. *AIC* (z ang. *Akaike Information Criterion*)[5] oraz *BIC* (z ang. *Bayesian Information Criterion*)[6]. Kryteria te, podobnie jak statystyka *Deviance*, są zbudowane w oparciu o funkcję log-wiarogodności, ale zawierają również „karę" za zwiększanie liczby regresorów:

- $AIC(M) = -2 \cdot \frac{l(a,b)^{\mathrm{M}}}{n} + \frac{2 \cdot p^M}{n}$,
- $BIC(M) = -2 \cdot \frac{l(a,b)^{\mathrm{M}}}{n} + \frac{\log(n) \cdot p^M}{n}$.

Gdzie:

- n to liczba obserwacji,
- p^M to liczba parametrów badanego modelu.

Widzimy zatem, że w przeciwieństwie do statystyki *Deviance*, kryteria informacyjne osiągają wartości niższe dla lepszych modeli i składają się z dwóch elementów: pierwszy jest związany z log-wiarogodnością modelu i ocenia jego dopasowanie do danych, a drugi jest oparty na liczbie parametrów i ocenia prostotę modelu, osiągając wyższe (i w związku z tym gorsze) wartości przy niepotrzebnym dodawaniu regresorów. Główną różnicą pomiędzy kryteriami jest fakt, że *BIC* stosuje bardziej surową karę za zwiększanie liczby predyktorów niż *AIC*, co skutkuje wybraniem mniejszej liczby zmiennych niezależnych do modelu. Warto zauważyć, że budowa kryteriów informacyjnych dobrze reprezentuje podstawową ideę wyboru modelu, czyli balansowanie dopasowania do danych z prostotą modelu.

Zanim przejdziemy do analizy danych, należy poruszyć jeszcze jedną kwestię. Przedstawione dotychczas statystyki służą nam do oceniania modeli, umożliwiając wybranie najlepszego z nich. Potrzebujemy jednak metod, które pozwolą nam w optymalny sposób przeprowadzić takie porównania modeli, bowiem już dla niewielkiej liczby potencjalnych regresorów istnieje olbrzymia liczba możliwych kombinacji i nie jesteśmy w stanie zbudować modelu dla każdej takiej kombinacji w celu ich porównania. Przykładem takiej metody jest tzw. selekcja typu *Forward*[7], którą zastosujemy w części analitycznej. Metoda ta ma postać algorytmu postępowania, w którym zaczynamy od zbudowania modelu regresji z samym *interceptem* i jego oceny z pomocą wybranego kryterium (np. *AIC* lub *BIC*). W następnym kroku budujemy p modeli zawierających po jednym regresorze i porównujemy z pomocą wybranego kryterium. Pozostajemy przy najlepszym z modeli i przechodzimy do kolejnego kroku, w którym tworzymy p-1 modeli, każdy będący rozszerzeniem poprzednio wybranego modelu o jeden z pozostałych regresorów. Wykonujemy analogicznie kroki aż do znalezienia modelu, dla którego wybrane kryterium osiągnęło po raz pierwszy lokalne minimum. W ten sposób istotnie przyspieszamy proces porównywania potencjalnych modeli w poszukiwaniu najlepszego.

2.3.2 Metoda Lasso.

W trakcie analizy danych przeprowadzonej w następnym rozdziale, wykorzystamy jeszcze jedno narzędzie statystyczne- tzw. metodę *Lasso*[8] (*Least absolute shrinkage and selection operator*). Jest to bardzo kompleksowa metoda, która prowadzi do uzyskania dobrego kandydata na model na podstawie całego zbioru danych. Polega na następującej idei: znalezienia jak „najmniejszych" wektorów parametrów ($||\alpha||_1 \to \min \ \wedge ||\beta||_1 \to \min$), dla których funkcja log-wiarogodności przyjmuje dostatecznie wysoką wartość ($l(a,b) \leqslant \delta$ dla pewnej stałej δ). Równoważnie możemy wyrazić to w następujący sposób:

$$\min_{a,b \in \mathbb{R}^p} (-l(a,b) + \lambda \sum_{i=1}^{p-1} (|a_i| + |b_i|))$$

Uzyskane w ten sposób estymacje parametrów α i β są skonstruowane w taki sposób, że część parametrów zostaje zupełnie wyzerowana, a pozostałe ulegają zbliżeniu do zera. W związku z tym sugerują nam które regresory należy włączyć do naszego modelu, a które należy uznać za nieistotne.

W praktycznym zastosowaniu metody *Lasso* kluczowym zagadnieniem jest wybór optymalnej wartości λ. Zwiększenie λ prowadzi do otrzymania estymatora parametrów o większej liczbie zer, co ułatwia interpretację i pozwala na zbudowanie prostszego modelu. Tymczasem niższa λ zapewnia estymator, który jest lepiej dopasowany do danych i dokładniej estymuje wektor parametrów. Bardzo popularnym i często stosowanym wraz z metodą *Lasso* narzędziem jest tzw. sprawdzian krzyżowy (z ang. *cross-validation*)[9]. Jest to algorytm postępowania, w którym dzielimy zbiór danych na K grup i dla każdego $i \in \{1, ..., K\}$ wykonujemy następujące operacje:[10]

- i-ta grupa zostaje przypisana jako zbiór testowy, a pozostałe K-1 grup jako zbiór treningowy.
- Na zbiorze treningowym zostaje zastosowana metoda *Lasso* t razy, dla t różnych wartości λ z pewnego ustalonego zakresu $\{\lambda_1, ..., \lambda_t\}$.
- Otrzymujemy w ten sposób t modeli, z pomocą których przeprowadzamy predykcję danych ze zbioru testowego.
- Dla każdej wartości λ_j; $j = 1, ..., t$ liczymy średni kwadratowy błąd predykcji.

Po zastosowaniu tego algorytmu otrzymujemy wartość $\lambda \in \{\lambda_1, ..., \lambda_t\}$, która jest najbardziej optymalna z perspektywy średniego kwadratowego błędu predykcji. Należy zaznaczyć, że stosowane są różne typy sprawdzianu krzyżowego o różnej wartości K. W następnej części wykorzystane zostanie tzw. *Leave-one-out Cross-validation*, w którym K równe jest liczbie obserwacji, zatem zbiory testowe składają się z dokładniej jednej obserwacji.

3 Analiza danych.

W tej części pracy przeprowadzona zostanie analiza danych, pochodzących z medycznego badania naukowego. Stanowi ona przykład praktycznego zastosowania teoretycznych założeń wielomianowej regresji logistycznej, które zostały opisane w poprzednim rozdziale. Analiza składać się będzie z opisu badania i wykorzystywanego zbioru danych, wyboru regresorów na podstawie przedstawionych wcześniej metod i porównania modeli pod względem predykcji nowych danych. Ostatecznie otrzymamy optymalny model, który pozwoli nam wysunąć pewne wnioski i sugestie skierowane do specjalistów medycznych.

3.1 Opis badania.

Analizowane dane pochodzą z badania przeprowadzonego retroaktywnie w latach 2019-2020, pod przewodnictwem Łukasza Antkowiaka, przez ośrodki medyczne z Katowic, Wrocławia, Gdańska, Sosnowca, Szczecina oraz Hamburga. Badaniu podlegali pacjenci z zespołem *Chiari I.* Jest to zespół związany z uwypuklaniem się migdałków móżdżku do otworu wielkiego, co z racji ucisku rdzenia kręgowego i pnia mózgu powoduje objawy zarówno ciasnoty wewnątrzczaszkowej (czyli m.in. bóle głowy i porażenia nerwów czaszkowych) jak i uszkodzenia rdzenia kręgowego (parestezje, drętwienia kończyn, zaniki mięśniowe, niedowłady). Leczenie polega na wykonaniu tzw. odbarczenia szczytowo-potylicznego, polegającego na wycięciu części kości potylicznej (technika *PFD - ang. Posterior Fossa Decompression*). Jest to podstawowa technika, którą często poszerza się o wszycie do worka oponowego pacjenta fragmentu powięzi (plastyka opony/duraplastyka) tak, żeby jeszcze bardziej zwiększyć miejsce w czaszce (technika *PFDD - ang. Posterior Fossa Decompression with Duraplasty*). Ostatecznie, najbardziej agresywną techniką jest poszerzenie dwóch powyższych o dodatkowe usunięcie migdałków móżdżku - czyli tego fragmentu mózgowia, który jest de facto odpowiedzialny za cały zespół *Chiari*. Ta technika nazywa się *PFDRT - ang. Posterior Fossa Decompression with Resection of the Tonsils*. Badanie miało na celu porównanie skuteczności tych 3 technik: *PFD*, *PFDD* i *PFDRT*, poprzez obserwowanie stanu pacjentów po operacji. Tymczasem nasza analiza skupi się wyłącznie na pacjentach, na których zastosowano technikę *PFDD* lub *PFDRT*, a jej celem będzie zbudowanie modelu przewidującego stan pacjenta po operacji na podstawie pewnych niezależnych zmiennych (wśród których może nie być informacji o stosowanej technice).

3.2 Opis zbioru danych.

Zbiór danych z badania dotyczy 73 anonimowych pacjentów, którzy opisywani są z pomocą 31 zmiennych. Pragniemy wykorzystać je do zbudowania modelu wielomianowej regresji logistycznej, który pozwoli przewidywać wynik operacji (zmienna objaśniana *Condition change* o trzech możliwych wartościach) na podstawie informacji o danym pacjencie (pewien podzbiór pozostałych 30 zmiennych, czyli naszych regresorów). Poniżej znajdują się dokładne opisy wszystkich zmiennych, a także tablica ich wzajemnych korelacji:

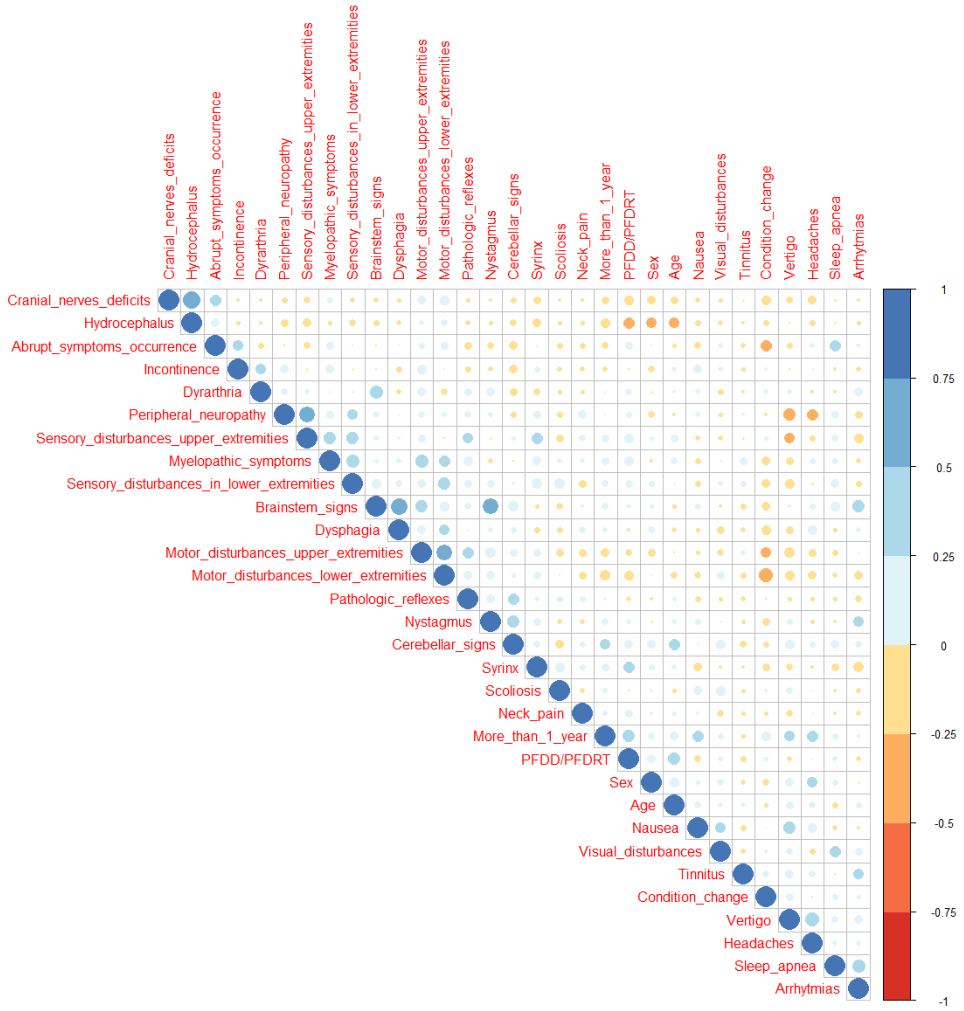
- *Sex* - płeć pacjenta oznaczona z pomocą wartości 1 (kobieta) i 0 (mężczyzna).

- *Age* - wiek pacjenta w latach.
- *Syrinx* - występowanie Syringomyelii, inaczej jamistości rdzenia - jest to nagromadzenie płynu w rdzeniu kręgowym, co powoduje ucisk neuronów rdzenia kręgowego, prowadząc do objawów typu zaburzenia czucia w kończynach lub niedowłady mięśniowe (1 - tak, 0 - nie).
- *More than one year* - czy objawy do momentu operacji trwały dłużej niż rok (1 - tak, 0 - nie).
- *Abrupt symptoms occurrence* - Czy objawy pojawiły się nagle, czy narastały stopniowo (1 - tak, 0 - nie).
- *Myelopathic symptoms* - Objawy mielopatii - często wynikają z obecności jamistości rdzenia. Są to objawy sugerujące często już nieodwracalne uszkodzenie neuronów rdzenia kręgowego - osłabienie mięśniowe, drętwienie kończyn, chwiejny chód (1 - tak, 0 - nie).
- *Peripheral neuropathy* - Objawy neuropatii obwodowej - w tym przypadku oznaczało to parestezje, czyli nieprawidłowe odczuwanie bodźców czuciowych (pieczenie, mrowienie itp). Często są mieszane z objawami mielopatii, ale mielopatia dotyczy rdzenia kręgowego, a neuropatia nerwów obwodowych (1 - tak, 0 - nie).
- *Sensory disturbances upper extremities* - Zaburzenia czucia w kończynach górnych - typowo brak czucia w danym obszarze, drętwienia kończyn (1 - tak, 0 - nie).
- *Sensory disturbances in lower extremities* - Zaburzenia czucia w dolnych kończynach - tak samo jak w górnych - brak czucia, drętwienie (1 - tak, 0 - nie).
- *Motor disturbances upper extremities* - Zaburzenia ruchowe w kończynach górnych - osłabienie mięśniowe, niedowład (1 - tak, 0 - nie).
- *Motor disturbances lower extremities* - Zaburzenia ruchowe w kończynach dolnych - podobnie jak w górnych - niedowłady, osłabienie mięśniowe (1 - tak, 0 - nie).
- *Vertigo* - Zawroty głowy (1 - tak, 0 - nie).
- *Nausea* - Nudności (1 - tak, 0 - nie).
- *Headaches* - Bóle głowy szeroko pojęte, bez specyfikacji w której części głowy ból był dokładnie zlokalizowany (1 - tak, 0 - nie).
- *Neck pain* - Ból karku - odróżniany od bólu podpotylicznego, ból karku jako umiejscowiony raczej nisko na karku, tymczasem ból podpotyliczny dotyczy potylicy i górnej części karku (1 - tak, 0 - nie).
- *Visual disturbances* - Zaburzenia widzenia, typowo mroczki przed oczami, chwilowa utrata widzenia w jednym oku (1 - tak, 0 - nie).
- *Tinnitus* - Szumy uszne (1 - tak, 0 - nie).

- *Nystagmus* - Oczopląs (1 - tak, 0 - nie).
- *Brainstem signs* - Objawy pniowe - typowo są to zaburzenia połykania (dysfagia) albo zaburzenia mowy (dysartria) (1 - tak, 0 - nie).
- *Dysphagia* - Zaburzenia połykania - objaw pniowy (1 - tak, 0 - nie).
- *Dyrarthria* - Zaburzenia mowy - objaw pniowy, choć nie specyficzny dla pnia, ponieważ może występować też w uszkodzeniu płatu skroniowego mózgu i innych szlaków istoty białej (1 - tak, 0 - nie).
- *Pathologic reflexes* - Odruchy patologiczne, wzmożone lub osłabione odruchy ścięgniste w kończynach. Świadczą o uszkodzeniu dróg ruchowych rdzenia kręgowego lub nerwów obwodowych (1 - tak, 0 - nie).
- *Cerebellar signs* - Objawy móżdżkowe, wynikające z ucisku/uszkodzenia móżdżku. Są to zaburzenia równowagi, niestabilny chód, zaburzenia koordynacji (1 - tak, 0 - nie).
- *Cranial nerves deficits* - Zaburzenia nerwów czaszkowych, czyli nerwów unerwiających obszar głowy i szyi. Są to np. zaburzenia czucia na twarzy, porażenie mięśni twarzy, brak czucia smaku, opadanie powieki, zez (1 - tak, 0 - nie).
- *Incontinence* - Nietrzymanie moczu - rzadki objaw, raczej wynikający z obecności zakotwiczenia rdzenia (ang. *tethered cord*) niż z *Chiari*. Spekuluje się, że zakotwiczenie rdzenia poprzez pociąganie rdzenia kręgowego w dół może powodować zespół *Chiari I*. Raczej objawy nietrzymania moczu nie są częste w *Chiari* (1 - tak, 0 - nie).
- *Sleep apnea* - Bezdech senny - również można go zaliczyć do objawów pniowych, gdyż świadczy o zaburzeniach w obrębie ośrodka oddychania w pniu mózgu (1 - tak, 0 - nie).
- *Arrhytmias* - Zaburzenia rytmu serca, które mogą wynikać z ucisku pnia mózgu, gdzie znajduje się ośrodek kontrolujący krążenie krwi (1 - tak, 0 - nie).
- *Scoliosis* - Boczne skrzywienie kręgosłupa (1 - tak, 0 - nie).
- *Hydrocephalus* - Poszerzenie układu komorowego na skutek zaburzenia przepływu płynu mózgowo-rdzeniowego z układu komorowego do przestrzeni podpajęczynówkowej. Na skutek ucisku drogi odpływu płynu mózgowo-rdzeniowego z komory czwartej dochodzi do gromadzenia się go w komorach i ich poszerzenia, co nazywamy wodogłowiem (1 - tak, 0 - nie).
- *PFDD/PFDRT* - Zastosowana technika operacyjna, *PFDD* (wartość 1) lub *PFDRT* (wartość 0). Typowy przebieg pierwszej ma następującą postać: pacjent leży na stole operacyjnym na brzuchu, z głową umocowaną w ramie Mayfielda (metalowa rama z 3 bolcami do umocowania czaszki). Głowa jest odgięta lekko w dół. Można też operować z pacjentem w pozycji siedzącej, ale to zwiększa ryzyko zatoru powietrznego podczas operacji. Potem nacina się skórę na potylicy (ok. 5 cm). Odciąga się na bok mięśnie, nawierca wiertarką szybkoobrotową 1-2 otwory w kości potylicy

i łączy je drugą wiertarką - kraniotomem. Po zdjęciu kości czaszki - ok 3x4 cm można zakończyć operację - jest to technika *PFD*. Aby zrobić z *PFD* technikę *PFDD*, należy naciąć oponę twardą w kształcie litery Y oraz wszyć w to nacięcie oponę syntetyczną lub powięź własną pacjenta (np. powięź szeroka uda/czepiec ścięgnisty na głowie). Natomiast w technice *PFDRT* wykonuje się analogiczne czynności, ale po nacięciu opony twardej wchodzi się narzędziami do worka oponowego i usuwa się migdałki móżdżku, które powodują ucisk pnia mózgu, a także utrudniają odpływ płynu mózgowo-rdzeniowego z komory czwartej do przestrzeni podpajęczynówkowej. W ten sposób usunięcie migdałków móżdżku prowadzi do odtworzenia prawidłowego przepływu płynu mózgowo-rdzeniowego oraz redukuje ucisk pnia mózgu. Po usunięciu migdałków móżdżku wykonuje się plastykę opony twardej, tak jak przy *PFDD*. Potem zszywa się mięśnie i skórę, ale bez ponownego przymocowania kości, które byłoby sprzeczne z podstawowym celem techniki - zwiększeniem miejsca w jamie czaszki.

- *Condition change* - zmienna oceniająca wyniki operacji (1 - pogorszenie stanu, 2 - brak zmian, 3 - poprawa stanu pacjenta).



Rysunek 1: Tablica korelacji.

Widzimy, że korelacje pomiędzy zmiennymi są generalnie niewielkie z nielicznymi wyjątkami, które nie dają jednak silnych podstaw do obaw przed wystąpieniem zjawiska multikolinearności.

3.3 Wybór regresorów.

Podstawowym problemem do rozwiązania jest odpowiednie dobranie regresorów. Widzimy bowiem, że mamy do czynienia ze stosunkowo niewielką liczbą obserwacji, zwłaszcza w porównaniu z liczbą zmiennych. Proporcje te mogą prowadzić do wystąpienia mocno niepożądanego zjawiska: nadmiernego dopasowania modelu do danych (ang. *Overfitting*), które polega na tym, że model silnie dopasowuje się do danych, na podstawie których został zbudowany, w związku z czym wykazuje słabe własności predykcyjne na nowych danych. *Overfitting* nie tylko oznacza, że model nie spełnia dobrze założonych przez nas celów, ale utrudnia również estymowanie parametrów regresji z pomocą metod numerycz-

nych. W przypadku naszego zbioru danych zjawisko to jest na tyle mocne, że nie jesteśmy w stanie zbudować modelu zawierającego wszystkie regresory. Do wyboru regresorów wykorzystamy metody opisane w części teoretycznej, z pomocą których uzyskamy trzech kandydatów.

Pierwszy model konstruujemy przy pomocy metody *Lasso* w oparciu o sprawdzian krzyżowy typu *Leave-one-out*. Metoda dokonała selekcji dwóch regresorów: *Abrupt symptoms occurrence* oraz *Motor disturbances lower extremities*. Następnie dwukrotnie przeprowadzamy selekcję typu *Forward*, z użyciem kryterium *AIC* oraz kryterium *BIC*:

Krok	*AIC*	*BIC*
0	144	148
1	133	142
2	131	144
3	129	147
4	128	151.8
5	124	151.9
6	123	156
7	120.3	157
8	120.6	161

Na kandydatów wybieramy te modele, dla których wartości kryteriów przyjęły po raz pierwszy minimum lokalne. W przypadku *AIC* osiągnęliśmy to mininum w kroku siódmym, a w przypadku *BIC* już w kroku pierwszym. W ten sposób otrzymujemy wszystkich kandydatów, których w dalszej części będziemy porównywać w poszukiwaniu najlepszego modelu:

- Model *Lasso - Abrupt symptoms occurrence, Motor disturbances lower extremities*,
- Model *AIC - Motor disturbances lower extremities, Sensory disturbances in lower extremities, Abrupt symptoms occurrence, Dysphagia, Neck pain, Syrinx, Incontinence*,
- Model *BIC - Motor disturbances lower extremities*.

3.4 Porównanie modeli.

Porównanie modeli rozpoczniemy od przedstawienia ich podstawowych cech w postaci tabel, zawierających predyktory modelu, estymacje ich parametrów (a, b) oraz odchylenia standardowe tych estymatorów ($\sigma(a)$, $\sigma(b)$), a także wartości statystyk pomagających w ocenie modelu: *Deviance*, *AIC* i *BIC*:

Tabela 1: Parametry modelu Lasso.

Zmienna	a	$\sigma(a)$	b	$\sigma(b)$
Intercept	1.25	0.60	2.39	0.55
Motor disturbances lower extremities	-2.04	0.87	-2.60	0.77
Abrupt symptoms occurrence	-1.34	1.02	-1.98	0.94

Tabela 2: Parametry modelu AIC.

Zmienna	a	$\sigma(a)$	b	$\sigma(b)$
Intercept	3.54	1.32	4.19	1.28
Dysphagia	-39.56	0.01	-2.74	1.66
Incontinence	28.24	0.70	26.63	0.70
Neck pain	26.22	793.47	11.75	733.30
Abrupt symptoms occurrence	-3.50	1.59	-3.53	1.32
Syrinx	-3.28	1.37	-1.98	1.24
Sensory disturbances in lower extremities	-1.58	1.40	0.03	0.95
Motor disturbances lower extremities	-1.16	1.19	-2.42	0.97

Tabela 3: Parametry modelu BIC.

Zmienna	a	$\sigma(a)$	b	$\sigma(b)$
Intercept	0.96	0.53	2.03	0.48
Motor disturbances lower extremities	-2.05	0.85	-2.62	0.73

Tabela 4: Statystyki modelów.

Model	*Deviance*	*AIC*	*BIC*
Model *Lasso*	121.0708	133.0708	146.8136
Model *AIC*	88.30463	120.3046	156.952
Model *BIC*	125.6929	133.6929	142.8548

Widzimy, że model *LASSO* i model *BIC* są bardzo zbliżone. Zbudowane są na bardzo niewielu regresorach i mają stabilne estymatory parametrów. Wykazują również niższe wartości kryterium *BIC* od trzeciego modelu. Tymczasem model *AIC* stanowi rozszerzenie pozostałych modelów o dodatkowe predyktory, dzięki którym wykazuje istotnie niższą wartość statystyki Deviance, ale w zamian za to niektóre estymatory jego parametrów są dosyć niestabilne (przy zmiennej *Neck pain* widzimy olbrzymie odchylenia standardowe).

Najważniejszym zadaniem naszego modelu jest sprawne przewidywanie zmiennej objaśnianej na podstawie nowych danych. Dlatego do podjęcia finałowej decyzji odnośnie wyboru najlepszego modelu potrzebny nam jest treningowy i testowy zbiór danych. W tym celu wybieramy losowo 80% obserwacji z pełnego zbioru w taki sposób, żeby zachować strukturę danych, czyli oryginalne proporcje wszystkich trzech stanów zmiennej objaśnianej (1 - 14 obserwacji, 2 - 16 obserwacji, 3 - 43 obserwacje). Otrzymujemy tym sposobem zbiór treningowy, na podstawie którego będziemy konstruować modele. Za to pozostałe 20% danych tworzy zbiór testowy, będący źródłem nowych danych. Posiadając te podzbiory danych, możemy dokonać porównania modeli pod względem trzech wartości:

- *Training Prediction*, czyli procent prawidłowych predykcji danych ze zbioru treningowego, z pomocą modelu zbudowanego na zbiorze treningowym

(czyli sytuacja, kiedy model przewiduje „znajome" dane),

- *Test Prediction*, czyli procent prawidłowych predykcji danych ze zbioru testowego, z pomocą modelu zbudowanego na zbiorze treningowym (czyli sytuacja, kiedy model przewiduje nowe dane),
- *Average Prediction*, czyli średnia wartość *Test Prediction*, uzyskana poprzez powtórzenie procesu losowego podziału pełnego zbioru danych na zbiór treningowy i testowy 10000 razy (przy czym regresory w modelach pozostają takie same, nie wykonujemy ponownie wyboru regresorów regresją typu *Forward* ani metodą *Lasso*).

Model	*Training Prediction*	*Test Prediction*	*Average Prediction*
Model *Lasso*	66.67%	56.25%	58.14813%
Model *AIC*	78.95%	62.5%	63.08937%
Model *BIC*	66.67%	56.25%	61.77875%

Na powyższej tabeli widzimy, że nasze modele radzą sobie dosyć dobrze z predykcją zarówno nowych danych, jak i estymacją danych, na podstawie których zostały skonstruowane. Zauważyć możemy, że modele *Lasso* i *BIC* osiągnęły bardzo podobne wyniki i dopiero uśredniona predykcja wskazuje ten drugi jako lepszy. Na tym etapie najlepszy wydaje się jednak model *AIC*.

Na koniec porównania przyjrzymy się dokładniej przeciętnej predykcji, która informuje nas o tym, jak dobrze model przewiduje nowe dane, ale nie jest mocno uzależniona od pojedynczo wylosowanego zbioru treningowego i testowego. Dla każdego modelu przedstawimy macierz rozmiaru 3x3, której element z i-tego wiersza i j-tej kolumny przedstawia przeciętny procent predykcji stanu j-tego, kiedy prawdziwy jest stan i-ty. Jest to tzw. tabela pomyłek w ujęciu procentowym, w którym wartości w każdym wierszu sumują się do 100%.

Tabela 5: Procentowa tabela pomyłek modelu *Lasso*.

	Przewidziany stan 1	Przewidziany stan 2	Przewidziany stan 3
Prawdziwy stan 1	50.15%	1.33%	48.52%
Prawdziwy stan 1	16.75%	0.01%	83.24%
Prawdziwy stan 1	12.54%	0.75%	86.72%

Tabela 6: Procentowa tabela pomyłek modelu *AIC*.

	Przewidziany stan 1	Przewidziany stan 2	Przewidziany stan 3
Prawdziwy stan 1	61.01%	1.86%	37.13%
Prawdziwy stan 2	6.55%	23.51%	69.45%
Prawdziwy stan 3	9.87%	8.76%	81.38%

Tabela 7: Procentowa tabela pomyłek modelu *BIC*.

	Przewidziany stan 1	Przewidziany stan 2	Przewidziany stan 3
Prawdziwy stan 1	64.23%	0%	35.77%
Prawdziwy stan 2	18.76%	0%	81.24%
Prawdziwy stan 3	11.61%	0%	88.39%

Zauważyć możemy, że wszystkie modele dokonują najdokładniejszej predykcji w sytuacji, gdy prawdziwa wartość zmiennej wynikowej wynosi 3 (poprawa stanu pacjenta). Trochę gorzej, ale nadal zadowalająco radzą sobie z prawidłowym przewidywaniem stanu 1 (pogorszenie stanu pacjenta). Tymczasem wyłącznie model *AIC* jest w stanie przewidywać stan 2 (brak zmian stanu pacjenta), chociaż procent prawidłowych predykcji jest niski. Dosyć zaskakujący wynik możemy dostrzec w przypadku modelu *BIC*, który zupełnie pomija ten stan zmiennej objaśnianej. Może to wynikać z występowania tylko jednego regresora w modelu i zdaje się wskazywać, że byłby to potencjalnie dobry model regresji logistycznej (co wymagałoby oczywiście przekształcenia przewidywanej zmiennej na binarną).

Po dokładnym porównaniu predykcyjnych zdolności kandydatów, za najbardziej optymalny możemy uznać model *AIC*. Warto zaznaczyć, że model *BIC* wykazuje podobne właściwości i jest prostszy ze względu na mniejszą liczbę regresorów, ale wyklucza go niezdolność do predykcji braku zmian w stanie pacjenta po operacji. W następnym rozdziale zajmiemy się opisaniem wniosków, które możemy wyciągnąć na podstawie powyższych informacji oraz wybranego modelu. Wpierw jednak poruszymy pewną kwestię. W trakcie naszej analizy wielokrotnie zaznaczana była istotność przebadania modeli pod kątem ich zdolności predykcyjnych dla nowych danych. Stąd brała się potrzeba stworzenia zbioru treningowego i testowego. W trakcie poszukiwania wstępnych kandydatów na model zastosowaliśmy jednak metody wyboru regresorów na pełnym zbiorze danych, co może wydawać się sprzeczne z intuicją czy powszechnymi praktykami. Zauważmy, że metody te stanowiły jedynie pewne ukierunkowanie w celu przyspieszenia analizy. Hipotetycznie można by skonstruować na podstawie zbioru treningowego wszystkie możliwe modele, których byłoby 2^{30} (tyle co podzbiorów całego zbioru regresorów), a następnie porównać je pod względem zdolności predykcyjnych. Wśród badanych modeli znalazłyby się oczywiście również te, które uzyskaliśmy w trakcie naszej analizy. Warto również zaznaczyć, że w przypadku wielu zbiorów danych zastosowanie metod wyboru regresorów może dawać takie same wyniki na podstawie pełnego zbioru danych, jak i zbioru treningowego. W naszym przypadku jednak liczba obserwacji jest na tyle niewielka, że zignorowanie 20% z nich mogło doprowadzić do utraty istotnej informacji, prowadząc do uzyskania gorszych kandydatów na model. Uwaga ta dotyczy przede wszystkim kryteriów *AIC* i *BIC*, ale analogicznie potraktowana została metoda *Lasso* w celu zachowania zgodności. Oznacza to, że po jednorazowym zastosowaniu jej na pełnym zbiorze danych, w trakcie dalszej analizy konsekwentnie konstruowaliśmy modele na podstawie dwóch regresorów, które wybraliśmy na samym początku.

4 Wnioski.

Wybrany przez nas model wskazuje nam zmienne, które wystarczająco dobrze przewidują stan pacjenta po operacji. Są to zmienne informujące o symptomach pacjenta: zaburzenia czucia i zaburzenia ruchowe w dolnych kończynach, jamistość rdzenia (Syringomyelia), ból karku, zaburzenia połykania oraz nietrzymanie moczu, a także zmienna informująca o nagłości wystąpienia symptomów. Estymatory parametrów pokazują, że największy wpływ na otrzymaną predykcję mają zmienne *Dysphagia*, *Neck pain* (przy tym regresorze należy jednak zaznaczyć bardzo duże odchylenie estymatora, które może zniekształcać wyciągane wnioski) oraz *Incontinence*, a pozostałe zmienne mają mniejszy, choć nadal istotny wpływ. Otrzymany przez nas model każe nam również wnioskować, że odrzucone zmienne nie były wystarczająco istotne do przewidywania wyniku operacji, co w szczególności sygnalizuje brak zauważalnych różnic pomiędzy wykorzystaniem techniki *PFDD* oraz *PFDRT* w tym kontekście. Podobnie mało istotne wydają się takie informacje jak płeć i wiek pacjenta, a także czas występowania objawów.

Efektem całej naszej analizy jest konkretny model regresji, dzięki któremu możemy osobno szacować prawdopodobieństwa poprawy, braku zmian lub pogorszenia długoterminowego stanu pacjenta po operacji na podstawie dowolnej kombinacji regresorów. W ten sposób utworzone zostały tabele predykcji, które znajdują się w ostatnim rozdziale. Pierwsza zawiera zbiór wartości predyktorów, dla których operacja jest mocno zalecana, bowiem szansa na poprawę jest bardzo wysoka przy małym prawdopodobieństwie innych wyników. Druga tabela zawiera kombinacje regresorów, dla których istnieje jednocześnie wysokie prawdopodobieństwo sukcesu, ale również dość duża szansa pogorszenia stanu pacjenta. Trzecia tabela zawiera kombinacje, dla których mocno odradza się przeprowadzenie operacji, tymczasem w ostatniej tabeli znajdują się kombinacje, przy których zarówno prawdopodobieństwo pogorszenia jak i poprawy stanu pacjenta są bardzo niskie, a najbardziej prawdopodobny jest brak zmian. Sytuacja ta nakazuje przeanalizować, czy przeprowadzenie operacji jest warte inherentnemu ryzyku, które wiąże się z każdym zabiegiem medycznym.

Na sam koniec niniejszej pracy zaznaczyć należy ograniczenia uzyskanych wyników. Cała analiza danych oparta była na stosunkowo małym zbiorze danych, który może niedokładnie reprezentować większą populację. Nawet najbardziej optymalny ze znalezionych modelów jest obarczony istotnym błędem predykcji, a w szczególności otrzymana macierz przeciętnych predykcji wykazuje, że nasz model najlepiej radzi sobie, kiedy prawdziwą wartością zmiennej objaśnianej jest poprawa stanu pacjenta (co nie jest zaskakujące, mając na uwadze proporcje wszystkich trzech stanów). Oznacza to, że przedstawione w poprzedniej sekcji estymacje prawdopodobieństw są skrzywione w kierunku poprawy stanu i przedstawiają operacje jako bardziej pomyślne, niż w rzeczywistości. Stąd najbardziej cenne informacje możemy wyciągnąć z tabeli kombinacji regresorów, dla których odradza się operowanie. Należy również zaznaczyć, że jeden z estymatorów parametrów naszego modelu wykazuje bardzo wysoką wariancję, co obarcza model dodatkowym błędem. Ze wszystkich tych powodów zalecane jest przeprowadzenie kolejnych badań na temat zespołu *Chiari I* oraz technik *PFDD* i *PFDRT*, w szczególności badań o większej liczbie obserwacji. Intencją niniejszej pracy jest nadanie wstępnego kierunku tym badaniom i dalszym analizom, a także dostarczenie specjalistom medycznym informacji, które

pomogą w podejmowaniu decyzji o przeprowadzeniu operacji.

5 Tablice predykcji.

Rysunek 2: Symptomy, dla których jest wysokie prawdopodobieństwo poprawy.

Motor	Sensory	Abrupt_symptoms	Dysphagia	Neck_pain	Syrinx	Incontinence	Pogorszenie	Brak_zmian	Poprawa
0	0	0	1	0	0	1	0.00	0.00	1.00
0	0	0	1	0	1	1	0.00	0.00	1.00
0	0	0	1	1	0	0	0.00	0.00	1.00
0	0	0	1	1	0	1	0.00	0.00	1.00
0	0	0	1	1	1	0	0.00	0.00	1.00
0	0	0	1	1	1	1	0.00	0.00	1.00
0	0	1	1	0	0	1	0.00	0.00	1.00
0	0	1	1	0	1	1	0.00	0.00	1.00
0	0	1	1	1	0	0	0.00	0.00	1.00
0	0	1	1	1	0	1	0.00	0.00	1.00
0	0	1	1	1	1	0	0.00	0.00	1.00
0	0	1	1	1	1	1	0.00	0.00	1.00
0	1	0	0	0	0	0	0.01	0.09	0.89
0	1	0	0	0	1	1	0.00	0.13	0.87
0	1	0	1	0	0	1	0.00	0.00	1.00
0	1	0	1	0	1	1	0.00	0.00	1.00
0	1	0	1	1	0	0	0.00	0.00	1.00
0	1	0	1	1	0	1	0.00	0.00	1.00
0	1	0	1	1	1	0	0.00	0.00	1.00
0	1	0	1	1	1	1	0.00	0.00	1.00
0	1	1	0	0	1	1	0.00	0.13	0.87
0	1	1	1	0	0	1	0.00	0.00	1.00
0	1	1	1	0	1	1	0.00	0.00	1.00
0	1	1	1	1	0	0	0.00	0.00	1.00

Rysunek 3: Symptomy, dla których jest wysokie prawdopodobieństwo poprawy (kont.).

Motor	Sensory	Abrupt_symptoms	Dysphagia	Neck_pain	Syrinx	Incontinence	Pogorszenie	Brak_zmian	Poprawa
0	1	1	1	1	0	1	0.00	0.00	1.00
0	1	1	1	1	1	0	0.00	0.00	1.00
0	1	1	1	1	1	1	0.00	0.00	1.00
1	0	0	1	0	0	1	0.00	0.00	1.00
1	0	0	1	0	1	1	0.00	0.00	1.00
1	0	0	1	1	0	0	0.00	0.00	1.00
1	0	0	1	1	0	1	0.00	0.00	1.00
1	0	0	1	1	1	0	0.00	0.00	1.00
1	0	0	1	1	1	1	0.00	0.00	1.00
1	0	1	1	0	0	1	0.00	0.00	1.00
1	0	1	1	0	1	1	0.00	0.00	1.00
1	0	1	1	1	0	0	0.00	0.00	1.00
1	0	1	1	1	0	1	0.00	0.00	1.00
1	0	1	1	1	1	0	0.01	0.00	0.99
1	0	1	1	1	1	1	0.00	0.00	1.00
1	1	0	1	0	0	1	0.00	0.00	1.00
1	1	0	1	0	1	1	0.00	0.00	1.00
1	1	0	1	1	0	0	0.00	0.00	1.00
1	1	0	1	1	0	1	0.00	0.00	1.00
1	1	0	1	1	1	0	0.00	0.00	1.00
1	1	0	1	1	1	1	0.00	0.00	1.00
1	1	1	1	0	0	1	0.00	0.00	1.00
1	1	1	1	0	1	1	0.00	0.00	1.00
1	1	1	1	1	0	0	0.00	0.00	1.00
1	1	1	1	1	0	1	0.00	0.00	1.00
1	1	1	1	1	1	0	0.00	0.00	1.00
1	1	1	1	1	1	1	0.00	0.00	1.00

Rysunek 4: Symptomy, dla których jest wysokie prawdopodobieństwo zarówno poprawy jak i pogorszenia.

Motor	Sensory	Abrupt_symptoms	Dysphagia	Neck_pain	Syrinx	Incontinence	Pogorszenie	Brak_zmian	Poprawa
0	0	0	0	0	1	0	0.09	0.11	0.80
0	0	0	1	0	0	0	0.19	0.00	0.81
0	1	0	0	0	1	0	0.09	0.03	0.88
0	1	0	1	0	0	0	0.19	0.00	0.81

Rysunek 5: Symptomy, dla których jest wysokie prawdopodobieństwo pogorszenia.

Motor	Sensory	Abrupt_symptoms	Dysphagia	Neck_pain	Syrinx	Incontinence	Pogorszenie	Brak_zmian	Poprawa
0	0	0	1	0	1	0	0.63	0.00	0.37
0	0	1	0	0	0	0	0.25	0.26	0.49
0	0	1	0	0	1	0	0.77	0.03	0.20
0	0	1	1	0	0	0	0.89	0.00	0.11
0	0	1	1	0	1	0	0.98	0.00	0.02
0	1	0	1	0	1	0	0.62	0.00	0.38
0	1	1	0	0	0	0	0.31	0.07	0.62
0	1	1	0	0	1	0	0.78	0.01	0.21
0	1	1	1	0	0	0	0.89	0.00	0.11
0	1	1	1	0	1	0	0.98	0.00	0.02
1	0	0	0	0	0	0	0.06	0.61	0.33
1	0	0	0	0	1	0	0.45	0.18	0.36
1	0	0	1	0	0	0	0.73	0.00	0.27
1	0	0	1	0	1	0	0.95	0.00	0.05
1	0	1	0	0	0	0	0.67	0.22	0.11
1	0	1	0	0	1	0	0.97	0.01	0.02
1	0	1	1	0	0	0	0.99	0.00	0.01
1	0	1	1	0	1	0	1.00	0.00	0.00
1	1	0	0	0	0	0	0.11	0.24	0.65
1	1	0	0	0	1	0	0.52	0.04	0.43
1	1	0	1	0	0	0	0.72	0.00	0.28
1	1	0	1	0	1	0	0.95	0.00	0.05
1	1	1	0	0	0	0	0.80	0.05	0.14
1	1	1	0	0	1	0	0.97	0.00	0.02
1	1	1	1	0	0	0	0.99	0.00	0.01
1	1	1	1	0	1	0	1.00	0.00	0.00

Rysunek 6: Symptomy, dla których jest wysokie prawdopodobieństwo braku zmian.

Motor	Sensory	Abrupt_symptoms	Dysphagia	Neck_pain	Syrinx	Incontinence	Pogorszenie	Brak_zmian	Poprawa
0	0	0	0	0	0	1	0	0.73	0.27
0	0	0	0	0	1	1	0	0.42	0.58
0	0	0	0	1	0	0	0	1.00	0.00
0	0	0	0	1	0	1	0	1.00	0.00
0	0	0	0	1	1	0	0	1.00	0.00
0	0	0	0	1	1	1	0	1.00	0.00
0	0	1	0	0	0	1	0	0.73	0.27
0	0	1	0	0	1	1	0	0.43	0.57
0	0	1	0	1	0	0	0	1.00	0.00
0	0	1	0	1	0	1	0	1.00	0.00
0	0	1	0	1	1	0	0	1.00	0.00
0	0	1	0	1	1	1	0	1.00	0.00
0	1	0	0	0	0	1	0	0.34	0.66
0	1	0	0	1	0	0	0	1.00	0.00
0	1	0	0	1	0	1	0	1.00	0.00
0	1	0	0	1	1	0	0	1.00	0.00
0	1	0	0	1	1	1	0	1.00	0.00
0	1	1	0	0	0	1	0	0.35	0.65
0	1	1	0	1	0	0	0	1.00	0.00
0	1	1	0	1	0	1	0	1.00	0.00
0	1	1	0	1	1	0	0	1.00	0.00
0	1	1	0	1	1	1	0	1.00	0.00

Rysunek 7: Symptomy, dla których jest wysokie prawdopodobieństwo braku zmian (kont.).

Motor	Sensory	Abrupt_symptoms	Dysphagia	Neck_pain	Syrinx	Incontinence	Pogorszenie	Brak_zmian	Poprawa
1	0	0	0	0	0	1	0	0.90	0.10
1	0	0	0	0	1	1	0	0.72	0.28
1	0	0	0	1	0	0	0	1.00	0.00
1	0	0	0	1	0	1	0	1.00	0.00
1	0	0	0	1	1	0	0	1.00	0.00
1	0	0	0	1	1	1	0	1.00	0.00
1	0	1	0	0	0	1	0	0.91	0.09
1	0	1	0	0	1	1	0	0.72	0.28
1	0	1	0	1	0	0	0	1.00	0.00
1	0	1	0	1	0	1	0	1.00	0.00
1	0	1	0	1	1	0	0	1.00	0.00
1	0	1	0	1	1	1	0	1.00	0.00
1	1	0	0	0	0	1	0	0.65	0.35
1	1	0	0	0	1	1	0	0.34	0.66
1	1	0	0	1	0	0	0	1.00	0.00
1	1	0	0	1	0	1	0	1.00	0.00
1	1	0	0	1	1	0	0	1.00	0.00
1	1	0	0	1	1	1	0	1.00	0.00
1	1	1	0	0	0	1	0	0.66	0.34
1	1	1	0	0	1	1	0	0.34	0.66
1	1	1	0	1	0	0	0	1.00	0.00
1	1	1	0	1	0	1	0	1.00	0.00
1	1	1	0	1	1	0	0	1.00	0.00
1	1	1	0	1	1	1	0	1.00	0.00

Literatura

[1] S. H. Walker and D. B. Duncan, "Estimation of the probability of an event as a function of several independent variables," *Biometrika*, vol. 54, no. 1/2, pp. 167–179, 1967.

[2] J. Engel, "Polytomous logistic regression," *Statistica Neerlandica*, vol. 42, no. 4, pp. 233–252, 1988.

[3] D. W. Hosmer and W. David, "Stanley lemeshow, rodney x. sturdivant," *Applied logistic regression*, 2013.

[4] P. X.-K. Song and P. X.-K. Song, *Correlated data analysis: modeling, analytics, and applications*, vol. 1. Springer, 2007.

[5] H. Akaike, *Information Theory and an Extension of the Maximum Likelihood Principle*, pp. 199–213. New York, NY: Springer New York, 1998.

[6] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[7] A. Ralston and H. S. Wilf, "Mathematical methods for digital computers," tech. rep., 1960.

[8] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the royal statistical society series b-methodological*, vol. 58, pp. 267–288, 1996.

[9] D. M. Allen, "The relationship between variable selection and data agumentation and a method for prediction," *Technometrics*, vol. 16, no. 1, pp. 125–127, 1974.

[10] T. Hastie, R. Tibshirani, and M. Wainwright, "Statistical learning with sparsity," *Monographs on statistics and applied probability*, vol. 143, p. 143, 2015.